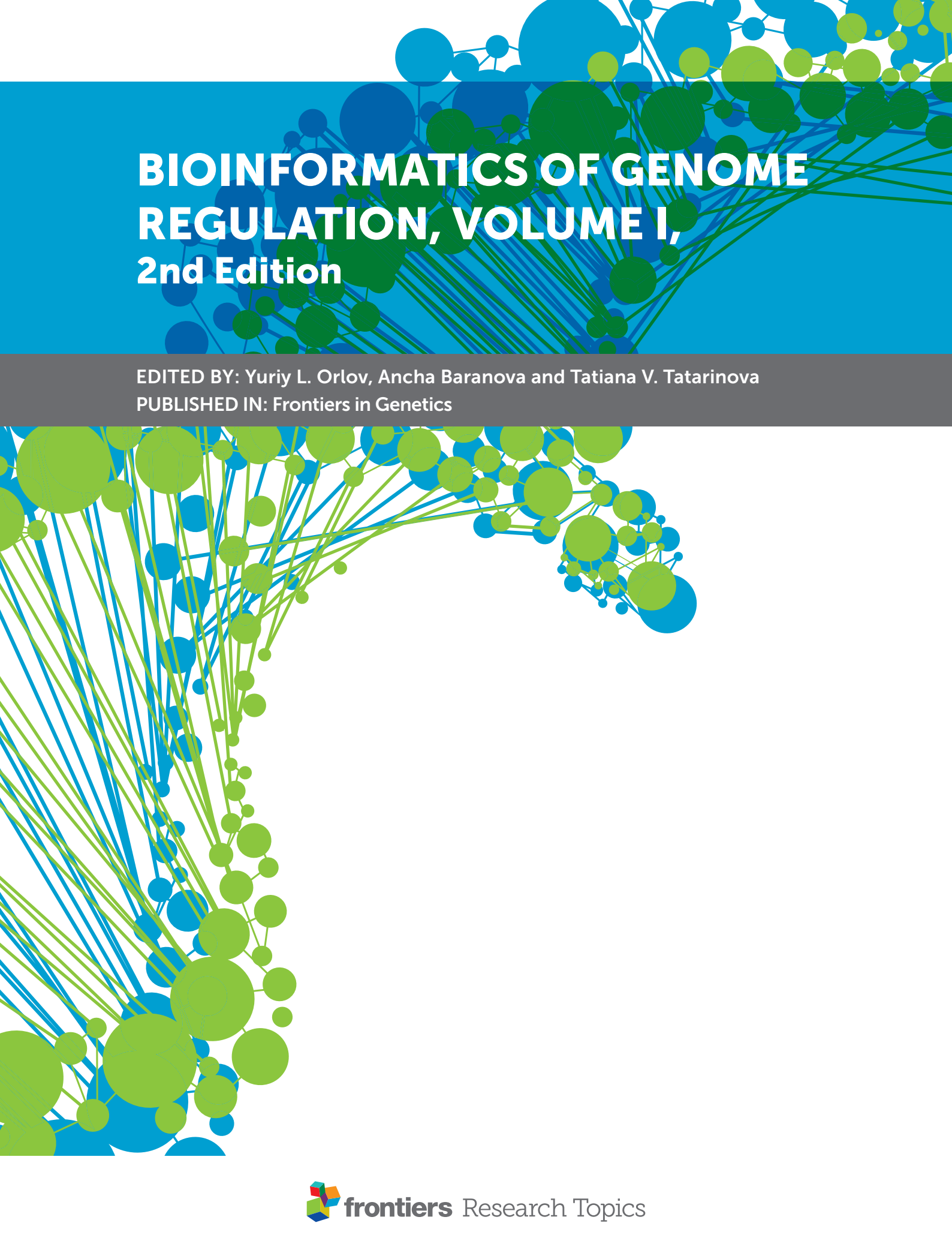




BIOINFORMATICS OF GENOME REGULATION, VOLUME I, 2nd Edition

EDITED BY: Yuriy L. Orlov, Ancha Baranova and Tatiana V. Tatarinova
PUBLISHED IN: Frontiers in Genetics



frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88974-142-7

DOI 10.3389/978-2-88974-142-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.

Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

BIOINFORMATICS OF GENOME REGULATION, VOLUME I, 2nd Edition

Topic Editors:

Yuriy L. Orlov, I. M. Sechenov First Moscow State Medical University, Russia

Ancha Baranova, George Mason University, United States

Tatiana V. Tatarinova, University of La Verne, United States

Publisher's note: In this 2nd edition, the following article has been updated: Orlov YL, Tatarinova TV, Oparina NY, Galieva ER and Baranova AV (2021) Editorial: Bioinformatics of Genome Regulation, Volume I. *Front. Genet.* 12:803273. doi: 10.3389/fgene.2021.803273

Citation: Orlov, Y. L., Baranova, A., Tatarinova, T. V., eds. (2022). Bioinformatics of Genome Regulation, Volume I, 2nd Edition. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-142-7

Table of Contents

- 05 Editorial: Bioinformatics of Genome Regulation, Volume I**
Yuriy L. Orlov, Tatiana V. Tatarinova, Nina Y. Oparina, Elvira R. Galieva and Ancha V. Baranova
- 08 In silico Prediction of miRNA Interactions With Candidate Atherosclerosis Gene mRNAs**
Dina Mukushkina, Dana Aisina, Anna Pyrkova, Alma Ryskulova, Siegfried Labeit and Anatoliy Ivashchenko
- 20 Structural and Functional Characteristics of miRNAs in Five Strategic Millet Species and Their Utility in Drought Tolerance**
Animikha Chakraborty, Aswini Viswanath, Renuka Malipatil, Abhishek Rathore and Nepolean Thirunavukkarasu
- 35 Bioinformatics Analysis of Actin Molecules: Why Quantity Does Not Translate Into Quality?**
Anna V. Glyakina and Oxana V. Galzitskaya
- 42 NETO2 Is Deregulated in Breast, Prostate, and Colorectal Cancer and Participates in Cellular Signaling**
Maria S. Fedorova, Anastasiya V. Snezhkina, Anastasiya V. Lipatova, Vladislav S. Pavlov, Anastasiya A. Kobelyatskaya, Zulfiya G. Guvatova, Elena A. Pudova, Maria V. Savvateeva, Irina A. Ishina, Tatiana B. Demidova, Nadezhda N. Volchenko, Dmitry Y. Trofimov, Gennady T. Sukhikh, George S. Krasnov and Anna V. Kudryavtseva
- 50 Characterization of Genome-Wide DNA Methylation and Hydroxymethylation in Mouse Arcuate Nucleus of Hypothalamus During Puberty Process**
Yihang Shen, Shasha Zhou, Xiaodong Zhao, Hua Li and Jieli Sun
- 58 ChIAMM: A Mixture Model for Statistical Analysis of Long-Range Chromatin Interactions From ChIA-PET Experiments**
Yibeltal Arega, Hao Jiang, Shuangqi Wang, Jingwen Zhang, Xiaohui Niu and Guoliang Li
- 68 Refinement of Draft Genome Assemblies of Pigeonpea (*Cajanus cajan*)**
Soma S. Marla, Pallavi Mishra, Ranjeet Maurya, Mohar Singh, Dhammaprakash Pandhari Wankhede, Anil Kumar, Mahesh C. Yadav, N. Subbarao, Sanjeev K. Singh and Rajesh Kumar
- 80 Identification of Molecular Markers Associated With the Pathophysiology and Treatment of Lupus Nephritis Based on Integrated Transcriptome Analysis**
Menghui Yao, Congcong Gao, Chunyi Zhang, Xueqi Di, Wenfang Liang, Wenbo Sun, Qianqian Wang and Zhaohui Zheng
- 92 An Immune Gene-Related Five-lncRNA Signature for to Predict Glioma Prognosis**
Xinzhuang Wang, Ming Gao, Junyi Ye, Qiuyi Jiang, Quan Yang, Cheng Zhang, Shengtao Wang, Jian Zhang, Ligang Wang, Jianing Wu, Hua Zhan, Xu Hou, Dayong Han and Shiguang Zhao

- 107 ***Mutation Frequency in Main Susceptibility Genes Among Patients With Head and Neck Paragangliomas***
Anastasiya V. Snezhkina, Maria S. Fedorova, Vladislav S. Pavlov, Dmitry V. Kalinin, Alexander L. Golovyuk, Elena A. Pudova, Zulfiya G. Guvatova, Nataliya V. Melnikova, Alexey A. Dmitriev, George S. Razmakhaev, Andrey A. Poloznikov, Galina S. Alekseeva, Andrey D. Kaprin, George S. Krasnov and Anna V. Kudryavtseva
- 113 ***Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies***
Nadezhda Biziukova, Olga Tarasova, Sergey Ivanov and Vladimir Poroikov
- 125 ***Comparative Expression Analysis of Stress-Inducible Candidate Genes in Response to Cold and Drought in Tea Plant [Camellia sinensis (L.) Kuntze]***
Lidiia S. Samarina, Alexandr V. Bobrovskikh, Alexey V. Doroshkov, Lyudmila S. Malyukova, Alexandra O. Matskiv, Ruslan S. Rakhmangulov, Natalia G. Koninskaya, Valentina I. Malyarovskaya, Wei Tong, Enhua Xia, Karina A. Manakhova, Alexey V. Ryndin and Yuriy L. Orlov
- 142 ***Sea Urchin as a Universal Model for Studies of Gene Networks***
Leonid Adonin, Anatoliy Drozdov and Nickolai A. Barlev
- 154 ***Gene Expression Changes and Associated Pathways Involved in the Progression of Prostate Cancer Advanced Stages***
Elena A. Pudova, George S. Krasnov, Anastasiya A. Kobelyatskaya, Maria V. Savvateeva, Maria S. Fedorova, Vladislav S. Pavlov, Kirill M. Nyushko, Andrey D. Kaprin, Boris Y. Alekseev, Dmitry Y. Trofimov, Gennady T. Sukhikh, Anastasiya V. Snezhkina and Anna V. Kudryavtseva
- 164 ***Predicting Genome Architecture: Challenges and Solutions***
Polina Belokopytova and Veniamin Fishman
- 179 ***ERBB1/2/3 Expression, Prognosis, and Immune Infiltration in Cutaneous Melanoma***
Shougang Liu, Rong Geng, Eryi Lin, Peizhen Zhao and Yongfeng Chen
- 195 ***Identification and Analysis of Potential Key Genes Associated With Hepatocellular Carcinoma Based on Integrated Bioinformatics Methods***
Zhuolin Li, Yao Lin, Bizhen Cheng, Qiaoxin Zhang and Yingmu Cai
- 209 ***Dynamical Modeling of the Core Gene Network Controlling Transition to Flowering in Pisum sativum***
Polina Pavlinova, Maria G. Samsonova and Vitaly V. Gursky
- 221 ***The Links of Ghrelin to Incretins, Insulin, Glucagon, and Leptin After Bariatric Surgery***
Daria Skuratovskaia, Maria Vulf, Nataliya Chasovskikh, Aleksandra Komar, Elena Kirienkova, Egor Shunkin, Pavel Zatolokin and Larisa Litvinova



Editorial: Bioinformatics of Genome Regulation, Volume I

Yuriy L. Orlov^{1,2,3,4*}, Tatiana V. Tatarinova⁵, Nina Y. Oparina⁶, Elvira R. Galieva³ and Ancha V. Baranova^{7,8}

¹Institute of Digital Medicine, I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia, ²Agrarian and Technological Institute, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia, ³Life Sciences Department, Novosibirsk State University, Novosibirsk, Russia, ⁴Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, ⁵Natural Science Division, La Verne University, La Verne, CA, United States, ⁶Institute of Medicine, University of Gothenburg, Göteborg, Sweden, ⁷School of Systems Biology, George Mason University, Fairfax, VA, United States, ⁸Research Centre for Medical Genetics, Moscow, Russia

Keywords: bioinformatics, computational genomics, gene expression, gene expression regulation, model organisms, genetics

Editorial on the Research Topic

Bioinformatics of Genome Regulation, Volume I

This “*Bioinformatics of Genome Regulation*” issue presents the studies in the field of bioinformatics of gene expression regulation. The materials were initially discussed in part at BGRS\SB-2020 (Bioinformatics of Genome Regulation and Structure\Systems Biology) multi-conference (<https://bgrssb.icgbio.ru/2020>) in Novosibirsk, Russia. The BGRS conference series is organized by the Institute of Cytology and Genetics SB RAS every other year since 1998. This issue consists of two parts—Volume I (<https://www.frontiersin.org/research-topics/14266/bioinformatics-of-genome-regulation-volume-i>) and Volume II (<https://www.frontiersin.org/research-topics/17947/bioinformatics-of-genome-regulation-volume-ii>). It continues the tradition of Research Topics at Frontiers in Genetics journal (<https://www.frontiersin.org/research-topics/8383/bioinformatics-of-genome-regulation-and-systems-biology>), which presents the works discussed at the BGRS-2018 meeting (Orlov et al., 2016; Tatarinova et al., 2019). Previous journal covering BGRS\SB conference were presented in the Journal of Bioinformatics and Computational Biology in 2012 (Orlov et al., 2015; Orlov et al., 2019a) and other platforms (Chen et al., 2017; Baranova et al., 2019; Orlov, 2019; Orlov, 2019b; Orlov et al., 2021a). Starting in 2018, actual research works on the mechanisms of gene expression regulation are presented in Frontiers in Genetics being extended as Volume II.

In Volume I of this Research Topic a total of 19 papers were arranged by two main areas—biomedical bioinformatics for human health and plant model studies. Biomedical papers start from bioinformatics applications to various cancers, including hepatocellular carcinoma, melanoma, brain tumors, prostate cancers, and paraganglioma.

Li et al. described a bioinformatics pipeline to reveal critical genes associated with hepatocellular carcinoma. The authors analyzed differentially expressed genes, followed by the Reconstruction of the protein-protein interaction (PPI) network. Eight hub genes significantly upregulated in carcinoma samples were highlighted and validated using GEPIA (Gene Expression Profiling Interactive Analysis) and Oncomine databases.

Fedorova et al. studied *NETO2* gene (neuropilin and tolloid-like 2) upregulation in diverse tumors, including ones originating in breast, prostate, and colorectal tissues. In addition, the authors evaluated *NETO2* functions in a short-lived fish model *Nothobranchius furzeri*.

OPEN ACCESS

Edited and reviewed by:

Richard D. Emes,
University of Nottingham,
United Kingdom

*Correspondence:

Yuriy L. Orlov
orlov@d-health.institute

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 27 October 2021

Accepted: 08 November 2021

Published: 06 December 2021

Citation:

Orlov YL, Tatarinova TV, Oparina NY,
Galieva ER and Baranova AV (2021)
Editorial: Bioinformatics of Genome
Regulation, Volume I.
Front. Genet. 12:803273.
doi: 10.3389/fgene.2021.803273

Liu et al. consider how cutaneous melanoma involves ERBB tyrosine kinase family members (ERBB receptor family) in its progression.

Wang et al. studied immune surveillance within the microenvironment in glioma. The authors highlighted long non-coding RNA (lncRNA) as the key in glioma progression.

Pudova et al. showed how the gene expression landscapes change during the progression of prostate cancer to its advanced stages. They described relevant networks and pathways pertinent to early recurrence. Rare neuroendocrine tumors were studied in Snezhkina et al. for the frequencies of the mutations within susceptibility genes such as *SDHx*. This work aids in understanding the immunochemistry analysis of *SDHx* genes in carotid paragangliomas reported earlier (Snezhkina et al., 2020).

Mukushkina et al. applied bioinformatics tools to study how miRNAs interact with other genes' products form atherosclerotic plaques.

Frontiers in Genetics' publications on human disease biomarkers are continued by the Research Topic "High-throughput sequencing-based investigation of chronic disease markers and mechanisms" (<https://www.frontiersin.org/research-topics/21036/high-throughput-sequencing-based-investigation-of-chronic-disease-markers-and-mechanisms>). Thus, Shi et al. (2021) studied molecular mechanisms related to alternative polyadenylation in gastric cancer; Chang et al. (2021) analyzed platinum-drug resistance mutations in advanced non-small cell lung cancer.

In this "Bioinformatics of Genome Regulation" Topic Yao et al. have reported the results of a search for molecular markers associated with complications of systemic lupus erythematosus.

Skuratovskaia et al. described the mechanisms regulating carbohydrate metabolism in Type 2 diabetes mellitus using the bioinformatics framework.

The approaches for analysis of gene expression regulation in human diseases from the population genetics point of view were highlighted in the Research Topic "Association Between Individuals' Genomic Ancestry and Variation in Disease Susceptibility" at Frontiers in Genetics (<https://www.frontiersin.org/research-topics/15891/association-between-individuals-genomic-ancestry-and-variation-in-disease-susceptibility>).

Kamenova et al. (2021) analyzed gene expression regulation by miRNA in Parkinson's disease. Zinchenko et al. (2021) studied rare hereditary diseases in Russia. Ramensky et al. (2021) discussed targeted sequencing of a set of clinically important genes associated with cardiovascular diseases.

Gozman et al. (2021) raised important actual problem of the role of genetic variance in disease severity in COVID-19 Patients. This problem is continuing to be actively discussed (Lu et al., 2021).

The following two articles in the Volume I highlight the mechanisms of epigenetic control revealed by gene network reconstruction in animal models. Shen et al. investigated the changes of DNA methylation and hydroxymethylation in mouse genome during puberty an emphasis on the activation of by hypothalamic gonadotropin-releasing hormone pathway.

Adonin et al. applied the methods of gene networks Reconstruction to a fascinating model organism—sea urchin *Strongylocentrotus purpuratus*.

The next group of articles in this Research Topic performed gene expression analysis in plants. This science field was presented at the bioinformatics conference series in Novosibirsk (Orlov et al., 2019b). In particular, Marla et al. performed the plant genome assembly from short sequencing reads of Pigeonpea (*Cajanus cajan*). Chakraborty et al. annotated miRNA functions in various millet species. Samarina et al. studied cold-resistance genes in the tea plant *Camellia sinensis*. Cold and drought stresses cause osmotic changes in the cells of the tea plant (Samarina et al., 2020). This study identified 45 stress-inducible candidate genes associated with cold and drought responses using homologous detection in related plant species. The gene network analysis revealed upregulated expression in the *ICE1*-related cluster of *bHLH* factors. Pavlinova et al. presents another application of network analysis in plant science; dynamical modeling of the core gene network controlling the transition to flowering in *Pisum sativum* was performed to extend previously developed non-linear regression models of the flowering in wild chickpea (Kozlov et al., 2019).

Finally, a set of novel computational techniques was developed for deciphering gene expression regulation in cells. Arega et al. presented a novel tool for 3D genomics modeling of long-range chromatin interactions, the ChIAMM algorithm, which utilizes ChIA-PET (Chromatin Interaction Analysis by Paired-End Tags sequencing) to estimate amounts of chromosome contacts (loops) mediated by a given transcription factor. The same authors' group has also described a 3D genome structure in cervical cancer cells (Adeel et al., 2021). The topic of genome architecture prediction based on 3D interaction maps in cell nuclei was further advanced by Belokopytova and Fishman. The authors reviewed high-throughput genome-wide chromatin profiling and chromosome contacts mapping using chromosome conformation capture techniques (Hi-C and ChIA-PET). The Research Topic "The Role of High-Order Chromatin Organization in Gene Regulation" (<https://www.frontiersin.org/research-topics/18088/the-role-of-high-order-chromatin7-organization-in-gene-regulation>) has been put together at Frontiers in Genetics by Drs. Fishman and Pindyurin.

In their brief report, Glyakina and Galzitskaya discuss bioinformatics modeling of actin molecules. Biziukova et al. utilized Machine Learning-based analysis of the scientific texts in HIV treatment to systematize information on small molecules, proteins, and genes related to the disease.

Overall, we are proud of the continuing Research Topic at Frontiers in Genetics we collated. We hope that you will find this paper collection a stimulating reading and consider coming to the next BGRS\SB conferences in Novosibirsk, Russia (<https://bgrssb.icgbio.ru/2022/>), and read the next continuing Research Topics in Frontiers (<https://www.frontiersin.org/research-topics/21036/high-throughput-sequencing-based-investigation-of-chronic-disease-markers-and-mechanisms>).

AUTHOR CONTRIBUTIONS

YO, TT, and AB organized the Research Topic as guest editors, supervised the reviewing of the manuscripts, NO and EG critically contributed both to the extension of the Topic and the reviewing process. All the authors wrote this Editorial paper. All authors contributed to the article and approved the submitted version.

FUNDING

The BGRS-2020 conference organization was supported in part by Novosibirsk State University, and the Institute of Cytology and

Genetics SB RAS, Novosibirsk, Russia. The population genetics studies were supported by the Russian Science Foundation (grant 19-15-00219). The publication has been prepared with the support of the RUDN University Strategic Academic Leadership Program (recipient YO).

ACKNOWLEDGMENTS

The guest editors are grateful to the BGRS conference organizing committee, Prof. N.A. Kolchanov, all the authors contributing to this special issue papers collection and thank all the reviewers who helped improve the manuscripts.

REFERENCES

- Adeel, M. M., Jiang, H., Arega, Y., Cao, K., Lin, D., Cao, C., et al. (2021). Structural Variations of the 3D Genome Architecture in Cervical Cancer Development. *Front. Cel Dev. Biol.* 9, 706375. doi:10.3389/fcell.2021.706375
- Baranova, A. V., Klimontov, V. V., Letyagin, A. Y., and Orlov, Y. L. (2019). Medical Genomics Research at BGRS-2018. *BMC Med. Genomics* 12 (Suppl. 2), 36. doi:10.1186/s12920-019-0480-0
- Chang, Y., Wang, Y., Li, B., Lu, X., Wang, R., Li, H., et al. (2021). Whole-Exome Sequencing on Circulating Tumor Cells Explores Platinum-Drug Resistance Mutations in Advanced Non-small Cell Lung Cancer. *Front. Genet.* 12, 722078. doi:10.3389/fgene.2021.722078
- Chen, M., Harrison, A., Shanahan, H., and Orlov, Y. (2017). Biological Big Bytes: Integrative Analysis of Large Biological Datasets. *J. Integr. Bioinform.* 14, 20170052. doi:10.1515/jib-2017-0052
- Gozman, L., Perry, K., Nikogosov, D., Klabukov, I., Shevlyakov, A., and Baranova, A. (2021). A Role of Variance in Interferon Genes to Disease Severity in COVID-19 Patients. *Front. Genet.* 12, 709388. doi:10.3389/fgene.2021.709388
- Kamenova, S., Aralbayeva, A., Kondybayeva, A., Akimniyazova, A., Pyrkova, A., and Ivashchenko, A. (2021). Evolutionary Changes in the Interaction of miRNA with mRNA of Candidate Genes for Parkinson's Disease. *Front. Genet.* 12, 647288. doi:10.3389/fgene.2021.647288
- Kozlov, K., Singh, A., Berger, J., Bishop-von Wettberg, E., Kahraman, A., Aydogan, A., et al. (2019). Non-linear Regression Models for Time to Flowering in Wild Chickpea Combine Genetic and Climatic Factors. *BMC Plant Biol.* 19 (Suppl. 2), 94. doi:10.1186/s12870-019-1685-2
- Lu, H., Chen, M., Tang, S., and Yu, W. (2021). Association of Coagulation Disturbances with Severity of COVID-19: a Longitudinal Study. *Hematology* 26 (1), 656–662. doi:10.1080/16078454.2021.1968648
- Orlov, Y. L. (2019). 5th International Scientific Conference of “Plant Genetics, Genomics, Bioinformatics, and Biotechnology” (24–29 June 2019, Novosibirsk, Russia). *J. Food Qual. Hazards Control* 6, 41. doi:10.18502/jfqhc.6.1.458
- Orlov, Y. L., Anashkina, A. A., Klimontov, V. V., and Baranova, A. V. (2021a). Medical Genetics, Genomics and Bioinformatics Aid in Understanding Molecular Mechanisms of Human Diseases. *Int. J. Mol. Sci.* 22, 9962. doi:10.3390/ijms22189962
- Orlov, Y. L., Baranova, A. V., and Markel, A. L. (2016). Computational Models in Genetics at BGRS/SB-2016: Introductory Note. *BMC Genet.* 17 (Suppl. 3), 155. doi:10.1186/s12863-016-0465-3
- Orlov, Y. L., Galieva, A. G., Orlova, N. G., Ivanova, E. N., Mozyleva, Y. A., and Anashkina, A. A. (2021b). Reconstruction of Gene Network Associated with Parkinson Disease for Gene Targets Search. *Biomed. Khim* 67 (3), 222–230. Russian. doi:10.18097/PBMC20216703222
- Orlov, Y. L., Hofestädt, R. M., and Kolchanov, N. A. (2015). Introductory Note for BGRS/SB-2014 Special Issue. *J. Bioinform. Comput. Biol.* 13, 1502001. doi:10.1142/S0219720015020011
- Orlov, Y. L., Hofestädt, R., and Tatarinova, T. V. (2019a). Bioinformatics Research at BGRS/SB-2018. *J. Bioinform. Comput. Biol.* 17, 1902001. doi:10.1142/S0219720019020013
- Orlov, Y. L., Salina, E. A., Eslami, G., and Kochetov, A. V. (2019b). Plant Biology Research at BGRS-2018. *BMC Plant Biol.* 19 (Suppl. 1), 56. doi:10.1186/s12870-019-1634-0
- Ramensky, V. E., Ershova, A. I., Zaichenoka, M., Kiseleva, A. V., Zharikova, A. A., Vyatkin, Y. V., et al. (2021). Targeted Sequencing of 242 Clinically Important Genes in the Russian Population from the Ivanovo Region. *Front. Genet.* 12, 1782. doi:10.3389/fgene.2021.709419
- Samarina, L. S., Malyukova, L. S., Efremov, A. M., Simonyan, T. A., Matskiv, A. O., Koninskaya, N. G., et al. (2020). Physiological, Biochemical and Genetic Responses of Caucasian tea (*Camellia sinensis*(L.) Kuntze) Genotypes under Cold and Frost Stress. *PeerJ* 8, e9787. doi:10.7717/peerj.9787
- Shi, X., Ding, K., Zhao, Q., Li, P., Kang, Y., Tan, S., et al. (2021). Suppression of CPSF6 Enhances Apoptosis through Alternative Polyadenylation-Mediated Shortening of the VHL 3'UTR in Gastric Cancer Cells. *Front. Genet.* 12, 707644. doi:10.3389/fgene.2021.707644
- Snezhkina, A. V., Kalinin, D. V., Pavlov, V. S., Lukyanova, E. N., Golovyuk, A. L., Fedorova, M. S., et al. (2020). Immunohistochemistry and Mutation Analysis of SDHx Genes in Carotid Paragangliomas. *Int. J. Mol. Sci.* 21 (18), 6950. doi:10.3390/ijms21186950
- Tatarinova, T. V., Chen, M., and Orlov, Y. L. (2019). Bioinformatics Research at BGRS-2018. *BMC Bioinformatics* 20 (Suppl. 1), 33. doi:10.1186/s12859-018-2566-7
- Zinchenko, R. A., Ginter, E. K., Marakhonov, A. V., Petrova, N. V., Kadyshchev, V. V., Vasilyeva, T. P., et al. (2021). Epidemiology of Rare Hereditary Diseases in the European Part of Russia: Point and Cumulative Prevalence. *Front. Genet.* 12, 678957. doi:10.3389/fgene.2021.678957

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Orlov, Tatarinova, Oparina, Galieva and Baranova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



In silico Prediction of miRNA Interactions With Candidate Atherosclerosis Gene mRNAs

Dina Mukushkina¹, Dana Aisina¹, Anna Pyrkova¹, Alma Ryskulova², Siegfried Labeit³ and Anatoliy Ivashchenko^{1*}

¹ Department of Biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan, ² Department of microbiology, Kazakh Medical University of Continuing Education, Almaty, Kazakhstan, ³ Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
First Moscow State Medical
University, Russia

Reviewed by:

Elvira Galieva,
Novosibirsk State University, Russia
Vladimir N. Uversky,
University of South Florida,
United States

*Correspondence:

Anatoliy Ivashchenko
a.iavashchenko@gmail.com

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 11 September 2020

Accepted: 15 October 2020

Published: 04 November 2020

Citation:

Mukushkina D, Aisina D,
Pyrkova A, Ryskulova A, Labeit S and
Ivashchenko A (2020) *In silico*
Prediction of miRNA Interactions With
Candidate Atherosclerosis Gene
mRNAs. *Front. Genet.* 11:605054.
doi: 10.3389/fgene.2020.605054

The involvement of genes and miRNAs in the development of atherosclerosis is a challenging problem discussed in recent publications. It is necessary to establish which miRNAs affect the expression of candidate genes. We used known candidate atherosclerosis genes to predict associations. The quantitative characteristics of interactions of miRNAs with mRNA candidate genes were determined using the program, which identifies the localization of miRNA binding sites in mRNA, the free energy interaction of miRNA with mRNA. In mRNAs of *GAS6* and *NFE2L2* candidate genes, binding sites of 21 miRNAs and of 15 miRNAs, respectively, were identified. In *IRS2* mRNA binding sites of 25 miRNAs were located in a cluster of 41 nt. In *ADRB3*, *CD36*, *FASLG*, *FLT1*, *PLA2G7*, and *PPARGC1A* mRNAs, clusters of miR-466, ID00436.3p-miR, and ID01030.3p-miR BS were identified. The organization of overlapping miRNA binding sites in clusters led to their compaction and caused competition among the miRNAs. The binding of 53 miRNAs to the mRNAs of 14 candidate genes with free energy interactions greater than -130 kJ/mole was determined. The miR-619-5p was fully complementary to *ADAM17* and *CD36* mRNAs, ID01593.5p-miR to *ANGPTL4* mRNA, ID01935.5p-miR to *NFE2L2*, and miR-5096 to *IL18* mRNA. Associations of miRNAs and candidate atherosclerosis genes are proposed for the early diagnosis of this disease.

Keywords: gene, atherosclerosis, miRNA, mRNA, association, binding sites cluster, marker

INTRODUCTION

The diagnosis, prevention and treatment of atherosclerosis are important tasks of modern medicine (Byrne et al., 2014; Churov et al., 2019; Solly et al., 2019; Shoeibi, 2020). In most cases, polygenic diseases, including atherosclerosis, develop when the expression of different combinations of candidate genes changes. As a result of a review of various publications, it seems that the approaches used to search for markers for the diagnosis and treatment of atherosclerosis, including miRNA, have not yet solved the problem of diagnosing and treating atherosclerosis (Chen et al., 2020; Li et al., 2020; Ryu et al., 2020; Sun et al., 2020; Wang W. et al., 2020; You et al., 2020). In human, it is known about 7000 miRNA and more than 20,000 genes, and it is unknown how many miRNA and genes from them participate in the development of atherosclerosis

(Byrne et al., 2014; Toba et al., 2014; Lu et al., 2018; Liu et al., 2020; Shi et al., 2020). As a rule, in publications several miRNA and some candidate genes are studied, so the combination of such attempts is large (Wang C. et al., 2020; Wang M. et al., 2020; Zhang et al., 2020). Considering the limitations of existing prediction methods, we can say that the establishment of adequate miRNA markers for diseases is unrealistic using existing approaches in the coming years. This statement is confirmed by the fact that over 20 years of studying miRNA participation in various diseases, diagnostic methods and therapeutic ways of treating diseases with miRNA are not used. Most researches study only the correlation between changes in miRNA concentration and expression of candidate genes, but this approach does not establish specific associations of miRNA and target gene. A number of publications argue that miRNA is the cause of disease without understanding that changes in miRNA concentration can occur through independent expression from intergenic regions or from host gene introns (Madrigal-Matute et al., 2013; Andreou et al., 2015; Feinberg and Moore, 2016; Solly et al., 2019). In all cases, miRNAs appear to modify the expression of target genes and not to cause the disease themselves. It is not taken into account that one miRNA can affect several or even 100s of genes, and one gene may be a target of several miRNAs (Atambayeva et al., 2017; Kondybayeva et al., 2018). In this article, we discuss what should be taken into consideration when reviewing the problem of miRNA interaction with candidate genes and show the need to use a systematic approach in establishing the most probable associations of miRNA and target genes.

It has been found that miRNAs, which are nanoscale regulatory biomolecules, are involved in many biological processes at all stages of the development of atherosclerosis, from early endothelial dysfunction to the erosion and rupture of an unstable atherosclerotic plaque (Cipollone et al., 2011; Madrigal-Matute et al., 2013; Andreou et al., 2015; Maitrias et al., 2015; Feinberg and Moore, 2016; Santovito et al., 2016; Chen et al., 2018; Ren et al., 2018; Wang et al., 2018). In addition, miRNAs are considered novel non-invasive biomarkers of the instability of atherosclerotic plaques and have been associated ischemic disorders (Churov et al., 2019). Their detection in the blood of patients may be a promising direction for the diagnosis of atherosclerosis complications such as ischemic stroke and myocardial infarction (Kanuri et al., 2018; Desjarlais et al., 2019; Moghaddam et al., 2019; Vargas-Alarcon et al., 2019; Velle-Forbord et al., 2019; Wiese et al., 2019; Kondybayeva et al., 2020). When the pathogenic role of a specific miRNA is confirmed, it can be considered a potential therapeutic target (Friedman et al., 2009; Fang et al., 2013; Moghaddam et al., 2019; Kazemi et al., 2020). According to the miRBase database, miRNAs have been found in many human tissues and are able to regulate the expression of more than 60% of all protein-coding genes (Kanuri et al., 2018; Ren et al., 2018; Wiese et al., 2019). Full complementarity between miRNA and mRNA results in degradation (Leidinger et al., 2012; Alagia and Eritja, 2016). However, incomplete complementarity is most often observed, in which case miRNAs suppress translation, generally by binding to the 3'-untranslated regions (3' UTRs) of mRNAs (Ivashchenko et al., 2014a,c; Alagia and Eritja, 2016; Atambayeva et al., 2017).

In addition, miRNAs can bind to other regions of target mRNAs, including the 5'-untranslated regions (5' UTR) and coding sequence (CDS) (Orom et al., 2008; Forman and Coller, 2010; Ivashchenko et al., 2013; Zhou and Rigoutsos, 2014; Niyazova et al., 2015; Kondybayeva et al., 2018; Yurikova et al., 2019). We used a program that effectively determines the quantitative characteristics of the interaction of miRNA with mRNA and allows us to identify fundamentally new properties of the binding of miRNA to mRNA. The aim of this work was to identify associations between miRNAs and mRNA candidate genes of atherosclerosis for use as markers for the diagnosis of this disease.

MATERIALS AND METHODS

The nucleotide sequences of 2565 miRNAs (we name this set as 'old miRNAs') were downloaded from the miRBase database¹ (Release 22.1) and 3707 miRNAs (we name this set as 'new miRNAs') obtained from a report by Londin et al. (2015). Due to this work the number of known miRNAs had more than doubled. The nucleotide sequences of genes were obtained from GenBank². A database of 68 candidate genes including the names of the genes and publication sources was compiled, confirming the associations of these genes with atherosclerosis (**Supplementary Table S1**). A search for the target genes of miRNAs was performed using the MirTarget program (Ivashchenko et al., 2014b). This program determines the following binding characteristics: the start of the miRNA binding site (BS) of mRNA; the locations of miRNA BS (3' UTR, 5' UTR, CDS); the interaction free energy (ΔG , kJ/mole); and nucleotide interaction schemes between miRNAs and mRNAs. The ratio of $\Delta G/\Delta G_m$ (%) was determined for each BS, where ΔG_m is equal to the free energy binding of miRNA with its full complementary nucleotide sequence. The MirTarget program found hydrogen bonds between adenine (A) and uracil (U), guanine (G) and cytosine (C), G and U, A and C. The distances between A and C were equal 1.04 nanometers, between G and C, and between A and U were equal 1.03 nanometers, between G and U equal to 1.02 nanometers (Leontis et al., 2002). The numbers of hydrogen bonds in the G-C, A-U, G-U, and A-C interactions were found to be 3, 2, 1 and 1, respectively (Kool, 2001; Lemieux and Major, 2002). The MirTarget program determines single miRNA BS in mRNA and miRNA BS which are in clusters (BS arranged with overlapping of nucleotide sequences of the same or several miRNAs) (Aisina et al., 2019). Predicted by the MirTarget program binding sites in over 30 genes were confirmed experimentally (Yurikova et al., 2019).

RESULTS

The BSs of the miRNAs and mRNAs of the target genes were not uniform along the length of the mRNAs. Both multiple and single BS were identified. BSs could be sequential or overlap with each other. Having overlapping nucleotide sequences in a cluster lead to the compaction of the mRNA

¹<http://mirbase.org>

²<http://www.ncbi.nlm.nih.gov>

sequence that is the target of several miRNAs. When the ΔG and $\Delta G/\Delta G_m$ values of miRNA BS are close to each other, it can be assumed that in the presence of equal miRNA concentrations, the miRNAs with a larger number of BS will be more likely to bind to the mRNAs of target genes. When the miRNA-mRNA interaction strength and the degree of their complementarity are similar, the miRNA has the highest concentration upon binding.

Characteristics of miRNA Interactions in the 5' UTRs of the mRNAs of Atherosclerosis Candidate Genes

BSs of miRNAs were identified in the 5' UTRs of 14 mRNAs. The *GAS6* mRNA contains a large cluster with BS of 21 different miRNAs, half of which have two or three BS, while the rest represent single BS (Table 1). The cluster size is 39 nt, starting at the 11 nt position and ending at position 49 nt. The total BS length in the cluster is 824 nt, where the degree of compaction is 21. This compaction allowed this number of miRNA BS to be located in a 5' UTR with a length of 153 nt. In an association analysis of seven miRNAs with the *GAS6* mRNA the free energy interaction of the miRNAs with the mRNA is more than -130 kJ/mole, which indicates that these associations are promising markers of atherosclerosis.

To prove that the identified clusters are reliable, on the example of *GAS6* gene showed that the cluster exists in the mRNA of orthologous genes and is conservative in its nucleotide composition (Figure 1). There are mainly new miRNAs in the clusters of BSs in 5' UTR region of mRNA of *GAS6*, *NFE2L2* and *SCAP* genes. In mRNA of *GAS6* gene the cluster includes 20 new miRNA and one old miRNA, in mRNA of *NFE2L2* gene the cluster contains 15 BSs of new miRNA; in mRNA of *SCAP* gene there is a cluster for binding of 13 new and one old miRNA (Table 1). In the other 11 target genes, there are 28 new and one old miRNAs are linked in 5' UTR region of mRNA (Supplementary Table S2). These data suggest that the use of new miRNAs significantly increased the number of effective gene expression modulators.

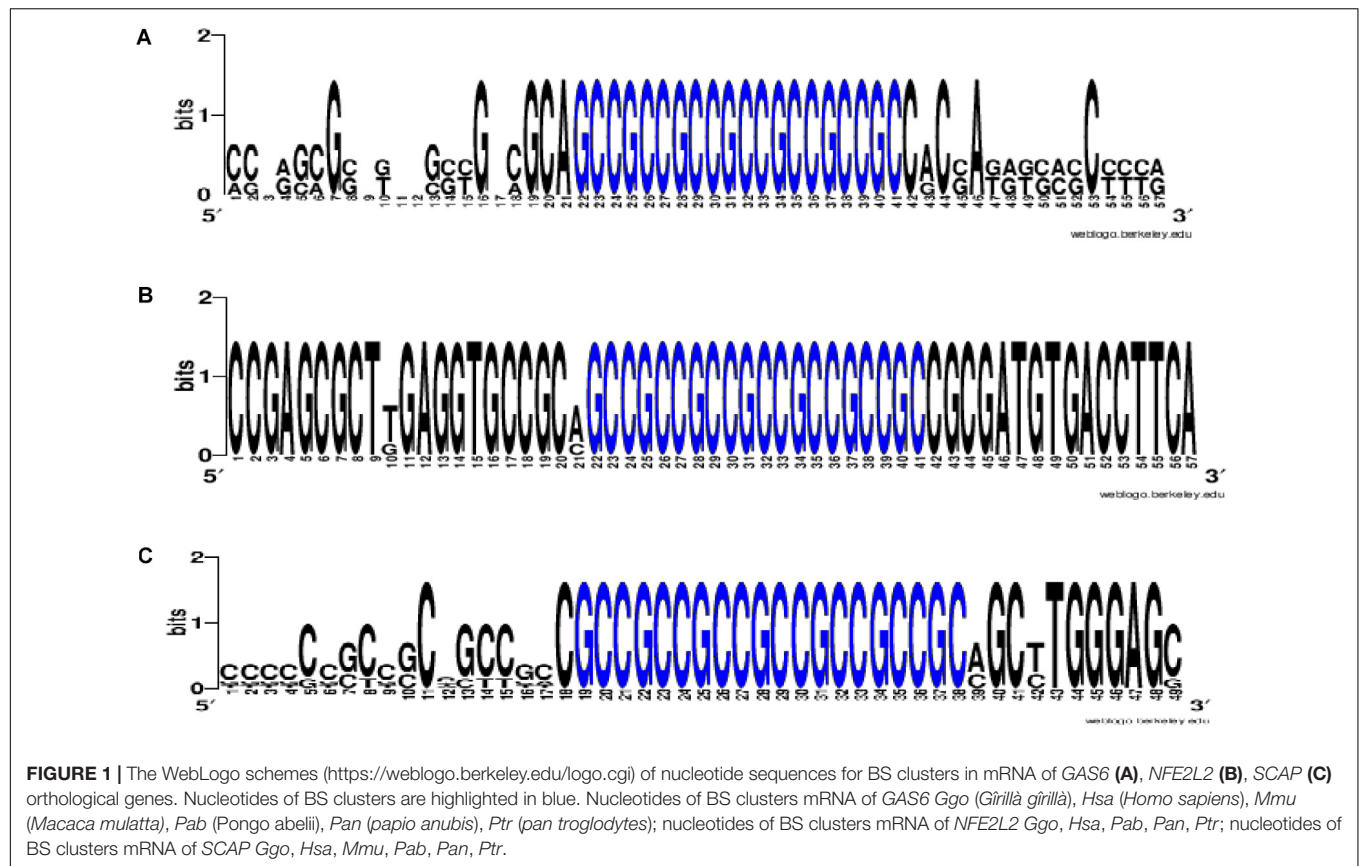
Some miRNAs have multiple BSs: ID00061.3p-miR, ID01155.3p-miR, ID01702.3p-miR, ID01804.3p-miR, ID02187.5p-miR, ID00296.3p-miR, ID01155.3p-miR by three, ID01641.3p-miR by four, ID00061.3p-miR, ID00296.3p-miR, ID01702.3p-miR by five BSs. These miRNAs are recommended as associations for the diagnosis of atherosclerosis considering the high free energy interaction with mRNA of candidate genes. ID01935.5p-miR has a complete complementarity interaction with mRNA of *NFE2L2* gene, which suggests that ID01935.5p-miR and *NFE2L2* association should be recommended as a marker. These miRNA correlations clearly demonstrate the predominant influence of new miRNAs on gene expression with BS in 5' UTR. Consequently, researchers actually do not receive information about the regulation of gene expression by new miRNAs, the effect of which is high when only the old miRNAs are studied.

The following circumstances need to be considered when evaluating the effects of many miRNAs in a cluster of BSs in

TABLE 1 | Characteristics of miRNAs interaction in the 5' UTR of mRNA of *GAS6*, *NFE2L2*, and *SCAP* genes.

miRNA	Start of site, nt	ΔG , kJ/mole	$\Delta G/\Delta G_m$, %	Length, nt
GAS6				
ID00061.3p-miR	17÷23 (3)	-129	94	22
ID00296.3p-miR	17	-144	92	25
ID00457.3p-miR	14÷20 (2)	-123÷-127	91÷94	22
ID00522.5p-miR	17	-125	89	23
ID01041.5p-miR	20	-132	90	24
ID01106.5p-miR	21	-132	89	24
ID01155.3p-miR	17÷23 (3)	-129	94	22
ID01641.3p-miR	17	-134	90	24
ID01702.3p-miR	16÷21 (3)	-136÷-142	86÷91	25
ID01804.3p-miR	11÷20 (3)	-136÷-142	85÷93	25
ID01873.3p-miR	20÷23 (2)	-123÷-125	94÷95	21
ID01879.5p-miR	22	-123	91	22
ID02064.5p-miR	24	-123	94	21
ID02084.3p-miR	22÷25 (2)	-129÷-132	86÷87	24
ID02187.5p-miR	15÷19 (3)	-123	89	23
ID02294.5p-miR	16÷19 (2)	-132÷-136	90÷93	24
ID02538.3p-miR	24	-123	92	22
ID02950.3p-miR	13	-125	89	23
ID03367.5p-miR	17÷23 (2)	-117	93	20
miR-3960	23	-115	92	20
NFE2L2				
ID01935.5p-miR	271	-142	100	24
ID00061.3p-miR	444÷450 (3)	-125÷-134	91÷97	22
ID00296.3p-miR	441÷448 (3)	-134÷-140	85÷89	25
ID00457.3p-miR	444	-123	91	22
ID00522.5p-miR	438	-125	89	23
ID01041.5p-miR	444÷445 (2)	-129÷-134	88÷91	24
ID01155.3p-miR	444÷450 (3)	-125÷-134	91÷97	22
ID01641.3p-miR	444÷448 (2)	-136	91	24
ID01702.3p-miR	444÷450 (3)	-138÷-144	86÷92	25
ID01804.3p-miR	438÷444 (3)	-138÷-144	87÷91	25
ID01873.3p-miR	444÷447 (2)	-121÷-123	92÷94	21
ID02187.5p-miR	439÷445 (2)	-123	89	23
ID02770.5p-miR	462	-115	92	20
ID02890.3p-miR	458	-119	89	23
ID03367.5p-miR	441÷453 (2)	-115	92	20
SCAP				
ID00061.3p-miR	102÷114 (5)	-125÷-132	91÷95	22
ID00296.3p-miR	99÷106 (5)	-140÷-144	89÷92	25
ID00756.3p-miR	105÷106 (2)	-123	89	23
ID01041.5p-miR	108	-129	88	24
ID01403.5p-miR	107	-121	89	23
ID01641.3p-miR	102÷108 (4)	-132÷-134	89÷90	24
ID01652.3p-miR	112	-125	89	23
ID01702.3p-miR	105÷112 (5)	-138÷-144	92÷96	24
ID01804.3p-miR	109	-134	91	23
ID01873.3p-miR	108	-125	95	21
ID02294.5p-miR	101	-129	88	24
ID03151.3p-miR	103	-115	93	20
ID03367.5p-miR	108÷111 (2)	-117	93	20
miR-3960	104÷106 (2)	-117	93	20

In the Table 1 and other Tables in parentheses indicate the number of miRNAs binding sites.



mRNA. Primarily, miRNA will bind with higher free energy interaction, so quantitative interaction characteristics are needed. The ratio of miRNA concentrations between alternative miRNAs has a significant importance (miRNAs, which bind in the cluster) and especially in relation to the concentration of mRNA. Obviously, the total concentration of alternative miRNAs should be lower than the mRNA concentration, otherwise the protein will not be synthesized. A decrease of concentration of any miRNA below mRNA concentration will not affect protein synthesis. At the same time, an increase of one miRNA above mRNA concentration can completely suppress gene expression. The mRNA of many genes contains clusters of BSs of several similar miRNAs. In this way, nature has optimized the expression of several genes under common miRNA control. It is necessary to consider that many genes are expressed in cells of different tissues to different degrees, which may affect the dependence of protein synthesis from miRNA. However, a large number of binding miRNA may not allow the target gene expression to increase significantly. Many miRNAs have been found in the blood and serum in combination with AGO proteins and within exosomes, which indicates their capability to circulate freely throughout the body and to reach many organs and tissues. It is necessary to remember that the synthesis of intronic miRNAs depends on the expression of host gene, while the synthesis of other miRNAs is made from transcripts of intergenic regions. The reasons for changes in the expression of such miRNA in case of disease should be known. There are known

changes of miRNA expression by several orders. Considering the above information, the reports are perceived cautiously that some researchers have been able to detect marker miRNA in specific diseases.

The above examples of presence of BS clusters in mRNA of some genes for several miRNAs suggest that such genes are expressed under the common control of the miRNA group and, consequently, these genes form a network of interconnected genes controlling key metabolic processes. To confirm the reliability of found miRNA and their BSs, it is possible to use data about the presence of such BSs or their clusters in orthological genes (**Figure 1**).

With the given examples we show the inadequacy of miRNA BSs establishment by programs based only on a miRNA seed sequence. The whole miRNA nucleotide sequence is important, which is confirmed by the conservation of miRNA nucleotide sequences and corresponding BS during millions of years. The nucleotide sequences of miRNA and their BSs have been conserved in the mRNA genes of animal and plant organisms over 10s of millions of years of evolution (Bari et al., 2013; Ivashchenko et al., 2013, 2014b; Yurikova et al., 2019).

ID00061.3p-miR, ID00296.3p-miR, ID00457.3p-miR,
ID00522.5p-miR, ID01041.5p-miR, ID01155.3p-miR,
ID01641.3p-miR, ID01702.3p-miR, ID01804.3p-miR,
ID01873.3p-miR, ID02187.5p-miR and ID03367.5p-miR interact
not only with GAS6 mRNA but also with *NFE2L2* mRNA, where

they form a large cluster together with other miRNAs. The cluster of miRNA BS in the *NFE2L2* mRNA is one of the largest clusters of miRNAs associated with atherosclerosis-related genes in which BS are formed in the 5' UTR. The cluster consists of 14 different miRNAs, ranging from the 438 to 482 nt positions, with a size of 45 nt. The total size of all binding sites is 669 nt, resulting in a degree of compaction of 15. Due to compaction, this length of BS is a small fraction of the 5' UTR length of 555 nt. A complete complementarity interaction of ID01935.5p-miR with *NFE2L2* mRNA was identified. Eight miRNAs interact with *NFE2L2* mRNA with a free energy of more than -130 kJ/mole. These associations are recommended as valuable markers of atherosclerosis. The above examples of associations of miRNAs and target genes show that determining the expression level of one or more miRNA without monitoring target gene expression will not provide adequate data on the specific association of these miRNAs with the disease.

The *ADAM10* mRNA contains BS for ID02761.3p-miR with overlapping nucleotide sequences (**Supplementary Table S2**). The *ADCY9* mRNA is characterized by the presence of two clusters, each of which is formed by two single BS. In addition to this gene, clusters were identified in the following mRNAs: *IRS2* and *PIN1*, which contain BS for two miRNAs. The *PNPLA3* and mRNAs have a cluster consisting of three single BS. The *CXCL12* mRNA contains a cluster of four different miRNAs, of which ID02036.3p-miR, ID01293.5p-miR, ID00417.3p-miR and ID02066.5p-miR have two BS. The cluster starts at 65 nt and ends at 93 nt, with a length of 29 nt and an average $\Delta G/\Delta G_m$ of 93%. The sum of the BS lengths of these miRNAs was 170 nt and was six times longer than the cluster length. Due to this compaction, all BS were located in a 5' UTR with a length of 92 nt. The *MMP2* mRNAs contained a cluster of six miRNA BS with a total length of 136 nt and a 5' UTR length of 311 nt. The *PNPLA3* mRNA contains a cluster of three miRNA BS. Among the clusters in the 5' UTRs of the target genes, one can distinguish the *SCAP* gene cluster formed by BS from 14 miRNAs from the 99 nt to the 136 nt position, in which both single BS and multiple BS are found, more specifically, seven multiple sites and seven single sites (**Table 1**). Of the 14 miRNAs with the exception of ID00756.3p-miR, ID01403.5p-miR, ID01652.3p-miR and ID03151.3p-miR all were in the group of miRNAs binding to the *GAS6* mRNA. Furthermore, ID00061.3p-miR, ID00296.3p-miR, and ID01702.3p-miR had five BS each, which significantly increased the likelihood of their binding compared to that of competing miRNAs. The total length of miRNA BS was 735 nt, and the degree of compaction was 19. Without compaction, these sites could not be located in this 255 nt long 5' UTR. This cluster is characterized by a maximum ΔG of -144 kJ/mole for the BS of ID00296.3p-miR and ID01702.3p-miR. Of all the interactions of miRNAs and the 5' UTRs of target mRNAs, five had a free energy of more than -130 kJ/mole. Convincing evidence of the effectiveness of the MirTarget program in determining the characteristics of miRNA binding to the 5' UTR, CDS and 3' UTR of mRNA is given in **Figure 2**. The diagrams show the interaction of nucleotides along the entire length of the miRNA in the BS of the target mRNA and the characteristics of this interaction.

Characteristics of miRNA Interactions With the CDS of mRNAs of Atherosclerosis Candidate Genes

Most genes contained more one BS for one miRNA in the CDS of mRNAs. There were some genes with several BS or multiple BS where clusters could form. It is worth noting that *IRS2* is characterized by the presence of three clusters and a large number of BS compared to other genes. This suggests that this gene is more susceptible to regulation by miRNA (**Table 2**). The first cluster was composed of 25 different miRNAs, 13 of which had two or more BS (ID00061.3p-miR, ID00457.3p-miR, ID00756.3p-miR, ID01155.3p-miR, ID01702.3p-miR, ID01804.3p-miR, ID01873.3p-miR, ID01879.5p-miR, ID02064.5p-miR, ID02187.5p-miR, ID03229.5p-miR, ID03367.5p-miR, miR-3960), and the rest were composed of single BS. The cluster was 41 nt long, starting at 2586 nt and ending at 2626 nt. The total length of the BS in the cluster was 1114 nt, while the degree of cluster compaction was 28. The 2-nd cluster was formed by single BS of seven miRNAs. The 3-day cluster consisted of the BS of nine different miRNAs. ID00296.3p-miR and ID01702.3p-miR had three binding sites. In general, the cluster was 41 nt long, extending from position 4304 to 4344 nt. The total length of all the BS of the three clusters was 1607 nt, which is 40% of the total CDS length of 4017 nt. Due to compaction, the total length of all clusters is 110 nt, that is, only 2.7%. Thirteen miRNAs interacted with the *IRS2* mRNA with a free energy of more than -130 kJ/mole, which gives reason to recommend these interactions as diagnostic markers of atherosclerosis. Each of the three clusters in the MR of the orthologous *IRS2* genes encoded different highly homologous oligopeptides (**Supplementary Table S3**).

Clusters consisting of single miRNA BS were found in the following genes (**Supplementary Table S4**): two BS in *ACE* with an average ΔG of -128 kJ/mole; five BS in *ADRB3* for five miRNAs with an average ΔG of -115 kJ/mole and BS in *FASLG* for ID00061.3p-miR, ID00296.3p-miR, ID01641.3p-miR, and ID01702.3p-miR. *SIRT1* contained two sites for ID03332.3p-miR and single BS for ID00278.3p-miR and ID00811.3p-miR, with an average ΔG of -129 kJ/mole.

TBC1D10B gene is involved in the functioning of cardiovascular endothelial cells (Li et al., 2017) and it is a target of miR-762 with fully complementary binding (**Supplementary Table S4**). This site encodes the APAPAPAPAPAPA oligopeptide, with the flanking oligopeptides AWVPGSAQTS and VTGSTVVVLTLL (**Supplementary Table S5**).

MRNA of *CDKN1C* gene (Rodriguez et al., 2007) contains three clusters consisting of both single and multiple BS. The first cluster consists of two BS for miR-762 and ID03129.3p-miR and has a size of 36 nt, extending from 738 to 753 nt. The second cluster is composed of 17 miR-762 BS, 3 ID00099.3p-miR BS, and 1 ID02682.5p-miR BS. ID00036.3p-miR, ID01075.3p-miR and ID00411.5p-miR cover a region of 31 nt, extending from the 888 nt position to 918 nt with an average ΔG value of -125 kJ/mole. The mRNA of the candidate gene *CDKN1C* has 17 sequentially located binding sites for miR-762 that encode the oligopeptide (AP)18, which indicates a strong

Gene; miRNA; start of binding site, nt; the miRNA region; the free energy, ΔG , kJ/mole; the $\Delta G/\Delta G_m$, %; length of miRNA, nt.	Gene; miRNA; start of binding site, nt; the miRNA region; the free energy, ΔG , kJ/mole; the $\Delta G/\Delta G_m$, %; length of miRNA, nt.
<i>ANGPTL4</i> ; ID01593.5p-miR; 259; CDS; -134; 100; 23 5' -AGCGCUCAGGGCGGACCCGUGCA-3' 3' -UCGCGAGUCCCGCCUGGGCAGCU-5' <i>IGF1</i> ; ID00470.5p-miR; 4054; 3'UTR; -108; 89; 23 5' -ACACACACACACACACACACACAC-3' 3' -UGUGUGUGCGUAUGUGUGCAUGU-5' <i>IRS2</i> ; ID02950.3p-miR; 2586; CDS; -132; 94; 23 5' -AGCCCAGGGCCGCCGCCGCCGCC-3' 3' -UCGGGCGCGCGCGCGCGCGG-5' <i>IRS2</i> ; ID00722.5p-miR; 2589; CDS; -117; 96; 20 5' -CCAGGCCGCCGCCGCCGCC-3' 3' -GGUCUCGGCGCGCGCGCGCA-5' <i>IRS2</i> ; ID02187.5p-miR; 2591; CDS; -123; 89; 23 5' -AGGGCGCGCGCGCGCGCGCGCG-3' 3' -CCUCAUUGGCGCGCGCGCGCG-5' <i>KLF2</i> ; ID01804.3p-miR; 269; CDS; -149; 93; 25 5' -CGGAGCCGCCGCCGCCGCCGCC-3' 3' -GCCCGCGCGCGCGCGCGCGCG-5' <i>NFE2L2</i> ; ID01935.5p-miR; 271; 5'UTR; -142; 100; 24 5' -UCAGCCCCGGCGCGGUGGGGGAU-3' 3' -AGUCGGGGCGCGCCACCCCCUA-5' <i>NOS1AP</i> ; ID00436.3p-miR; 2100; 5'UTR; -104; 89; 23 5' -GUGUGUGUGUACGUGUGUGUGUG-3' 3' -CACACACGCAUAUAUACACACAU-5' <i>NR4A2</i> ; ID00470.5p-miR; 2606; 3'UTR; -110; 91; 23 5' -ACACACACACAUAACACACACACA-3' 3' -UGUGUGUGCGUAUGUGUGCAUGU-5' <i>PDE4D</i> ; ID00296.3p-miR; 411; CDS; -138; 88; 25 5' -CCCUGCCGCCGCCGCCGCCGCC-3' 3' -GGGUGGGGGCGCGCGCGCGCG-5' <i>PPARA</i> ; ID00470.5p-miR; 9032; 3'UTR; -108; 90; 23 5' -ACACACACACACACACACACACA-3' 3' -UGUGUGUGCGUAUGUGUGCAUGU-5'	<i>ICAM1</i> ; ID00436.3p-miR; 2989; 3'UTR; -102; 87; 23 5' -GUGUGUGUGUGUGUGUAUGUGUA-3' 3' -CACACACGCAUAUAUACACACAU-5' <i>IRS2</i> ; ID02240.3p-miR; 1885; CDS -115; 90; 20 5' -CUCCUACCCGCCGCCGCCCG-3' 3' -GGGGCGGGCGCGCGCGGGC-5' <i>IRS2</i> ; ID03221.5p-miR; 2587; CDS; -121; 98; 20 5' -GCCCAGGGCCGCCGCCGCCCG-3' 3' -CGGGUCUCGGCGCGCGCGGC-5' <i>IRS2</i> ; ID01804.3p-miR; 2590; CDS; -155; 97; 25 5' -CAGGGCCGCCGCCGCCGCCGCC-3' 3' -GCCCCGGCGCGCGCGCGCGCG-5' <i>KLF2</i> ; ID00722.5p-miR; 268; CDS; -117; 96; 20 5' -CCGGAGCCGCCGCCGCCGCC-3' 3' -GGUCUCGGCGCGCGCGCGCA-5' <i>LDLR</i> ; ID00376.5p-miR; 2452; CDS; -108; 96; 20 5' -CACCCUGGGCUCACCACGG-3' 3' -GUGGGGACCUAGAGUGGUACC-5' <i>NLRP3</i> ; ID00662.3p-miR; 3638; CDS; -102; 92; 20 5' -UCUGACCUCAGCCAGAGCC-3' 3' -AGACUGGAGGUCAGUCCAG-5' <i>NOS1AP</i> ; ID02513.5p-miR; 4598; 3'UTR; -102; 91; 22 5' -GGUGGUGUGUGUGUGUGUUU-3' 3' -UUACCGUACACACACACACAAA-5' <i>OLRI</i> ; ID00470.5p-miR; 1506; 3'UTR; -108; 89; 23 5' -ACACACACACACACACACACA-3' 3' -UGUGUGUGCGUAUGUGUGCAUGU-5' <i>PPARA</i> ; ID00470.5p-miR; 9026; 3'UTR; -108; 89; 23 5' -ACACACACACACACACACACA-3' 3' -UGUGUGUGCGUAUGUGUGCAUGU-5' <i>TGFBI</i> ; ID01382.3p-miR; 2062; 3'UTR; -113; 93; 20 5' -GCCCCGCCCGCCCGCCCC-3' 3' -CGGGGCGGGGAAGAGCGGG-5'

FIGURE 2 | Convincing evidence of the effectiveness of the MirTarget program in determining the characteristics of miRNA binding to the 5' UTR, CDS and 3' UTR of mRNA. The diagrams show the interaction of nucleotides along the entire length of the miRNA in the BS of the target mRNA and the characteristics of this interaction. Non-canonical base pairs are in bold.

dependence of gene expression on this miRNA (**Supplementary Table S4**). Orthologous primate genes (**Supplementary Table S6**) encode similar oligopeptides with an AP dipeptide number up to 33. In all primate cases, miR-762 BSs are located between the conserved flanking nucleotide sequences that encode the

conserved AAPVAVAVLA and DAAPQESAEQ oligopeptides (**Supplementary Table S6**). These animals can be used to study the role of miR-762 in the development of atherosclerosis. A similar type of organization of miRNA binding multiple sites has been found in the mRNAs of many genes involved in the

development of cardiovascular and other diseases (Kondybayeva et al., 2018; Aisina et al., 2019; Yurikova et al., 2019).

KLF2 is characterized by the presence of two clusters consisting of the BS for 14 and 2 different miRNAs. The first of these clusters is formed by single BS starting from the 264 to 296 nt position with a size of 33 nt (Table 2). The total length of all BS is 305 nt. The degree of compaction of the binding sites in the cluster is nine. The second cluster of this gene is formed only by two single BS. In the CDS of *KLF2* mRNA BS were detected for ID00061.3p-miR, ID00457.3p-miR, ID01155.3p-miR, ID01702.3p-miR, ID01804.3p-miR, ID02187.5p-miR, ID03367.5p-miR which have BS in the 5' UTR of *GAS6* mRNA. This indicates an overall control of these genes expression by these miRNAs. The maximum $\Delta G/\Delta G_m$ value (100%) among these genes was detected at the ID01593.5p-miR BS, which interact with the *ANGPTL4* mRNA.

PDE4D formed two clusters. The first cluster contained the BS for ID00061.3p-miR, ID01155.3p-miR, ID03064.3p-miR, miR-3960 and ID01702.3p-miR, which had three BS with a free energy value of -134 kJ/mole. ID01641.3p-miR, miR-3960 and ID01702.3p-miR BS were found in the second cluster. The size of the first cluster was 34 nt, extending from position 336 to 369 nt. The second cluster consisted of five multiple BS and eight single BS; the cluster size was 49 nt, extending from 391 to 439 nt. The total length of miRNA BS was 491 nt, and the degree of compaction was 10 (Supplementary Table S4).

There were 39 BS for new miRNAs and two BS for old miRNA in *IRS2* mRNA. Two clusters of binding sites in the *KLF2* mRNA interacted with 14 new miRNAs and one old miRNA. In addition, there were 15 BS for new miRNA and two BS for miR-3960 in the *PDE4D* mRNA. A total in CDS mRNA of 40 target genes BS were identified for 148 new miRNAs and 32 old miRNAs.

Some miRNAs had more than two BS: ID01641.3p-miR, ID02064.5p-miR, ID02187.5p-miR, ID02770.5p-miR, ID03324.3p-miR, ID03367.5p-miR – 3 BS; ID00296.3p-miR, ID01377.3p-miR, ID01804.3p-miR, miR-3960 – 4 BS; ID00457.3p-miR, ID01155.3p-miR – 5 BS and ID00061.3p-miR, ID01702.3p-miR – 6 BS. Most of them had free energy interaction above -130 kJ/mole.

Three clusters of miRNA BS in CDS mRNA of orthologic genes encode different oligopeptides because they have different reading frames. Oligopeptides encoded by BS are clearly defined by conservative oligopeptides, which flank oligopeptides encoded by BS. Apparently, flanking oligopeptides carry out the important functional role of a protein, and oligopeptides coded by BS in the first and second clusters change for the reason of an acceptability of incomplete complementary interactions of miRNA and mRNA.

Characteristics of miRNA Interactions in the 3' UTRs of mRNAs of Atherosclerosis Candidate Genes

The 40 genes involved in the development of atherosclerosis were characterized by BS in the 3' UTRs. A feature of miRNA interactions with atherosclerosis candidate genes is the presence of BS for miR-466, ID00436.3p-miR, and ID01030.3p-miR in

TABLE 2 | Characteristics of miRNAs interaction in the CDS of mRNA of *IRS2* and *KLF2* genes.

miRNA	Start of site, nt	ΔG , kJ/mole	$\Delta G/\Delta G_m$, %	Length, nt
<i>IRS2</i>				
ID00061.3p-miR	2589÷2601 (4)	$-125\div-136$	91÷98	22
ID00296.3p-miR	2589	-136	86	25
ID00457.3p-miR	2592÷2595 (2)	$-123\div-132$	91÷97	22
ID00522.5p-miR	2589	-127	91	23
ID00756.3p-miR	2587÷2593 (2)	$-123\div-125$	89÷91	23
ID01041.5p-miR	2592	-138	94	24
ID01155.3p-miR	2595÷2601 (4)	$-125\div-136$	91÷98	22
ID01574.5p-miR	2591	-127	90	23
ID01702.3p-miR	2589÷2599 (4)	$-136\div-140$	90÷93	25
ID01778.3p-miR	2594	-134	90	24
ID01804.3p-miR	2589÷2596 (5)	$-140\div-155$	88÷97	25
ID01873.3p-miR	2592÷2595 (2)	-123	94	21
ID01879.5p-miR	2594÷2600 (2)	-123	91	22
ID01895.5p-miR	2586	-134	90	24
ID02052.5p-miR	2589	-132	89	24
ID02064.5p-miR	2594÷2603 (4)	$-129\div-132$	90÷91	23
ID02187.5p-miR	2591÷2593 (2)	$-123\div-125$	89÷91	23
ID02294.5p-miR	2594	-134	91	24
ID02950.3p-miR	2586	-132	94	23
ID03064.3p-miR	2589	-140	92	24
ID03221.5p-miR	2587	-121	98	20
ID03229.5p-miR	2590÷2593 (2)	-121	90	22
ID03305.5p-miR	2589	-115	95	20
ID03367.5p-miR	2592÷2598 (3)	-117	93	20
miR-3960	2594÷2597 (2)	-115	92	20
ID01702.5p-miR	3602	-138	88	25
ID01574.5p-miR	3603	-127	90	23
ID01804.3p-miR	3599	-140	88	25
ID01106.5p-miR	3603	-129	87	24
ID02229.3p-miR	3605	-125	95	21
ID02499.3p-miR	3605	-121	93	21
miR-3960	3604	-115	92	20
ID00061.3p-miR	4316	-125	91	22
ID00296.3p-miR	4304÷4310 (3)	$-134\div-144$	85÷92	25
ID00457.3p-miR	4319	-123	91	22
ID01155.3p-miR	4316	-125	91	22
ID01641.3p-miR	4310	-134	90	24
ID01702.3p-miR	4310÷4316 (3)	$-134\div-136$	89÷90	25
ID01804.3p-miR	4313	-138	87	25
ID02064.5p-miR	4321	-129	90	23
ID02187.5p-miR	4314	-125	91	23
<i>KLF2</i>				
ID00061.3p-miR	274	-129	94	22
ID00457.3p-miR	271	-127	94	22
ID00722.5p-miR	268	-117	96	20
ID01155.3p-miR	274	-129	94	22
ID01377.3p-miR	273	-121	95	20
ID01445.3p-miR	265	-115	92	20
ID01702.3p-miR	266	-136	90	24
ID01804.3p-miR	269	-149	93	23
ID01895.5p-miR	264	-140	94	24
ID02187.5p-miR	270	-123	89	23
ID02950.3p-miR	265	-134	95	23
ID03221.5p-miR	266	-121	98	20
ID03367.5p-miR	271	-115	92	20
miR-4787-5p	269	-123	92	22

their mRNAs (Table 3 and Supplementary Table S7). The BS of these miRNAs can be single or multiple and could form a cluster. For example, *CD36* mRNA contains six BS for ID00436.3p-miR and five BS for miR-466 and ID01030.3p-miR. *FASLG* mRNA contains six BS for miR-466 and ID00436.3p-miR and five BS for ID01030.3p-miR. The mRNA of the *FLT1* gene contains eight BS for miR-466 and seven BS for ID00436.3p-miR and ID01030.3p-miR. Another feature of candidate genes is the presence of clusters of binding sites for miR-5095 and miR-619-5p in their mRNA. There are two clusters of BS of miR-5095 and miR-619-5p located 12 nucleotides apart in *IL18* mRNA. In these clusters, the start sites of miR-5095 and miR-619-5p BS are located six nucleotides apart, which is probably due to their common origin. In the mRNAs of *CD36* and *ADAM17* completely complementary BS with miR-619-5p and BS between the mRNA of *IL18* with miR-5096 were revealed. The BS of miR-1273f and miR-1273e form a cluster in the mRNAs of *IGF1* gene (Supplementary Table S7).

The 3' UTRs of the mRNAs of many candidate genes contain only single miRNA BS that do not form clusters: *ABO*, *ACE*, *ADAM17*, *ADAM33*, *ANGPT2*, *APOL1*, *CD59*, *FADS2*, *FOXP3*, *GPR132*, *HNF1A*, *IGF1R*, *LDLR*, *LPCAT3*, *NOS1AP*, *PNPLA3*, *PPARA*, *SOAT1*, *SOCS3*, *TFPI*, *TNC*, and *ZBTB46*. Some genes are characterized by clusters consisting of two single BS: *BRCA1*, *F11R*, *IGF1*, *ITGA2*, and *TNFSF10*. Some genes were characterized by mixed clusters formed by both single and multiple BS. *NR4A2* had a cluster consisting of both single BS of ID00470.5p-miR and multiple BS of ID02299.5p-miR (Supplementary Table S7).

FASLG was characterized by a cluster formed from a miRNA with single BS (ID02868.3p-miR) and three miRNAs with multiple BS consisting of six BS for miR-466, five for ID01030.3p-miR and six for ID00436.3p-miR, extending from 1602 to 1637 nt with an average ΔG value of -108 kJ/mole. In the 3' UTR of *FLT1*, there was a cluster of three multiple BS consisting of seven BS for ID01030.3p-miR, seven for ID00436.3p-miR and

TABLE 3 | Characteristics of miRNAs interaction in the 3' UTR of mRNA of atherosclerosis candidate genes.

Gene	miRNA	Start of site, nt	ΔG , kJ /mole	$\Delta G/\Delta G_m$, %	Length, nt
<i>ADRB3</i>	ID02868.3p-miR	2442	-115	92	23
	miR-466	2451	-110	95	23
	ID00436.3p-miR	2456	-108	93	23
	ID01030.3p-miR	2456÷2462 (2)	-113	93	23
<i>CD36</i>	miR-466	3530÷3538 (5)	-106÷-108	91÷93	23
	ID00436.3p-miR	3533÷3543 (6)	-104	89	23
	ID01030.3p-miR	3529÷3539 (5)	-108	89	23
	ID01727.5p-miR	3542	-106	91	23
	miR-619-5p	4041	-121	100	22
	miR-5096	4107	-104	92	21
	miR-619-5p	4168	-117	96	22
	miR-5585-3p	4175	-108	93	22
<i>FASLG</i>	ID02868.3p-miR	1602	-113	90	23
	miR-466	1603÷1613 (6)	-106÷-108	91÷93	23
	ID00436.3p-miR	1604÷1614 (6)	-104÷-106	89÷91	23
	ID01030.3p-miR	1604÷1612 (5)	-108	89	23
<i>FLT1</i>	miR-466	6910÷6936 (8)	-106÷-108	91÷93	23
	ID00436.3p-miR	6913÷6925 (7)	-104	89	23
	ID01030.3p-miR	6909÷6923 (7)	-108÷-110	89÷91	23
<i>ICAM1</i>	miR-466	2988	-106	91	23
	ID01030.3p-miR	2987	-108	89	23
	ID01360.3p-miR	3022	-104	91	21
	ID00367.5p-miR	3025	-110	90	22
<i>IL18</i>	miR-1273g-3p	3031	-115	98	21
	miR-5095	811÷823 (2)	-110	95	21
	miR-619-5p	817÷829 (2)	-119		22
<i>PLA2G7</i>	miR-5096	890÷903 (4)	-102÷-113	91÷100	21
	miR-466	1643÷1651 (5)	-106÷-108	91÷93	23
	ID00436.3p-miR	1646÷1654 (5)	-104	89	23
<i>PPARGC1A</i>	ID01030.3p-miR	1646÷1652 (4)	-108	89	23
	miR-466	2806÷2822 (2)	-106	91	23
	ID00436.3p-miR	2809÷2825 (3)	-104÷-108	89÷93	23
	ID01030.3p-miR	2811	-115	95	23
	ID01727.5p-miR	2824	-104	89	23

eight for miR-466, extending from 6909 to 6959 nt and with an average ΔG value of -107 kJ/mole. *IL18* had a cluster of two miR-5095 and two miR-619-5p sequences, with a size of 41 nt and an average ΔG value of -114 kJ/mole. *MTHFR* mRNA contained one cluster, consisting of ID01811.5p-miR, miR-5095 and miR-619-5p with an average ΔG value of -114 kJ/mole. *PPARGC1A* contained a cluster formed by four different miRNAs, including two miR-466 BS and three ID00436.3p-miR, ID01030.3p-miR, and ID01727.5p-miR sequences. The size of this cluster was 43 nt, extending from 2806 to 2848 nt. There were some genes with more than one cluster. The *PLA2G7* mRNA had a cluster that consisted of five miR-466, four ID01030.3p-miR and five ID00436.3p-miR sites and a length of 35 nt, extending from 1643 to 1677 nt, with an average ΔG value of -106 kJ/mole.

ICAM1 was characterized by the presence of two clusters, the first of which consisted of two single BS (ID01030.3p-miR and miR-466), extending from 2987 to 3011 nt. The second cluster consisted of three BS of different miRNAs. ID00470.5p-miR formed BS in *OLR1* and *NR4A2* mRNAs. The cluster of miR-1273f and miR-1273e BS was formed in *IGF1* mRNA.

There were miRNAs that had BS for more than one mRNA. The unique miR-619-5p had BS in the mRNAs of *ADAM17*, *ADAM33*, *APOL1*, *BRCA1*, *F11R*, *IL18*, *ITGA2*, *LDLR*, *MTHFR*, *PNPLA3*, *SOAT1*, and *TNFSF10*. ID00436.3p-miR and ID01030.3p-miR formed clusters that consisted of multiple BS for *ADRB3*, *F11R*, *FASLG*, *FLT1*, *PLA2G7*, and *PPARGC1A* mRNA. ID01727.5p-miR had BS for *NOS1AP* and *PPARGC1A* mRNAs.

The *CD36* mRNA cluster, consisting of BS to two different miRNAs, miR-619-5p and miR-5585-3p, extended from 4168 to 4197 nt. The *CD36* mRNA cluster consisted of both single BS and multiple BS. In general, the cluster consisted of five ID01030.3p-miR, five miR-466, six ID00436.3p-miR and one ID01727.5p-miR sequences. The size was 38 nt, extending from 3529 to 3566 nt, with an average ΔG value of -106 kJ/mole.

Associations of miRNA and target genes having BS in 3' UTR mRNA are very different from those in 5' UTR and CDS. In the 3' UTR mRNA there are no clusters of BS over four miRNAs. The number of old and new miRNAs is comparable. The total number of binding sites of new miRNA is 57 and 63 old miRNA per mRNA of 40 target genes. This is probably due to the fact that the GC content of old miRNAs is comparable to that of 3' UTRs, while in 5' UTR and CDS it is about 10% higher. In the new miRNA GC content is comparable to that in 5' UTR and CDS, so the number of BS for new miRNA in these regions of mRNA is much larger.

The choice of miRNA-gene associations is difficult to offer as miRNAs such as miR-619-5p, miR-5095-, miR-5096- and miR-5585-3p, which have targets in many genes and most likely play the role of stabilizers of protein expression. Similar to miR-466, ID00436.3p-miR and ID01030.3p-miR, which have multiple BS in the mRNA of many genes, which also gives them the role of stabilizing the expression of their targets. A significant increase in the concentration of any these miRNAs will lead to numerous metabolic disorders and to different diseases consequently. If about 100–200 key miRNAs are identified with miRNA-chips, the range of candidate genes will be significantly

reduced because with bioinformatic approaches it could be done easily and quickly.

DISCUSSION

Since it would be advisable to have lower costs for the diagnosis of the disease, it is desirable to minimize the number of associations of miRNA and target genes. However, in the case of polygenic diseases, such methods could be difficult to develop because it is not known which gene and which miRNA cause a particular patient's disease. Therefore, now it is necessary to select a list of priority genes for diagnosis and then check which miRNA could influence protein expression. With the MirTarget program, this is easily to install and, if alternative miRNAs are involved, it is also simple to identify miRNAs competitiveness. The results of the present research demonstrate the effectiveness of this approach.

It was found that the BS of some miRNAs formed clusters in the 5' UTR and CDS sequences of mRNA, resulting in competition between these miRNAs for binding to mRNA and, accordingly, for suppressing the expression of the target gene. The quantitative characteristics of miRNA BS with mRNA make it possible to predict which miRNAs can more efficiently bind to mRNA at equal miRNA concentrations. At different concentrations of these miRNAs, kinetic equations can be used to predict their suppressive effect on the expression of the target gene. The revealed associations of several miRNAs that bind to mRNAs of different candidate genes make it possible to predict the effect of these miRNAs on the corresponding genes, which can be expressed to different degrees. For example, *GAS6*, *NFE2L2*, and *IRS2* are targets of several identical miRNAs and bind with these miRNAs depending on their concentrations in the cell. In cells of different tissues, the ratio of miRNA and target gene concentrations will be different, and the effect of the miRNAs on these target genes will also be different. This brief discussion of the different interactions of miRNAs and candidate genes under different circumstances shows the complexity of these networks and begins to paint a picture of the interactions of miRNA groups with groups of target genes. Without this assumption, it is difficult to predict or evaluate the involvement of a single miRNA or a single gene in the development of a disease. From the identified interactions, it is necessary to choose those that control more genes and miRNAs and reflect the development of atherosclerosis. A number of mRNAs and miRNAs strongly interacting with each other are proposed in this study. For example, *PDE4D* and ID00296.3p-miR; *SCAP* and ID00296.3p-miR and ID01702.3p-miR; *ADCY9* and ID00296.3p-miR; *IRS2* and ID01702.3p-miR, ID03064.3p-miR, ID01804.3p-miR, and ID00296.3p-miR; *NFE2L2* and ID01935.5p-miR, ID00296.3p-miR, ID01702.3p-miR, and ID01804.3p-miR. In addition, miRNAs that bind with mRNAs of candidate genes with complete complementarity are of interest. Within addition to this condition, it is necessary to consider miRNAs interacting with a large number of genes, since such miRNAs can simultaneously control the expression of many genes. Genes that are regulated by a large number of miRNAs can also be

considered as effective markers. Unfortunately, our knowledge is still insufficient to determine which candidate genes play a key role in the development of the disease, and it is not known which miRNAs are the most likely signals for the development of atherosclerosis. However, one thing is obvious – it is necessary to simultaneously study miRNA and gene interactions in order to improve the likelihood of obtaining a valid result. From the established interactions, it is first necessary to consider the interactions between candidate genes with multiple miRNAs that form clusters of BS and miRNA interactions with high free energy. For example, *IRS2* is the target of 41 miRNAs with BS located in the CDS, of which ID00061.3p-miR, ID00296.3p-miR, ID00457.3p-miR, ID01041.5p-miR, ID01155.3p-miR, ID01641.3p-miR, ID01702.3p-miR, ID01778.3p-miR, ID01804.3p-miR, ID01895.5p-miR, ID02052.5p-miR, ID02064.5p-miR, ID02950.3p-miR, and ID03064.3p-miR have free energy interactions above -130 kJ/mole. The mRNA of *GAS6* interacts with 21 miRNAs, of which those with ID00296.3p-miR, ID01041.5p-miR, ID01106.5p-miR, ID01641.3p-miR, ID01702.3p-miR, ID01804.3p-miR, ID02084.3p-miR, and ID02294.5p-miR are more efficient. Associations of this type were observed for *NFE2L2*, *ADCY9*, *PNPLA3*, *SCAP*, *KLF2* and *PDE4D*, which bind to many miRNAs with high free energy. Another type of association is the interaction of one or more miRNA with many genes. For example, miR-466, ID00436.3p-miR and ID01030.3p-miR bind with the mRNAs of *ADRB3*, *CD36*, *FASLG*, *FLT1*, *PLA2G7*, and *PPARGC1A*. In addition, miR-619-5p targets several candidate atherosclerosis genes: *CD36*, *IL18*, *ADAM17*, *ADAM33*, *APOL1*, *BRCA1*, *F11R*, *ITGA2*, *LDLR*, *MTHFR*, *PNPLA3*, *SOAT1*, and *TNFSF10*. The mRNA of *ANGPTL4* is fully complementary with ID01593.5p-miR; the mRNAs of *ADAM17* and *CD36* with miR-619-5p; the mRNA of *IL18* with miR-5096; and the mRNA of *NFE2L2* with ID01935.5p-miR.

Here, miRNA interactions with candidate atherosclerosis genes have been established, which consist of one gene and several miRNAs, one gene and one miRNA, one miRNA and several genes, and two or more miRNAs with two or more

candidate genes. The revealed cluster organization of miRNA BS in mRNA candidate genes contributes to a more accurate diagnosis of the participation of competing miRNAs in the development of atherosclerosis.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

AI and DM conceived of the study, drafted the manuscript, and gave final approval of the version to be published. SL, DA, AP, and AR conceived of the study and drafted the manuscript. All authors made substantial contributions to acquisition of data, to interpretation and modification of the data, were involved in subsequent rounds of revisions, and read and approved the final manuscript.

FUNDING

The work was carried out with the financial support of the Ministry of Education and Science of the Republic of Kazakhstan within the framework of the grant no. AP05132460. We thank the European Union Horizon 2020 research and innovation program (grant agreement no. 645648) ‘Muscle Stress Relief.’

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.605054/full#supplementary-material>

REFERENCES

- Aisina, D., Niyazova, R., Atambayeva, S., and Ivashchenko, A. (2019). Prediction of clusters of miRNA binding sites in mRNA candidate genes of breast cancer subtypes. *PeerJ*. 7:e8049. doi: 10.7717/peerj.8049
- Alagia, A., and Ertija, R. (2016). siRNA and RNAi optimization. *Wiley Interdiscip. Rev. RNA*. 7, 316–329. doi: 10.1002/wrna.1337
- Andreou, I., Sun, X., Stone, P. H., Edelman, E. R., and Feinberg, M. W. (2015). MiRNAs in atherosclerotic plaque initiation, progression, and rupture. *Trends Mol. Med.* 21, 307–318. doi: 10.1016/j.molmed.2015.02.003
- Atambayeva, S., Niyazova, R., Ivashchenko, A., Pyrkova, A., Pinsky, I., Akimniyazova, A., et al. (2017). The binding sites of miR-619-5p in the mRNAs of human and orthologous genes. *BMC Genomics* 18:428. doi: 10.1186/s12864-017-3811-6
- Bari, A., Orazova, S., and Ivashchenko, A. (2013). miR156- and miR171-binding sites in the protein-coding sequences of several plant genes. *BioMed Res. Int.* 2013:307145. doi: 10.1155/2013/307145
- Byrne, M. M., Murphy, R. T., and Ryan, A. W. (2014). Epigenetic modulation in the treatment of atherosclerotic disease. *Front. Genet.* 5:364. doi: 10.3389/fgene.2014.00364
- Chen, W., Yu, F., Di, M., Li, M., Chen, Y., Zhang, Y., et al. (2018). MicroRNA-124-3p inhibits collagen synthesis in atherosclerotic plaques by targeting prolyl 4-hydroxylase subunit alpha-1 (P4HA1) in vascular smooth muscle cells. *Atherosclerosis* 277, 98–107. doi: 10.1016/j.atherosclerosis.2018.08.034
- Chen, F., Ye, X., Jiang, H., Zhu, G., and Miao, S. J. (2020). MicroRNA-151 attenuates apoptosis of endothelial cells induced by oxidized low-density lipoprotein by targeting interleukin-17A (IL-17A). *Cardiovasc. Transl. Res.* doi: 10.1007/s12265-020-10065-w
- Churov, A., Summerhill, V., Grechko, A., Orekhova, V., and Orekhov, A. (2019). MicroRNAs as potential biomarkers in atherosclerosis. *Int. J. Mol. Sci.* 20:E5547. doi: 10.3390/ijms20225547
- Cipollone, F., Felicioni, L., Sarzani, R., Uchino, S., Spigonardo, F., Mandolini, C., et al. (2011). A unique microRNA signature associated with plaque instability in humans. *Stroke* 42, 2556–2563. doi: 10.1161/STROKEAHA.110.597575

- Desjarlais, M., Dussault, S., Rivard, F., Harel, S., Sanchez, V., Hussain, S. N. A., et al. (2019). Forced expression of microRNA-146b reduces TRAF6-dependent inflammation and improves ischemia-induced neovascularization in hypercholesterolemic conditions. *Atherosclerosis* 289, 73–84. doi: 10.1016/j.atherosclerosis.2019.08.010
- Fang, Z., Du, R., Edwards, A., Flemington, E. K., and Zhang, K. (2013). The sequence structures of human microRNA molecules and their implications. *PLoS One* 8:e54215. doi: 10.1371/journal.pone.0054215
- Feinberg, M. W., and Moore, K. J. (2016). MicroRNA regulation of atherosclerosis. *Circ. Res.* 118, 703–720. doi: 10.1161/CIRCRESAHA.115.306300
- Forman, J. J., and Collier, H. A. (2010). The code within the code: microRNAs target coding regions. *Cell Cycle* 9, 1533–1541. doi: 10.4161/cc.9.8.11202
- Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* 19, 92–105. doi: 10.1101/gr.082701.108
- Ivashchenko, A., Berillo, O., Pyrkova, A., and Niyazova, R. (2014a). Binding sites of miR-1273 family on the mRNA of target genes. *Biomed. Res. Int.* 2014:e11. doi: 10.1155/2014/620530
- Ivashchenko, A., Berillo, O., Pyrkova, A., Niyazova, R., and Atambayeva, S. (2014b). MiR-3960 binding sites with mRNA of human genes. *Bioinformatics* 10, 423–427. doi: 10.6026/97320630010423
- Ivashchenko, A., Berillo, O., Pyrkova, A., Niyazova, R., and Atambayeva, S. (2014c). The properties of binding sites of miR-619-5p, miR-5095, miR-5096 and miR-5585-3p in the mRNAs of human genes. *Biomed. Res. Int.* 2014:e8. doi: 10.1155/2014/720715
- Ivashchenko, A., Issabekova, A. S., and Berillo, O. A. (2013). miR-1279, miR-548j, miR-548m, and miR-548d-5p binding sites in CDSs of paralogous and orthologous PTPN12, MSH6, and ZEB1 Genes. *Biomed. Res. Int.* 2013:e10. doi: 10.1155/2013/902467
- Kanuri, S. H., Ipe, J., Kassab, K., Gao, H., Liu, Y., Skaar, T. C., et al. (2018). Next generation MicroRNA sequencing to identify coronary artery disease patients at risk of recurrent myocardial infarction. *Atherosclerosis* 278, 232–239. doi: 10.1016/j.atherosclerosis.2018.09.021
- Kazemi, F. T., Tavakoli, S., Ahmadi, R., Moradi, N., Fadaei, R., Mohammadi, A., et al. (2020). Evaluation of IP10 and miRNA 296-a expression levels in peripheral blood mononuclear cell of coronary artery disease patients and controls. *DNA Cell Biol.* 39, 1678–1684. doi: 10.1089/dna.2020.5650
- Kondybayeva, A., Akimniyazova, A., Kamenova, S., Duchshanova, G., Aisina, D., Goncharova, A., et al. (2020). Prediction of miRNA interaction with mRNA of stroke candidate genes. *Neurol. Sci.* 41, 799–808. doi: 10.1007/s10072-019-04158-x
- Kondybayeva, A. M., Akimniyazova, A. N., Kamenova, S. U., and Ivashchenko, A. T. (2018). The characteristics of miRNA binding sites in mRNA of ZFXH3 gene and its orthologs. *Vavilov J. Genet. Breed.* 22, 438–444. doi: 10.18699/VJ18.380
- Kool, E. T. (2001). Hydrogen bonding, base stacking, and steric effects in DNA replication. *Annu. Rev. Biophys. Biomol. Struct.* 30, 1–22. doi: 10.1146/annurev.biophys.30.1.1
- Leidinger, P., Keller, A., and Meese, E. (2012). MicroRNAs – important molecules in lung cancer research. *Front. Genet.* 2:104. doi: 10.3389/fgene.2011.00104
- Lemieux, S., and Major, F. (2002). RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.* 30, 4250–4263. doi: 10.1093/nar/gkf540
- Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-watson-crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* 30, 3497–3531. doi: 10.1093/nar/gkf481
- Li, Q., Cui, H. H., Yang, Y. J., Li, X. D., Chen, G. H., Tian, X. Q., et al. (2017). Quantitative proteomics analysis of ischemia/reperfusion injury-modulated proteins in cardiac microvascular endothelial cells and the protective role of tongxinluo. *Cell Physiol. Biochem.* 41, 1503–1518. doi: 10.1159/000470806
- Li, X., He, X., Wang, J., Wang, D., Cong, P., Zhu, A., et al. (2020). The Regulation of Exosome-Derived miRNA on Heterogeneity of Macrophages in Atherosclerotic Plaques. *Front. Immunol.* 11:2175. doi: 10.3389/fimmu.2020.02175
- Liu, D., Song, J., Ji, X., Liu, Z., Li, T., and Hu, B. (2020). PRDM16 upregulation induced by MicroRNA-448 inhibition alleviates atherosclerosis via the TGF- β signaling pathway inactivation. *Front. Physiol.* 11:846. doi: 10.3389/fphys.2020.00846
- Londin, E., Loher, P., Telonis, A. G., Quann, K., Clark, P., Jing, Y., et al. (2015). Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1106–1115. doi: 10.1073/pnas.1420955112
- Lu, Y., Thavarajah, T., Gu, W., Cai, J., and Xu, Q. (2018). Impact of miRNA in Atherosclerosis. *Arterioscler. Thromb. Vasc. Biol.* 38, e159–e170. doi: 10.1161/ATVBAHA.118.310227
- Madrigal-Matute, J., Rotllan, N., Aranda, J. F., and Fernandez-Hernando, C. (2013). MicroRNAs and atherosclerosis. *Curr. Atheroscler. Rep.* 15:322. doi: 10.1007/s11883-013-0322-z
- Maitrias, P., Metzinger-Le Meuth, V., Massy, Z. A., M'Baya-Moutoula, E., Reix, T., Caus, T., et al. (2015). MicroRNA deregulation in symptomatic carotid plaque. *J. Vasc. Surg.* 62, 1245–1250. doi: 10.1016/j.jvs.2015.06.136
- Moghaddam, A. S., Afshari, J. T., Esmaili, S. A., Saburi, E., Joneidi, Z., and Momtazi-Borojeni, A. A. (2019). Cardioprotective microRNAs: lessons from stem cell-derived exosomal microRNAs to treat cardiovascular disease. *Atherosclerosis* 285, 1–9. doi: 10.1016/j.atherosclerosis.2019.03.016
- Niyazova, R., Berillo, O., Atambayeva, S., Pyrkova, A., Alybayeva, A., and Ivashchenko, A. (2015). MiR-1322 binding sites in paralogous and orthologous genes. *BioMed. Res. Int.* 1:e7. doi: 10.1155/2015/962637
- Orom, U. A., Nielsen, F. C., and Lund, A. H. (2008). MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol. Cell.* 30, 460–471. doi: 10.1016/j.molcel.2008.05.001
- Ren, K., Zhu, X., Zheng, Z., Mo, Z. C., Peng, X. S., Zeng, Y. Z., et al. (2018). MicroRNA-24 aggravates atherosclerosis by inhibiting selective lipid uptake from HDL cholesterol via the post-transcriptional repression of scavenger receptor class B type I. *Atherosclerosis* 270, 57–67. doi: 10.1016/j.atherosclerosis.2018.01.045
- Rodriguez, I., Coto, E., Reguero, J. R., Gonzalez, P., Andres, V., Lozano, I., et al. (2007). Role of the CDKN1A/p21, CDKN1C/p57, and CDKN2A/p16 genes in the risk of atherosclerosis and myocardial infarction. *Cell Cycle* 6, 620–625. doi: 10.4161/cc.6.5.3927
- Ryu, J., Ahn, Y., Kook, H., and Kim, Y. K. (2020). The roles of non-coding RNAs in vascular calcification and opportunities as therapeutic targets. *Pharmacol. Ther.* e107675. doi: 10.1016/j.pharmthera.2020.107675
- Santovito, D., Egea, V., and Weber, C. (2016). Small but smart: MicroRNAs orchestrate atherosclerosis development and progression. *Biochim. Biophys. Acta* 1861, 2075–2086. doi: 10.1016/j.bbali.2015.12.013
- Shi, X., Xie, X., Sun, Y., He, H., Huang, H., Liu, Y., et al. (2020). Paeonol inhibits NLRP3 mediated inflammation in rat endothelial cells by elevating hyperlipidemic rats plasma exosomal miRNA-223. *Eur. J. Pharmacol.* 885:173473. doi: 10.1016/j.ejphar.2020.173473
- Shoebi, S. (2020). Diagnostic and theranostic microRNAs in the pathogenesis of atherosclerosis. *Acta Physiol.* 228:E13353. doi: 10.1111/apha.13353
- Solly, E. L., Dimasi, C. G., Bursill, C. A., Psaltis, P. J., and Tan, J. T. M. (2019). MicroRNAs as therapeutic targets and clinical biomarkers in atherosclerosis. *J. Clin. Med.* 8:E2199. doi: 10.3390/jcm8122199
- Sun, B., Shan, Z., Sun, G., and Wang, X. (2020). MicroRNA-183-5p acts as a potential diagnostic biomarker for atherosclerosis and regulates the growth of vascular smooth muscle cell. *J. Chin. Med. Assoc.* doi: 10.1097/JCMA.0000000000000433
- Toba, H., Lindsey, M. L., and Chilton, R. J. (2014). Applications of miRNA technology for atherosclerosis. *Curr. Atheroscler. Rep.* 16:386. doi: 10.1007/s11883-013-0386-9
- Vargas-Alarcon, G., Perez-Hernandez, N., Rodriguez-Perez, J. M., Fragoso, J. M., Cardoso-Saldana, G., Vazquez-Vazquez, C., et al. (2019). MRE11A polymorphisms are associated with subclinical atherosclerosis and cardiovascular risk factors. a case-control study of the GEA mexican project. *Front. Genet.* 10:530. doi: 10.3389/fgene.2019.00530
- Velle-Forbord, T., Eidlaug, M., Debik, J., Sæther, J. C., Follstad, T., Nauman, J., et al. (2019). Circulating microRNAs as predictive biomarkers of myocardial infarction: evidence from the HUNT study. *Atherosclerosis* 289, 1–7. doi: 10.1016/j.atherosclerosis.2019.07.024
- Wang, C., Yang, W., Liang, X., Song, W., Lin, J., Sun, Y., et al. (2020). MicroRNA-761 modulates foam cell formation and inflammation through autophagy in the progression of atherosclerosis. *Mol. Cell. Biochem.* 474, 135–146. doi: 10.1007/s11010-020-03839-y

- Wang, M., Li, J., Cai, J., Cheng, L., Wang, X., Xu, P., et al. (2020). Overexpression of MicroRNA-16 alleviates atherosclerosis by inhibition of inflammatory pathways. *Biomed Res. Int.* 2020:8504238. doi: 10.1155/2020/8504238
- Wang, W., Ma, F., and Zhang, H. (2020). MicroRNA-374 is a potential diagnostic biomarker for atherosclerosis and regulates the proliferation and migration of vascular smooth muscle cells. *Cardiovasc. Diagn. Ther.* 10, 687–694. doi: 10.21037/cdt-20-444
- Wang, Y., Zhu, J., Handberg, A., Overvad, K., Tjønneland, A., Rimm, E. B., et al. (2018). Association between plasma CD36 levels and incident risk of coronary heart disease among Danish men and women. *Atherosclerosis* 277, 163–168. doi: 10.1016/j.atherosclerosis.2018.08.045
- Wiese, C. B., Zhong, J., Xu, Z. Q., Zhang, Y., Ramirez Solano, M. A., Zhu, W., et al. (2019). Dual inhibition of endothelial miR-92a-3p and miR-489-3p reduces renal injury-associated atherosclerosis. *Atherosclerosis* 282, 121–131. doi: 10.1016/j.atherosclerosis.2019.01.023
- You, L., Chen, H., Xu, L., and Li, X. (2020). Overexpression of miR-29a-3p suppresses proliferation, migration, and invasion of vascular smooth muscle cells in atherosclerosis via targeting TNFRSF1A. *Biomed. Res. Int.* 2020:9627974. doi: 10.1155/2020/9627974
- Yurikova, O. Y., Aisina, D. E., Niyazova, R. E., Atambayeva, S. A., Labeit, S., and Ivashchenko, A. T. (2019). The interaction of miRNA-5p and miRNA-3p with the mRNAs of Orthologous Genes. *Mol. Biol.* 53, 692–704. doi: 10.1134/S0026898419040189
- Zhang, R., Song, B., Hong, X., Shen, Z., Sui, L., and Wang, S. (2020). microRNA-9 inhibits vulnerable plaque formation and vascular remodeling via suppression of the SDC2-dependent FAK/ERK signaling pathway in mice with atherosclerosis. *Front. Physiol.* 11:804. doi: 10.3389/fphys.2020.00804
- Zhou, H., and Rigoutsos, I. (2014). MiR-103a-3p targets the 5' UTR of GPRC5A in pancreatic cells. *RNA* 20, 1431–1439. doi: 10.1261/rna.045757.114

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Mukushkina, Aisina, Pyrkova, Ryskulova, Labeit and Ivashchenko. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Structural and Functional Characteristics of miRNAs in Five Strategic Millet Species and Their Utility in Drought Tolerance

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Charu Lata,
National Institute of Science
Communication and Information
Resources (CSIR), India
Vetriventhan Mani,
International Crops Research Institute
for the Semi-Arid Tropics
(ICRISAT), India
Anil Kumar,
G. B. Pant University of Agriculture
and Technology, India

*Correspondence:

Nepolean Thirunavukkarasu
tnepolean@gmail.com;
nepolean@millets.res.in
Abhishek Rathore
a.rathore@cgiar.org

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 20 September 2020

Accepted: 05 November 2020

Published: 08 December 2020

Citation:

Chakraborty A, Viswanath A,
Malipatil R, Rathore A and
Thirunavukkarasu N (2020) Structural
and Functional Characteristics of
miRNAs in Five Strategic Millet
Species and Their Utility in Drought
Tolerance. *Front. Genet.* 11:608421.
doi: 10.3389/fgene.2020.608421

Animikha Chakraborty¹, Aswini Viswanath¹, Renuka Malipatil¹, Abhishek Rathore^{2*} and Nepolean Thirunavukkarasu^{1*}

¹ Genomics and Molecular Breeding Lab, Indian Council of Agricultural Research-Indian Institute of Millets Research, Hyderabad, India, ² Statistics, Bioinformatics and Data Management, International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India

Millets are the strategic food crops in arid and drought-prone ecologies. Millets, by virtue of nature, are very well-adapted to drought conditions and able to produce sustainable yield. Millets have important nutrients that can help prevent micro-nutrient malnutrition. As a result of the adverse effect of climate change and widespread malnutrition, millets have attained a strategic position to sustain food and nutritional security. Although millets can adapt well to the drought ecologies where other cereals fail completely, the yield level is very low under stress. There is a tremendous opportunity to increase the genetic potential of millet crops in dry lands when the genetics of the drought-tolerance mechanism is fully explained. MicroRNAs (miRNAs) are the class of small RNAs that control trait expression. They are part of the gene regulation but little studied in millets. In the present study, novel miRNAs and gene targets were identified from the genomic resources of pearl millet, sorghum, foxtail millet, finger millet, and proso millet through *in silico* approaches. A total of 1,002 miRNAs from 280 families regulating 23,158 targets were identified using different filtration criteria in five millet species. The unique as well as conserved structural features and functional characteristics of miRNA across millets were explained. About 84 miRNAs were conserved across millets in different species combinations, which explained the evolutionary relationship of the millets. Further, 215 miRNAs controlling 155 unique major drought-responsive genes, transcription factors, and protein families revealed the genetics of drought tolerance that are accumulated in the millet genomes. The miRNAs regulating the drought stress through specific targets or multiple targets showed through a network analysis. The identified genes regulated by miRNA genes could be useful in developing functional markers and used for yield improvement under drought in millets as well as in other crops.

Keywords: microRNAs, millets, drought stress, conserved genes, functional genes

INTRODUCTION

MicroRNAs (miRNAs) are small, single-stranded, non-coding, endogenous RNA of size varying from 21 to 24 nucleotides mainly involved in post-transcriptional gene regulation (Zhang et al., 2014; Aravind et al., 2017). They are highly conserved in matured form, and the conserved nature has made it a molecule of interest in several plant growth, development, and stress regulatory studies without any species boundaries. miRNAs regulate gene expression by targeting specific genes that are involved in biological processes, such as development and metabolic process, as well as target-specific transcription factors (TFs). Interaction between the miRNA-mRNA target is more important as it induces variation in the gene being expressed. Regulation of gene expression occurs in several ways, such as miRNA-directed mRNA cleavage, translational repression, chromatin remodeling, or epigenetic modification (Kumar et al., 2018). Single miRNA itself can target several genes involved in the same cellular signaling pathway (Alptekin et al., 2017).

Studies have shown that miRNAs and their targets are highly conserved across all major lineages of plant species, including dicots and monocots (Jones-Rhoades and Bartel, 2004). There are at least four theories about the origin of miRNAs: (1) inverted duplication events in the gene sequences; (2) duplication events from the protein-coding genes; (3) derived from the transposable elements, such as miniature inverted-repeat transposable elements (MITEs); and (4) accumulation of mutations in the inverted repeats and selection (Cuperus et al., 2011; Nozawa et al., 2012). Evolutionarily conserved miRNAs are mostly encoded by gene families. This, coupled with miRNA-mRNA target interaction, results in overlapping functions of miRNAs belonging to the same families (Jones-Rhoades et al., 2006).

miRNA families can be divided into two categories, based on the nature of their conservation and function. The first type is ancient in terms of evolution, highly conserved in the system with a high degree of expression. Conserved miRNAs are ubiquitous with low sequence variation and play an important role in basic biological functions through regulation of transcription factors and genes. The second type is relatively young in terms of evolution and expresses only when induced under specific conditions. Since they recently evolved to perform a specific function, the sequence variation is high in such types (Qin et al., 2014). Even though miRNA families are conserved among plants, they are more specific to the species of plant, physiological stages, type of organ/tissues, and stress conditions (Sun, 2012; Banerjee et al., 2016; Sunkar et al., 2017). This conserved nature over long evolutionary distances suggests the role of the evolutionarily conserved mechanism of miRNA in gene regulation (Molnar et al., 2007).

Abiotic stresses negatively impact plant growth, development, and productivity by altering the gene expression patterns. In order to cope with stress conditions, plants have developed several mechanisms over time, including the intricate interactions between stress-responsive elements and various molecular and biochemical factors affecting growth and development (Razmjoo et al., 2008). The plant molecular

responses to abiotic stresses involve interactions and crosstalk with many molecular pathways, including miRNA-mediated regulatory pathways (Bej and Basak, 2014). miRNA-mediated regulation involves a change in self-concentration and modifying the mRNA expression. These regulations, in turn, change the protein expression when exposed to stress (Ding et al., 2009; Wang et al., 2014).

Under drought stress, plants have evolved a series of protective mechanisms to withstand adverse conditions (Jaworski et al., 2010). Plants produce an array of gene regulation responses, which include triggering the expression of several stress-related genes, accumulation of osmotically active metabolites, and biosynthesis of specific proteins (Nepolean et al., 2014; Mittal et al., 2017a,b). Many studies have shown that the miRNAs were important modulators of drought tolerance in plants, where they modify the translation of target mRNAs that contain sequences that are complementary to the mature miRNAs. A study conducted in wild emmer wheat produced differential expression patterns of 13 miRNAs in response to drought stress (Kantar et al., 2011; Aravind et al., 2017).

Millets are a group of small grain crops of the family Poaceae, widely grown in the arid and semi-arid tropical regions of Asia and Africa. They are highly favored for food sustainability, owing to climate-resilient features, such as diverse adaptation to arid, semi-arid, and humid conditions. They tend to be less prone to biotic and abiotic stresses, and they can be grown in marginal lands (Kole et al., 2015). Compared to other cereals, millets show exceptional tolerance toward diverse abiotic stresses including drought, salinity, and heat stresses (Bandyopadhyay et al., 2017). In earlier studies on pearl millet, miRNAs were identified using non-pearl millet genomes (Jaiswal et al., 2018; Kumar et al., 2018). miRNAs in sorghum for drought (Katiyar et al., 2015; Hamza et al., 2016) and foxtail for drought (Wang et al., 2016) and dehydration stress (Yadav et al., 2019) using NGS approaches gave rise to a group of differentially regulating miRNAs. In our experiment, we selected five important millet species, namely pearl millet, sorghum, foxtail millet, finger millet, and proso millet, to mine the miRNAs using the latest genomic resources and *in silico* approaches and to track their stress-responsive features at the cellular, molecular, and physiological levels. We structurally characterized the miRNAs and identified conserved domains, such as sequence signatures in mature miRNAs across millet species. We identified target mRNAs regulated by the miRNAs and annotated the functional genes and transcription factors. We explained comprehensively how specific miRNAs regulating various genes in response to drought stress can be used to increase the productivity of millets and other crops.

MATERIALS AND METHODS

miRNA Reference and Genomic Data

Known microRNAs and their precursor sequences from different plants were obtained from miRNA databases PNRD (Yi et al., 2015) and miRBase (<http://www.miRbase.org/>). Of the 16,436 miRNAs we obtained, a total of 5,906 plant miRNAs served as reference after removal of the redundant sequences. Most of these

miRNAs were identified or verified by experiments, and others were computationally predicted as their close homologs.

Genomic sequences in the form of expressed sequence tags (ESTs), genome survey sequences (GSS), and whole-genome sequences (WGS) were retrieved from the NCBI (<http://www.ncbi.nlm.nih.gov/>) for five millet species: pearl millet [*Pennisetum glaucum* (L.) R. Br.], 5265 ESTs, 4105 GSS, and WGS; sorghum [*Sorghum bicolor* (L.) Moench], 80461 GSS and WGS; finger millet [*Eleusine coracana* (L.) Gaertn.], 2021 ESTs; foxtail millet [*Setaria italica* (L.) Beauv.], 66052 ESTs and 96975 GSS; and proso millet (*Panicum miliaceum* L.), 216 ESTs. The reference genome sequence assemblies used in our study for pearl millet, sorghum, finger millet, foxtail millet, and proso millet are GCA_002174835.2, GCF_000003195.3, GCA_002180455.1, GCA_000263155.2, and GCA_002895445.2, respectively (<https://www.ncbi.nlm.nih.gov/assembly>).

Pre-processing

Distinct pre-processing steps were taken to draw miRNA candidate sequences from the respective millet genomes. The BLAST version 2.6.0+ alignment tool was used for BLASTn to find the homologs. All known hairpin loop sequences were used as reference in the BLAST search against the genome with accurate parameters as follows: word-size 11 and E value cut-off 10^{-3} and 100 percent identity value, with a maximum three mismatches allowed and default settings for the remaining parameters. Flanking regions from both sides of the matched sequences were cut to a length of 70 nt and scanned by a sliding window of 100 nt (Wang et al., 2005). Duplicated sequences were discarded, and the remaining query sequences were searched using protein and nucleotide databases filtered to remove the presence of rRNA, tRNA, and mRNA, leaving the candidate sequences to be treated as miRNA precursors.

Structure Prediction

The secondary structure of the candidate precursor sequences was predicted by RNA fold (Hofacker et al., 1994) using the Vienna R package (Lorenz et al., 2011). The folding structures prediction uses the minimum free energy (MFE) algorithm and base pairing probability matrix. Sequences satisfying the criteria are as follows: (1) precursors having no more than three mismatches with previously known plant miRNAs; (2) the secondary structure should be folding into a perfect or near-perfect stem loop hairpin; (3) the mature sequence should be located in one of the arms of the stem loop; (4) presence of loop in the miRNA sequence is not allowed; (5) only the sequences with MFE lower than -20 kcal/mol are kept, and (6) A+U nucleotide content is 25–70% (Patanun et al., 2013).

The miRNAs position on the stem loop structure were predicted by the MirDup (Leclercq et al., 2013), using the training plant model for the candidate mature miRNAs. The classifier uses 10-fold cross validation. The ranking method was performed using the information gain evaluator in WEKA. The MirDup uses the random forest classifier trained with an unlimited maximum depth of the trees.

Potential Target Identification

Putative gene targets were identified by complementarity between miRNA and mRNA sequence. The targets against all miRNAs of five millet species were identified using the plant miRNA analysis psRNA Target Tool (<http://plantgrn.noble.org/psRNATarget/analysis>) (Dai and Zhao, 2011). The genome and EST sequences of the crops were taken as input, considering the seed region 2–13 nt, maximum UPE 25, and expectation value 5 at the most. miRNA and mRNA complementary sites were scored 0.5 or 1, according to the G:U match or non-match. Additionally, no more than two consecutive mismatches and no more than four mismatches between mature miRNA and potential target (Chai et al., 2015) were allowed. Due to lack of annotation of the millets, protein coding genes were identified from the putative target mRNAs using BLASTx from either the same crop or from other plants, such as *Arabidopsis* or rice.

Gene Ontology

The gene ontology (GO) base is continually evolving as biological knowledge increases and the curation of biological process develops. Target genes were subjected to functional annotation to reveal the miRNA-mediated gene regulatory network on biological processes, cellular components, and molecular functions. GO annotation was performed using DAVID (<https://david.ncifcrf.gov/>) for individual crops to understand the diverse function. The DAVID functional annotation clustering tool provided a module-centric approach for functional analysis of large gene lists. The grouping of data is based on functional categories and co-expression profiles, such as genes in the same pathway. A gene-term matrix gives different categorical clusters, such as GO biological process, GO molecular function, and GO cellular component and pathways (Huang et al., 2007). The calculation of over-representation of GO terms was done by applying the Fisher's exact test for count data and *p*-value (Benjamini and Hochberg, 1995). **Figure 1** shows the comprehensive workflow used for miRNA.

Gene Network

A network of drought targets of the miRNAs was created using Cytoscape (<https://cytoscape.org/>) (Shannon et al., 2003) to point out the hub genes of drought responsiveness.

RESULTS

Identification of Pre-miRNAs

The species-specific miRNAs were found by precise excision of the stem-loop precursor and single strands with a length of ~ 22 nt (Zhang et al., 2009). After BLASTn search against plant miRNAs from miRbase, the hits without protein coding sequences, tRNAs, and rRNAs were kept for secondary structure analysis. We identified 10,376, 9,064, 8,748, 8,173, and 3,055 homologous sequences in sorghum, foxtail, pearl, finger millet, and proso millet, respectively (**Table 1**). The homologous sequences were filtered by removing other RNA matches (tRNA, mRNA, rRNA), which resulted in the removal of 8,173, 3,493, 771, 478, and 27 candidate sequences from finger

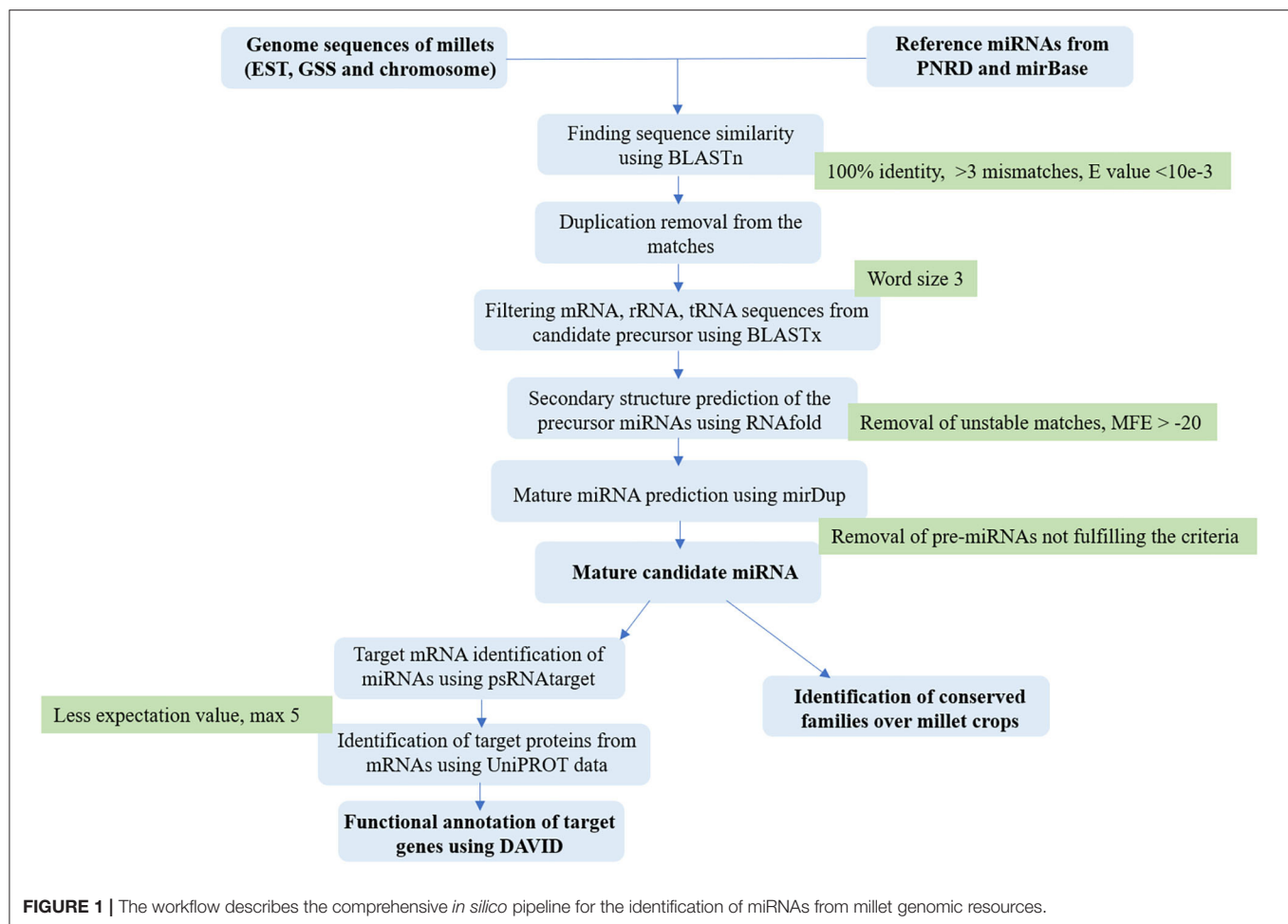


TABLE 1 | Pre- and mature miRNAs identified through *in silico* tools in five millet species.

Crop	Blast matches	Filtered-off tRNA, rRNA, and mRNA homologs	Entries without duplication	Stable match	Rejected miRNA candidates	Final mature miRNAs
Finger millet	8,173	8,173	0	75	58	17
Foxtail millet	9,064	771	8,293	1,615	1,242	373
Pearl millet	8,748	478	8,270	808	626	178
Proso millet	3,055	27	3,028	464	21	404
Sorghum	10,376	3,493	6,883	815	781	34

millet, sorghum, foxtail millet, pearl millet, and proso millet, respectively (**Table 1**).

The stability of the hairpin must be high with the lowest free energy of all other alternative folds for that sequence, which were predicted by RNA fold (Mathews et al., 2010). The minimal folding free energy (MFE) index is a major feature to distinguish putative precursor miRNAs from other RNAs. It was observed that more than 90% of the miRNAs had an MFE value < -30 (Zhang et al., 2006). Stable fold structures with MFE -20 were sorted out from the remaining structures producing 1,615, 815, 804, 434, and 75 stable precursor miRNAs from foxtail, sorghum, pearl millet, proso millet, and finger millet, respectively (**Table 1**).

miRNA Clusters in Genome

Among the candidate miRNAs, unique mature miRNAs were identified from foxtail millet, pearl millet, proso millet, and sorghum (**Figure 2**). Of these families, some were frequently present in different chromosomal regions in respective millets. In foxtail millet, a maximum number of miRNAs was observed in chromosome 1 (145 miRNAs), followed by chromosomes 2, 3, and 4 (50 miRNAs). Of the families, miR156 and miR157 were observed 9 times higher in chromosome 1. In pearl millet, miR167 and miR169 were the most frequent families, located in chromosomes 2 and 1, respectively, and chromosome 1 had the maximum miRNA families count among all chromosomes.

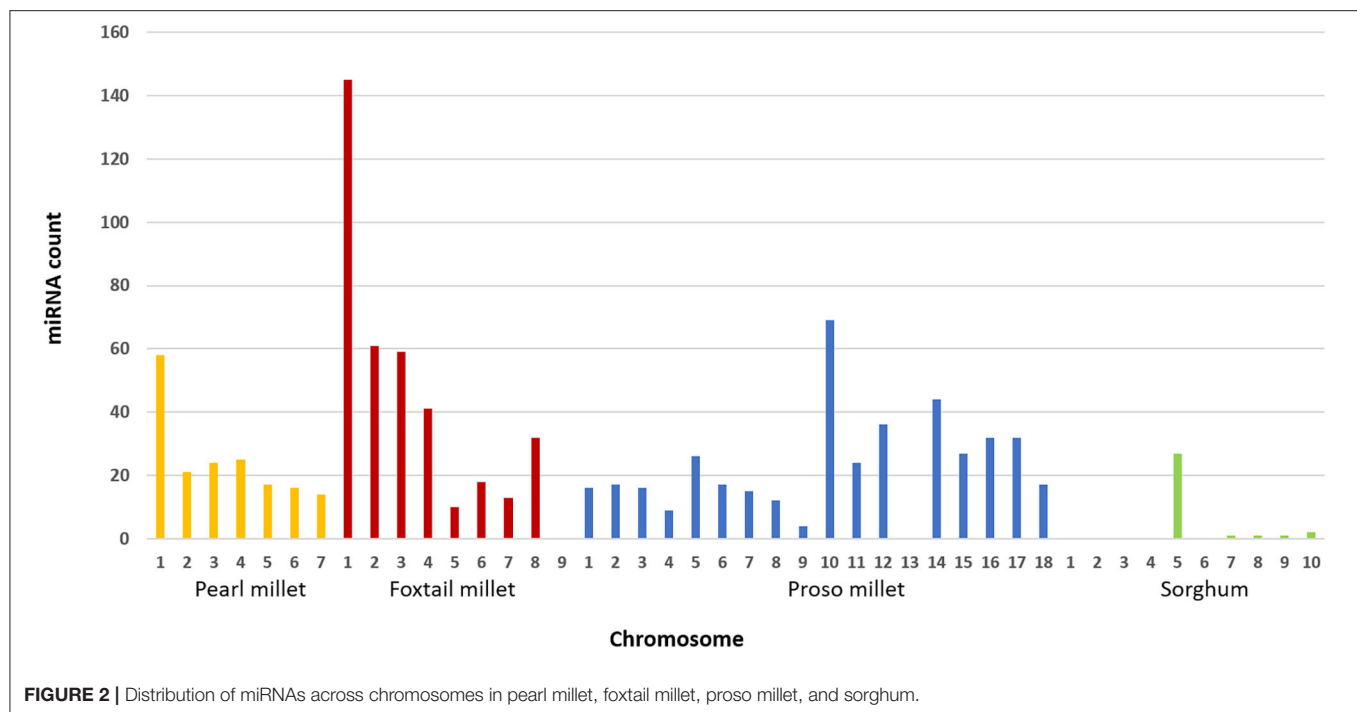


FIGURE 2 | Distribution of miRNAs across chromosomes in pearl millet, foxtail millet, proso millet, and sorghum.

Chromosome 10 in proso millet was the largest contributor with 68 miRNA families, and the major families were miR396, miR167, and miR166. In sorghum, chromosome 5 had the highest number of families. In finger millet, 14 unique families were identified, but the chromosome-specific miRNA count could not be tagged since genome was distributed in scaffolds.

Sequence Characteristics of New miRNAs

Transformation from precursor to mature form includes one major feature, the presence of mature 5' and 3' stems. Tools, such as Mature Bayes and the MirDup tool were used to bring out the mature form from the stable precursor miRNA molecule (Gkirtzou et al., 2010). A total of 404, 373, 174, 34, and 17 mature miRNAs were obtained from proso millet, foxtail millet, pearl millet, sorghum, and finger millet, respectively (Table 1).

The conserved miRNA sequences identified were variable in length across families, and in some families, the members were of uniform size. There were two classes of precursors with different structural properties. The most abundant class included precursors that had only two strongly conserved regions or blocks that consisted of the miRNA and miRNA* (* = complementary sequence of hairpin loop). The foldbacks of these precursors contained a short stem, consisting mainly of the miRNA/miRNA duplex. The consensus structures of miRNA families, such as miR156, miR160, miR170, miR171, miR395, and others are shown in Figure 3. A second and less frequent class, which includes the miRNA families miR159/319 and miR394, displays four conserved sequence blocks (Dezulian et al., 2006).

Conserved miRNAs Across Millets

The present study has revealed 77 miRNAs were observed commonly repeating in pearl millet, proso millet, and foxtail

millet (Figure 4). A smaller number of identified miRNAs in finger millet was not included in the analysis. Previously known reference miRNAs from sorghum and foxtail millet were also considered, which infers only three were common between old and newly discovered miRNA in sorghum, while for foxtail millet, all the miRNAs were new. The frequently co-occurring families (66) showed a conserved sequence pattern and consensus folding in the secondary structure. In our study, foxtail millet, proso millet, and pearl millet showed 27.9% conserved pattern in identified miRNAs, whereas foxtail and proso millet had more than 50% similarity (Figure 4). About 28–37% of miRNA families were matching among foxtail millet, sorghum, and proso millet. There was no conserved found in all millets together. In finger millet, we found only two miRNA families (miR845 and miR1873) common between pearl millet and proso millet due to incomplete genome and lack of annotation.

Target Prediction

In all five millets, most of the genes were identified as a result of specific stress conditions, such as drought, cold, salt, and water deficit. A total number of 7,090, 4,063, 1,754, 238, and 121 unique targets were identified in foxtail, sorghum, pearl millet, finger millet, and proso millet, respectively (Table 2). In pearl millet, miR134 acts on translation initiation factor like 5A-1/2, miR164a on heat shock TF, and miR172 on TF SUI1, which are evolutionarily conserved proteins (Koia et al., 2013). Some miRNAs control different physiological functions, such as miR10302, which was observed for regulating MYR1 protein related to flowering time under low light intensity in *Arabidopsis* (Zhao and Beers, 2013) and miR34 functions in seedling salt stress in broccoli (Tian et al., 2014).

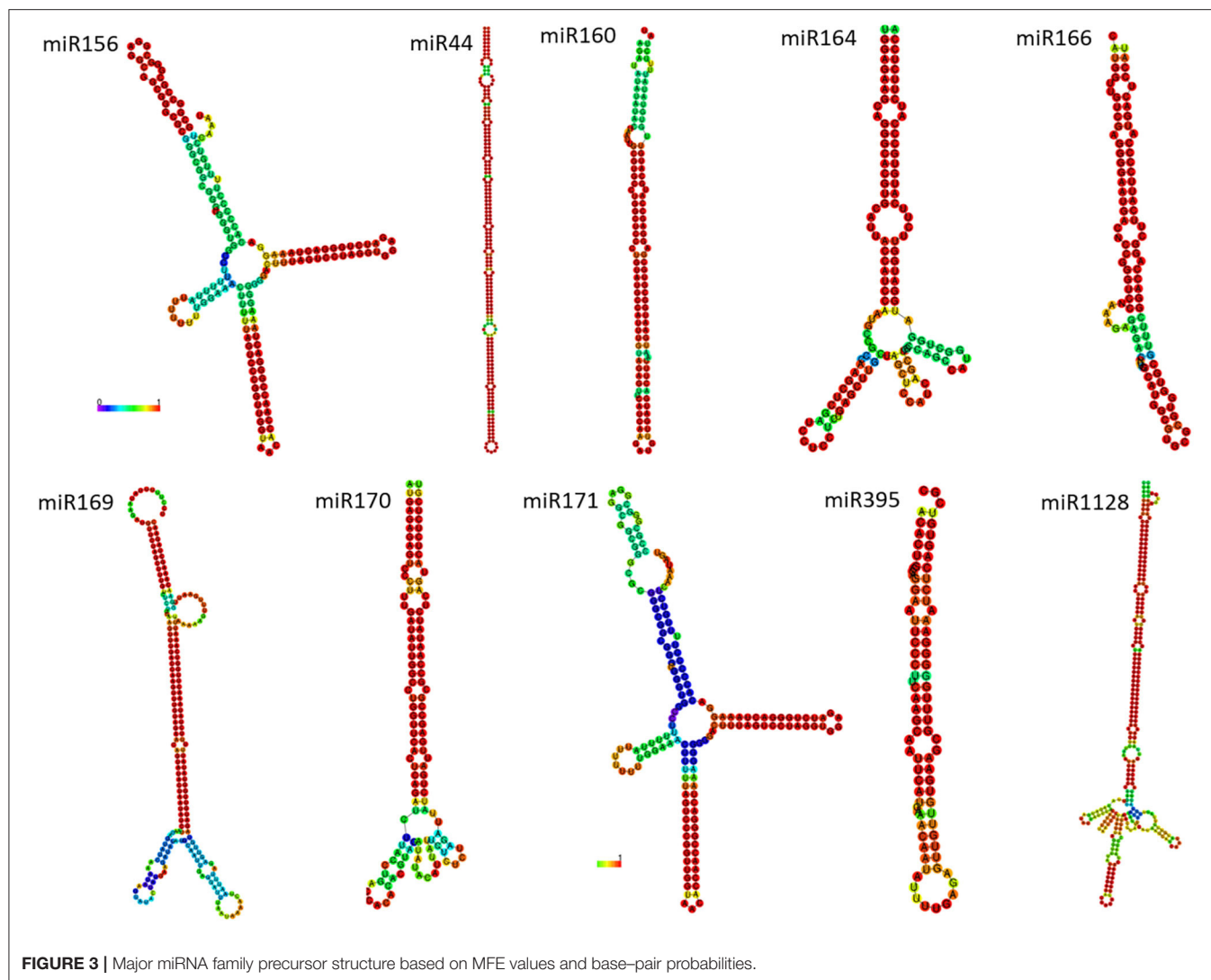


FIGURE 3 | Major miRNA family precursor structure based on MFE values and base-pair probabilities.

Out of 121 mRNA targets, 104 played essential roles in stress responses in proso millet, which was mostly targeted by miR44, miR156, miR159, miR160, miR166, miR169, miR171, miR172, and miR177. Enzymes, such as dioxygenase, kinases, dehydrin, and malate translocator operating on heat shock and ripening pathways were controlled by miR81, miR148, and miR164. Sorghum had 4,063 unique mRNA targets, mostly involved in different binding proteins, transcription factors, stress-related HSPs, and dehydration stress-related proteins, along with aminopeptidases and kinases.

In all millets, the miRNAs were mostly targeting the TFs, stress-related genes, metal-deficit factors, enzymes, and pathogenic factors other than normal metabolic and physiological regulators. Although the majority of plant miRNA targets were captured by the cut-off of expectation value <5 and mismatches 0 to 3, several authentic targets were missed due to lack of annotation.

GO Annotation

The target gene sets were subjected to GO analysis, which covers three domains (biological process, cellular component, and molecular function) to interpret the underlying functions of miRNAs (Figure 5). The biological process category showed a maximum number of genes participated in cellular process (545, 424, 153, 150, and 48 in sorghum, pearl millet, foxtail millet, finger millet, and proso millet, respectively), followed by metabolic processes, biological regulation, and cellular component. In cellular component, most of the genes were grouped in the cell, cell part, organelle, and protein-containing complex (2685, 2219, 1281, 102, and 98 genes in sorghum, foxtail, pearl, proso millet, and finger millet, respectively). Molecular function revealed the underlying activity of the genes, such as binding, catalytic activity, transcription regulation, transporter activity, molecular function regulation, and structural molecule activity.

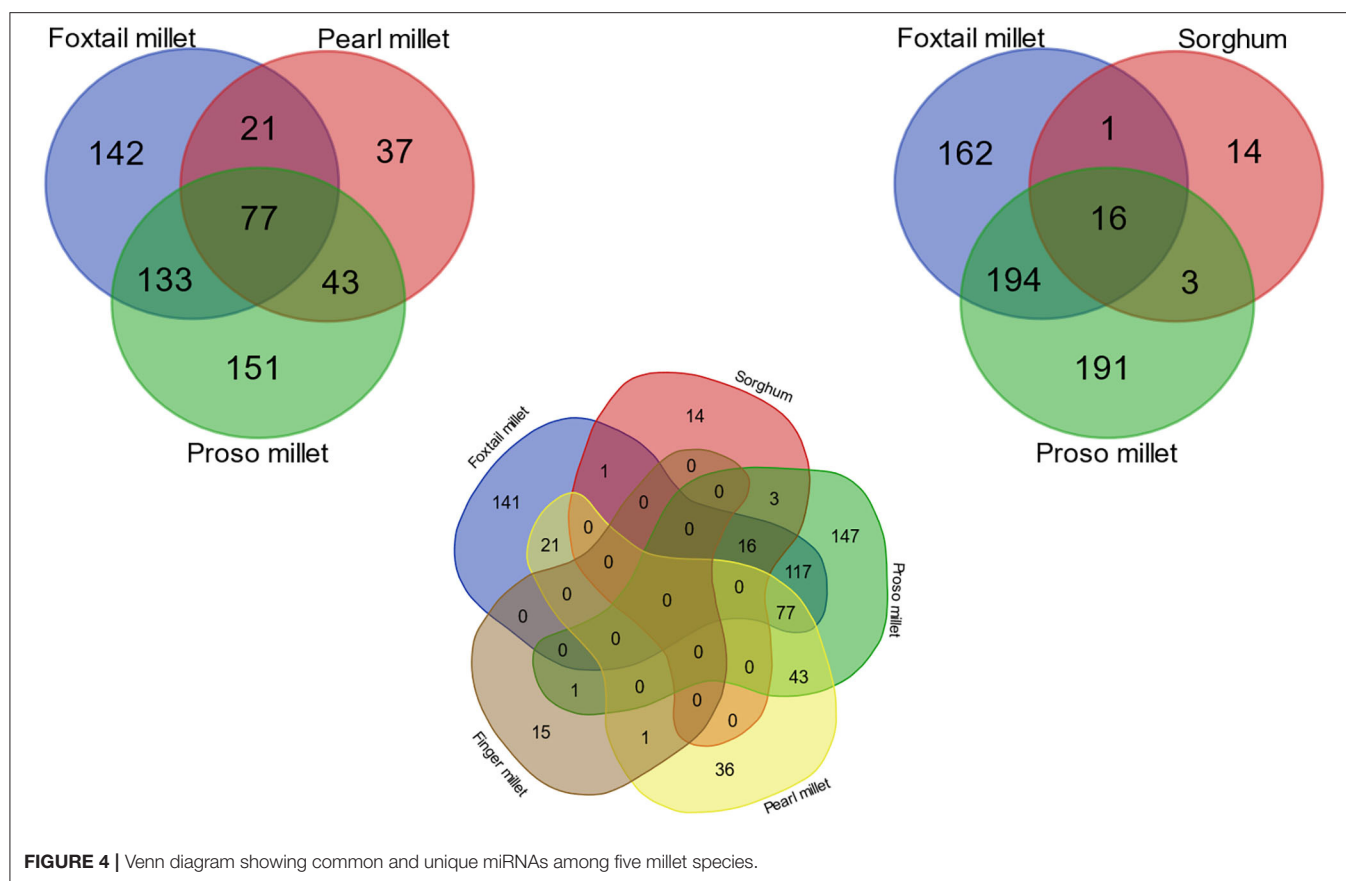


FIGURE 4 | Venn diagram showing common and unique miRNAs among five millet species.

TABLE 2 | miRNA-mediated genes, transcription factors, proteins, and enzymes identified in the genome of five millet species.

Feature	Foxtail millet	Sorghum	Proso millet	Pearl millet	Finger millet
Genome size (mbp)	423	666	851	1,817	1,195
Total gene target	13,714	5,882	319	2,977	266
Unique gene target	7,090	4,063	121	1,754	238
Transcription factor	902	193	–	24	14
Enzymatic activity	1,445	763	19	87	64
Stress regulatory	62	516	253	621	28
Carrier protein	43	10	21	102	21
Growth/physiological factor	39	–	–	–	92

DISCUSSION

Conserved miRNAs Across Millets

Identification of conserved miRNAs families on evolutionary basis among plants has provided a powerful approach to understanding gene regulation among related species. Identical miRNA sequences exist in closely and distantly related plant species. Many of the miRNAs discovered were found in a wide range of plant groups, from mosses to angiosperms. Some miRNA families exist broadly within the angiosperms, including eudicots and monocots, dating back to at least the early

Cretaceous. Several miRNA families also pre-date the divergence of gymnosperms and angiosperms (305 million years) and the divergence between vascular plants and mosses (490 million years). These results indicated that miRNA sequences are highly conserved across great phylogenetic distances and that similar selection pressures have been active in the regulation of gene expression in plant cells since the earliest stages of their evolution (Zhang et al., 2006).

Evolutionarily conserved miRNAs are mostly encoded by the same gene families, and the members of these families are physically clustered in the entire plant genome (Tanzer et al., 2005). The mature miRNA sequences are highly conserved among the same family members, and this extends to the entire stem-loop precursor duplex. Axtell and Bartel (2005) identified the occurrence of certain miRNA families (miRNA159 and miRNA319) common in ten different plant species and similarly the expression of miRNA165 and miRNA166 in nine plant species. We analyzed each millet with individually discovered potential miRNAs in other millets to find the orthologous candidates. All five millets species belong to same sub-family *Panicoidae*, in which pearl millet, finger millet, foxtail millet, and proso millet belong to the tribe *Panicaceae*, while sorghum belongs to another tribe, *Andropogoneae* (Vetriventhan et al., 2020).

Similar to mature miRNA sequences that are conserved across different plant species, the targets are also specific and conserved within the plant families by possessing highly

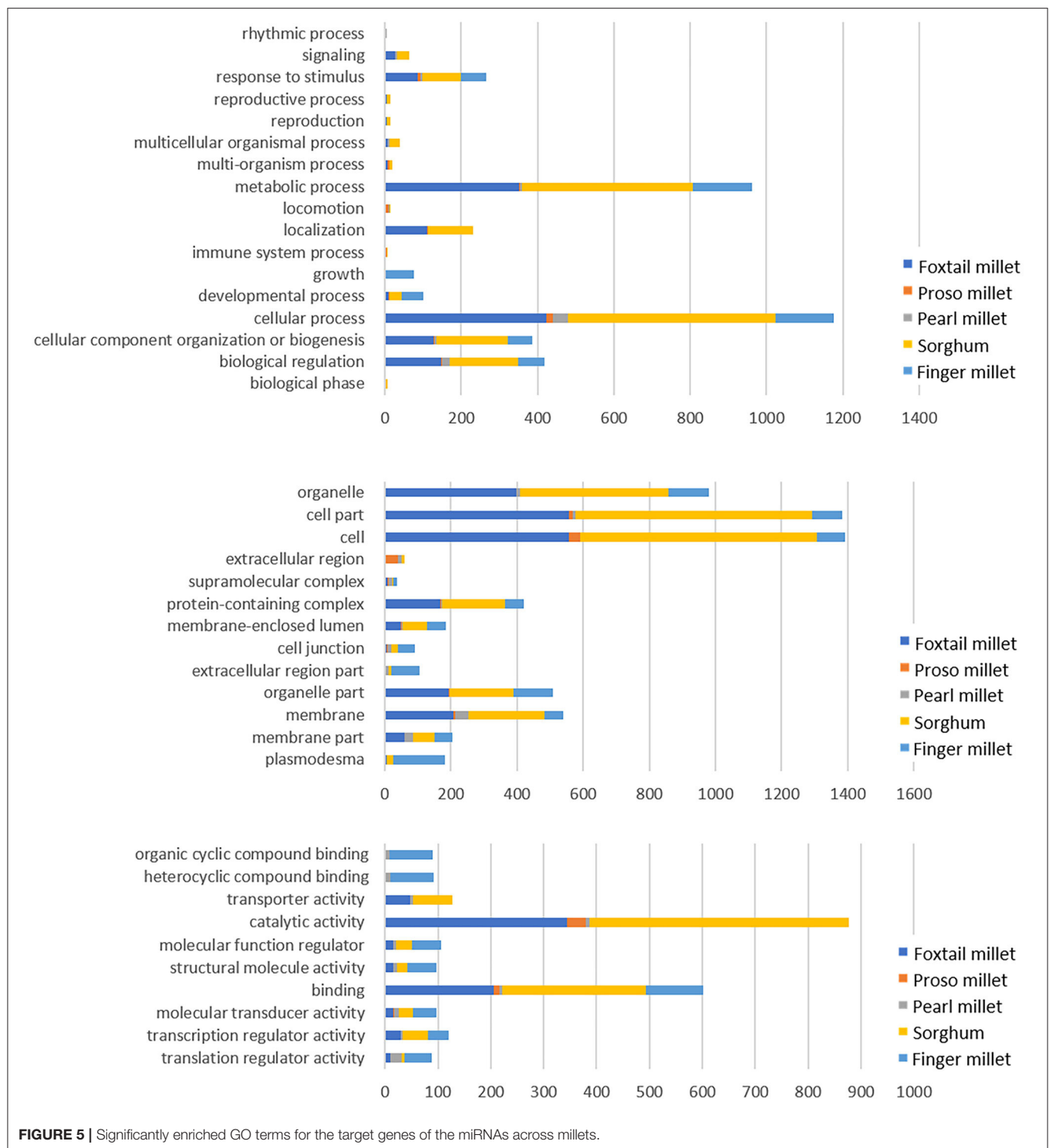


FIGURE 5 | Significantly enriched GO terms for the target genes of the miRNAs across millets.

conserved sequences at their complementary sites as we found in millet species. Although there are many nucleotide changes among the targets of different plant species, the sequences of the complementary sites are highly conserved. This is consistent with the study conducted by Floyd and Bowman (2004), in which class III HD-Zip genes targeted by miRNA 166 had specific

conserved target regions. It was observed that the sequence similarity of miRNA coupled with specific target results in overlapping functions of miRNAs belonging to the same families (Jones-Rhoades et al., 2006). Families, such as miR156, miR157, miR159, miR165, miR166, miR172, miR390, and so on were highly conserved in all vascular plants studied (You et al., 2017).

miRNAs are highly conserved in the plant kingdom, irrespective of the time of evolutionary divergence. Many families of miRNA are orthologous and homologous in different plant species spanning the breadth of green plant phylogeny (Dezulian et al., 2006). A comparative genomic experiment revealed that finger millet with sorghum produced 69 pairs of syntenic miRNA precursors, which were conserved between them, thereby indicating the evolutionary relationship of miRNA families across different species (Yi et al., 2013). Comparative analysis of miRNAs identified in *A. trichopoda* revealed that several miRNA families orthologous and paralogous with other crops. Among the conserved miRNAs identified in *A. trichopoda*, the miR407 family had three orthologs in *A. thaliana*, *Zea mays*, and *Gossypium hirsutum*. Similarly, there were two orthologs for *A. trichopoda* miR417 with *A. thaliana* and *Oryza sativa* (Hajieghrari et al., 2015). Relative to sorghum, proso millet and pearl millet are evolutionarily the closest species, with their supposed common ancestor dating back ~27 Mya, followed by ~8.3 Mya between pearl millet and foxtail (Singh et al., 2017). All millets possess considerable morphological differences but are evolutionarily well-related and thus share common structural and functional similarities.

miRNA and Genomic Targets

Prediction of the mRNA targets of the miRNAs identified from the millets will improve our understanding of the functions and regulation of these miRNAs. The conserved miRNAs, such as miR156, miR166, miR165, miR169, miR393, miR395, miR160, miR170, miR171, and miR172 were identified with 50–100 targets in different millets. We identified 902, 194, 24, and 14 TF targets in foxtail millet, sorghum, pearl millet, and finger millet, respectively, which were related to plant development,

phase change in growth, and other molecular functions. A set of 74 miRNA families was identified, targeting different stress regulatory factors, such as drought stress, dehydration stress, oxidative stress, and salt stress. Another important functional group of the predicted targets of miR155, miR156, miR169, miR172, miR2180, and miR2118 families were associated with enzymes, such as kinases, phosphate synthase, acetyltransferase, acid dehydratase, and hydrolase. Most of the miRNA targets were classified into the binding category and appeared to be involved in membrane, steroid, nucleic acid, protein, and ion binding. Around 30 miRNA families were involved in diverse molecular functions, including DNA binding, zinc ion binding, oxidoreductase activity, catalytic activity, protein kinase activity, and transferase activity in all five crops. Our target prediction results confirmed the widely held view that most plant miRNA targets encode TFs, which operate different mechanisms in the millets.

Drought-Responsive Target Prediction

Our study on identification of miRNA families targeting drought-related factors revealed that 89 targets were found in pearl millet regulated by 29 miRNAs. In proso millet, 64 miRNAs were found targeting mRNAs related to stresses, including drought-related pathways. Around 400 drought-specific mRNA targets were found in sorghum and foxtail millet controlling diverse pathways. In finger millet, 25 mRNAs specific to drought and water-deficient stress were identified to target nine different miRNAs. It was observed that many of the miRNAs families were both functionally and structurally conserved among the species, indicating a broad conservation of the regulatory roles in millets.

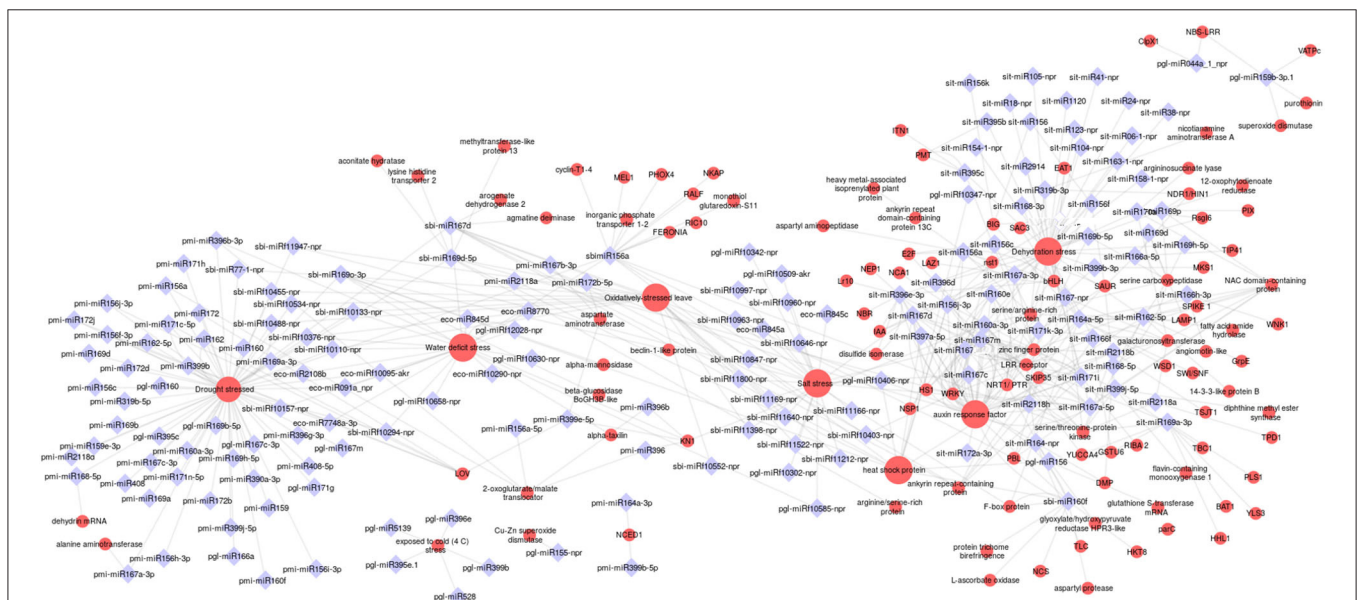


FIGURE 6 | Stress-specific interactions explained through miRNA-gene interconnected networks. The red circle indicates the drought mRNAs, the blue diamond indicates the miRNAs. The size of the red circle indicates the degree of connectivity of the target genes.

Network of Drought-Responsive miRNA and Their Targets

Our prediction of target genes revealed that multiple genes can be targeted by one specific miRNA, which suggested that the miRNA research should focus on networks more than on individual connections between miRNA and strongly predicted targets. To investigate the relationship between drought-responsive miRNA and their targets, a network analysis was carried out using the Cytoscape platform. The analysis incorporated all the non-conserved and conserved miRNAs of millet species belonging to 22 families. Targets involved in stress tolerance or plant development, such as genes encoding transcription factors, protein kinases, and phosphatases, and hormone-responsive factors were considered in the network analysis (**Figure 6**). It was found that the conserved miR156, miR160, and miR167

targeted up to 4 mRNAs in pearl and proso millet, and 20 in sorghum and foxtail millet. miR2108, miR170, and miR171 targeted important genes, such as auxin response factors (ARFs), NTR/PTR, NAC domain, and heat stress factors. Important TFs, such as WRKY and bHLH were regulated by miR399 and miR396 in foxtail and sorghum, respectively. Analysis of the network showed that miR160, miR156 in sorghum, and miR155 in foxtail millet had maximum connectivity. Drought stress-related mRNAs were one of the hubs (high-degree node in network) in the network with 86 connectivity. ARFs and dehydration stress were observed with a connectivity of 46 miRNAs each. Different TFs, such as bHLH, nst1, and WRKY were identified as the semi-hub targets. Six to ten miRNAs were identified as cross-interacting with drought-responsive factors and ARFs and water-deficit stress. Approximately 10 cross-talking miRNAs were identified between dehydration stress and TFs (bHLH and WRKY).

TABLE 3A | Species-specific microRNA families associated with drought-responsive traits.

miRNA	Stress-related mRNAs function	Species
miR104, miR105, miR1120, miR123, miR150, miR154, miR158, miR163, miR38, miR41, miR6, miR156, miR171, miR2108, miR164, miR395, miR167, miR170, miR160, miR319, miR169, miR172, miR165	Dehydration stress	Foxtail millet
miR396	Dehydration stress	Sorghum
miR1873, miR10095	Drought stress	Finger millet
miR399	Drought stress	Foxtail millet
miR160, miR164, miR167, miR169, miR171	Drought stress	Pearl millet
miR165, miR395	Drought stress	Pearl millet, proso millet
miR2108, miR1432, miR159, miR160, miR162, miR168, miR169, miR319, miR390, miR393, miR408, miR167, miR156, miR171, miR172, miR399	Drought stress	Proso millet
miR167, miR10294, miR156, miR169, miR156, miR396	Drought stress	Sorghum
miR10290	Drought stress, oxidative stress, water-deficit stress	Finger millet
miR11947	Drought stress, oxidatively stressed leaves, water-deficit stress	Pearl millet
miR10110, miR10133, miR10157, miR10376	Drought stress, oxidatively stressed leaves, water-deficit stress	Sorghum
miR2108, miR91	Drought stress, water-deficit stress	Finger millet
miR156, miR169, miR160, miR10403	Oxidative stress	Sorghum
miR167, miR77	Oxidative stress, water-deficit stress	Sorghum
miR845, miR8770, miR7748	Water-deficit stress	Finger millet
miR156	Water-deficit stress	Sorghum

miRNA-Mediated Gene Regulation of Drought Tolerance

Plants respond to several environmental stresses, among which drought is the major stress that limits the yield of many crops. Regulation of gene expression through miRNA and its target complementarity has made plants that tolerate drastic effects caused by drought. In our study, different categories of targets were identified across millets as being specific to that particular family of miRNA. These include different enzymes, such as protein kinases, peroxidase, amino/carboxy peptidases, stress-associated proteins (HSPs, RNA binding proteins), and drought-specific TFs (ARF, NAC family, MAD box, WRKY, bHLH, and ZFs). The important miRNA families related to drought along with the millet-specific target genes are presented in **Tables 3A, 3B**.

During the stress conditions, excess concentration of ROS is accumulated within the cells, resulting in cellular oxidative damage. Studies have shown the involvement of peroxidase in ROS scavenging under drought-stressed conditions. In sorghum, the peroxidase family targeting six miRNA families under drought conditions were reported (Katiyar et al., 2015). We have also found sorghum (sbi-miR160), foxtail millet (sit-miR169p, sit-miR171n, and sit-miR395c), and pearl millet (pgl-miR156) miRNAs regulating different peroxidase families. Plants exposed to drought or heat stress produce HSPs, which play a crucial role in protecting from stress conditions (Wang et al., 2004). They function as molecular chaperones, facilitating in folding of proteins, which is important for plants to cope with drought stress (Ford et al., 2011). We found miRNAs pgl-miR156 and pgl-miR10347 mediating HSPs. In sorghum, miRNA families targeting HSP were under down-regulation during drought stress (Katiyar et al., 2015).

Our study identified that the pmi-miR164, pmi-miR399 were involved in targeting NCED protein. It encodes 9-cis-epoxycarotenoid dioxygenase, which is negatively correlated with ABA accumulation, whereas in some species, its expression increased along with ABA accumulation under water stress (Changan et al., 2018). The sbi-miR164a (sorghum) and

TABLE 3B | Species-specific microRNA families associated with drought-responsive genes, transcription factors, and other molecules.

miRNA	Drought-responsive genes, transcription factors, and enzymes	Species
miR156	ARF, E2F, HSP, nst1, EAT1, EMB1444, LRBPK, BIG, bHLH, WRKY, HIPK, SRPK, kinase byr2-like, SAUR, BRG3, LAZ1, ANK, GLYR1, NCA1, NRT1/PTR, Znf, STK	Foxtail millet
miR160	Serine carboxypeptidase, KN1, YUCCA4, STK, NSP1, Znf, WRKY, XET, ARF	
miR162	BIG, 14-3-3-protein, NAC, HSP, STK, SRPK, GrpE	
miR164	Carboxylesterase, F-box, GAUT, ANK, GLDC, HPR3, TLC, DMP, LRR receptor, NRT1/PTR, PBL, STK, bHLH, NAC, SRPK, Znf, SWI/SNF, syntaxin, BIG, LAMP1, WSD1, WRKY, XET, 11S globulin seed storage protein	
miR165	AMOT, ARF, NDR1/HIN1, GSTU6, PBL, Znf, WNK1, parC, STK, TIP41, LRR receptor, carboxylesterase, WRKY, NAC	
miR167	LRBPK, LAZ1, Znf, IAA, NRT1/PTR, ARF, NAC, HS1, LRR receptor, SRPK, SAC3, PBL, STK	
miR168	bHLH, carboxylesterase, ARF, NAC, serine carboxypeptidase, DMP, NRT1/PTR	
miR169	12-oxophytodienoate reductase, NAC, NDR1/HIN1, PIX, STK, serine carboxypeptidase, YLS3, YUCCA4, BAT1, GSTU6, HHL1, LRR receptor, PLS1, Znf, nst1, LAMP1, Rsgl6	
miR170	GAUT, argininosuccinate lyase, NAC, Rsgl6	
miR171	ANK, ARF, SAUR, carboxylesterase, LRBPK, SKIP35, SPIKE 1, Znf, LRR receptor, RIBA 2, GAUT, NRT1/ PTR	
miR172	ARF, bHLH, Znf, STK	
miR2108	GAUT, TPD1, STK, ARF, HSP, LRR receptor, NRT1/ PTR, ANK, Znf, SKIP35, TSJT1	
miR319	Carboxylesterase, kinase byr2-like, NDR1/HIN1, aminotransferase, SRPK, PMT, Znf, ITN1	
miR397	Znf, auxin, HSP, LRR receptor	
miR399	LRBPK, NAC, syntaxin, TBC1, MKS1, WRKY, DPH, Znf, PBL, SAUR, STK	
miR10347	HSP	Pearl millet
miR155	Cu-Zn superoxide dismutase, elf-4A	
miR156	Glutathione S-transferase mRNA, HSP, RC12A	
miR159	Purothionin, NBS-LRR, superoxide dismutase, VATPc, protein kinase	
miR44	CipX1, NBS-LRR	
miR156	LOV	
miR164	NCED1	
miR167	Alanine aminotransferase, aspartate aminotransferase, LOV	
miR168	Dehydrin mRNA	
miR172	LOV, aspartate aminotransferase	
miR399	NCED1, LOV, aspartate aminotransferase, KN1	
miR10403	WRKY, Oxidatively stressed	Sorghum
miR156	Cyclin-T1-4, F-box, FERONIA, STK, FERONIA, GRX11, NKAP, HSP, cyclin-T1-4, F-box, MEL1, methylmalonate-semialdehyde dehydrogenase, PBL, PHOX4, RALF, RIC10, RING-H2 finger	
miR160	ARF, ACT domain-containing protein ACR4, aspartyl protease, HSP, HKT8, L-ascorbate oxidase, LECRK91, LRRP, CP12, F-box, NIR1, MYOB, NCS, NRT1/ PTR, TBL	
miR167	Agmatine deiminase, arginate dehydrogenase 2, HSP	
miR167	ANK, ARF	
miR169	BoGH3B, flotillin-like protein 2, MAN2A, alpha-taxilin, ATG6, METTL13, MYOB, STK LECRK4, HSP, WRKY, aconitate hydratase, arginine/serine-rich protein	
miR396	ARF, NEP1, IAA, Lr10, NBR, KN1, LRR receptor, NRT1/ PTR, nst1, bHLH, disulfide isomerase, WRKY, YLS7, BRG3, XET, E2F	

sit-miR2118d (foxtail millet) regulating protein NRT1/PTR family in *Arabidopsis* was suggested as a tolerance mechanism for abiotic stress through reallocation of nitrate to plant roots (Corratge-Faillie and Lacombe, 2017). The pgl-miR155 regulating Cu/ZnSOD transcripts found in pearl millet had an antioxidant protector system under water stress conditions in *L. corniculatus* (Borsani et al., 2001).

The sit-miR156, sit-miR164, sit-miR166, sit-miR167, sit-miR171, and sit-miR393 in foxtail millet and sbi-miR169 and

sbi-miR156 in sorghum were regulating the serine-threonine protein kinases (STPKs). It is observed that the phosphorylation state of several protein kinases changes when exposed to drought stress, implying their regulation in the drought-response signaling pathway. A change in concentration of free amino acid also has an impact on induction of drought stress in many plants. In *Brassica* leaves, the activities of alanine aminotransferase and aspartate amino transferase led to an overall decrease in protein synthesis under drought stress (Good and Zaplachinski, 1994).

The pmi-miR399, pmi-miR2118, and pmi-miR167 from proso millet, sbi-miR169 from sorghum, pgl-miR167 and pgl-miR528 from pearl millet, and sit-miR393, sit-miR396d, sit-miR397a, sit-miR156a, sit-miR160a, sit-miR164a, sit-miR166, and sit-miR172 from foxtail millet were found potentially regulating different aminotransferases.

It was observed that during drought stress, miRNA target genes code for specific TFs, which mediate the regulation of drought tolerance (Nepolean et al., 2014; Tang and Chu, 2017; Mittal et al., 2018). TF-mediated gene regulation includes several physiological and signaling pathways, such as abscisic acid (ABA)-mediated response, auxin signaling, osmotic, and antioxidant production (Ding et al., 2013). We found that many miRNAs are involved in TF-mediated gene regulation in different millet species in response to drought stress. NAC factors played an important role in drought tolerance through ABA signaling pathways (Wang et al., 2016; Aravind et al., 2017). It was found that NAC factors in foxtail millet was controlled by miR164. In rice, the overexpression of stress-responsive *NAC1* (*SNAC1*) in guard cells reduced the transpiration losses by increased stomatal closure (Singh et al., 2015). ARFs are the key elements mediated by auxins, which contribute to the drought stress tolerance. In cowpea, the upregulation of miR160a and miR160b targeted different ARFs, thereby resulting in drought tolerance. We also observed that sit-miR160a-3p especially targeted several ARF family members, including ARF10, ARF13, ARF18, and ARF22. The sbi-miR167 and sbi-miR156 targeting ARFs were found to play a major role in the process of plant growth and development. A study on sweet potato revealed that IbARF5 increased the contents of carotenoids and enhanced drought tolerance in transgenic *Arabidopsis* (Kang et al., 2018).

The role of bHLH57 in tolerance to drought, salt, and oxidative stresses was identified in finger millet (Babitha et al., 2015). The up-regulation of miRNA targeting bHLH when cowpea was exposed to drought stress (Barrera-Figueroa et al., 2011) was also observed. The expression of bHLH122 was recorded under drought and osmotic stress conditions in *Arabidopsis* (Liu et al., 2014). We also identified the regulation of bHLH by sit-miR156a, sit-miR164a, sit-miR168, sit-miR172m, and sit-miR396e (foxtail millet) and sbi-miR160f (sorghum). ZF proteins were associated with different developmental and other stress responses (Golldack et al., 2011). The miRNAs pgl-miR164 in pearl millet, sbi-miR169d in sorghum, and sit-miR167d, sit-miR2118d, sit-miR399j, sit-miR397a, and sit-miR395c in foxtail millet targeting ZFs were identified from our comparative study. Under drought conditions, the down-regulation of miRNA *vun_cand030*, which targeted ZF, was recorded in cowpea (Barrera-Figueroa et al., 2011).

Nuclear factor Y (NFY) is a major TF induced at the time of drought as we found in foxtail millet, in which the sit-miR169 family targeted NFYA 4, 5, 7, and 10. Li et al. (2008) reported that the NFYA was induced by ABA-dependent manner during drought stress. In *Arabidopsis*, miR169 targeting NFYA was down-regulated, but in cowpea, it induced the expression under drought stress (Barrera-Figueroa et al., 2011). In tomato, over-expressing miR169c exhibited better tolerance

to drought due to reduced stomata opening (Zhang et al., 2011).

Kelch repeat-containing F-box proteins are known to be involved in response to both biotic and abiotic stresses (Sun et al., 2010). Studies conducted in sorghum revealed that the presence of several drought-responsive miRNAs targeting Kelch repeat-containing F-box protein (Katiyar et al., 2015). The expression of many F-box proteins was also noticed in cowpea under drought stress (Jia et al., 2012). sbi-miR156a, sbi-miR160f, sbi-miR169d, and sit-miR156a, sit-miR395b, and sit-miR164a were found targeting F-BOX proteins in our study. WRKY acts as positive regulators of ABA signaling in several stress responses. Its involvement in heat stress in sunflower was controlled negatively by miR396 (Giacomelli et al., 2012). We have also identified sbi-miR169d, sit-miR156, sit-miR160a, sit-miR164, sit-miR396, and sit-miR166 targeting the WRKY.

We identified pgl-miR155, pgl-miR156 specific to MADS box proteins. Several studies have started to identify various members of the MADS-box gene family as an important molecular component involved in different types of stress responses. Studies showed that the MADS-box genes act as critical negative regulators of growth, improving plant survival, while others function as positive regulators of stress tolerance, associated with regulating the maintenance of primary metabolism, ABA signaling, ROS homeostasis, and detoxification processes through antioxidant enzymatic activities (Causier et al., 2002; Jia et al., 2018; Castelán-Muñoz et al., 2019; Zhao et al., 2020).

CONCLUSION

Novel miRNAs were identified by exploring the genomic resources of pearl millet, sorghum, foxtail millet, finger millet, and proso millet, and by comparing the miRNAs among millet species through a series of *in silico* approaches. Structural and functional classification of the identified miRNAs explained the unique and common features among the five millet species. The gene targets of miRNA were identified, and based on the GO annotation, they were classified into several functional groups. The drought-responsive gene targets regulated by the miRNAs were identified, and their role in drought tolerance were comprehensively explained. The genes can be further explored in trait improvement programs to enhance the productivity of millets in arid and drought-prone ecologies. Considering the conservative nature of miRNAs, the results of our experiment can also be used in other crops to understand the drought stress mechanism.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

NT planned and designed the research. AC, AV, RM, and AR conducted the experiments and data analysis. All authors contributed equally to the manuscript preparation, read, and approved the final manuscript.

FUNDING

The experiment was funded in part by the Bill & Melinda Gates Foundation (project no. 1011165) and the ICAR-Indian

Institute of Millets Research (project no. CI/2018-23/120). The funding agencies had no role in the study design, data collection and analysis, the decision to publish, or the preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.608421/full#supplementary-material>

REFERENCES

- Alptekin, B., Akpinar, B. A., and Budak, H. (2017). A comprehensive prescription for plant miRNA identification. *Front. Plant Sci.* 7:2058. doi: 10.3389/fpls.2016.02058
- Aravind, J., Rinku, S., Pooja, B., Shikha, M., Kaliyugam, S., Mallikarjuna, M. G., et al. (2017). Identification, characterization and functional validation of microRNAs in subtropical maize inbreds. *Front. Plant Sci.* 8:941. doi: 10.3389/fpls.2017.00941
- Axtell, M. J., and Bartel, D. P. (2005). Antiquity of microRNAs and their targets in land plants. *Plant Cell* 17, 1658–1673. doi: 10.1105/tpc.105.032185
- Babitha, K. C., Vemanna, R. S., Nataraja, K. N., and Udayakumar, M. (2015). Overexpression of EcbHLH57 transcription factor from *Eleusine coracana* L. in tobacco confers tolerance to salt, oxidative and drought stress. *PLoS ONE* 10:e0137098. doi: 10.1371/journal.pone.0137098
- Bandyopadhyay, T., Muthamilarasan, M., and Prasad, M. (2017). Millets for next generation climate-smart agriculture. *Front. Plant Sci.* 8:1266. doi: 10.3389/fpls.2017.01266
- Banerjee, A., Roychoudhury, A., and Krishnamoorthi, S. (2016). Emerging techniques to decipher microRNAs (miRNAs) and their regulatory role in conferring abiotic stress tolerance of plants. *Plant Biotechnol. Rep.* 10, 185–205. doi: 10.1007/s11816-016-0401-z
- Barrera-Figueroa, B. E., Gao, L., Diop, N. N., Wu, Z., Ehlers, J. D., Roberts, P. A., et al. (2011). Identification and comparative analysis of drought-associated microRNAs in two cowpea genotypes. *BMC Plant Biol.* 11:127. doi: 10.1186/1471-2229-11-127
- Bej, S., and Basak, J. (2014). MicroRNAs: the potential biomarkers in plant stress response. *Am. J. Plant Sci.* 5, 748–759. doi: 10.4236/ajps.2014.55089
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
- Borsani, O., Diaz, P., Agius, M. F., Valpuesta, V., and Monza, J. (2001). Water stress generates an oxidative stress through the induction of a specific Cu/Zn superoxide dismutase in *Lotus corniculatus* leaves. *Plant Sci.* 161, 757–763. doi: 10.1016/S0168-9452(01)00467-8
- Castelán-Muñoz, N., Herrera, J., Cajero-Sánchez, W., Arrizubieta, M., Trejo, C., García-Ponce, B., et al. (2019). MADS-box genes are key components of genetic regulatory networks involved in abiotic stress and plastic developmental responses in plants. *Front. Plant Sci.* 10:853. doi: 10.3389/fpls.2019.00853
- Causier, B., Kieffer, M., and Davies, B. (2002). MADS-box genes reach maturity. *Science* 296, 275–276. doi: 10.1126/science.1071401
- Chai, J., Feng, R., Shi, H., Ren, M., Zhang, Y., and Wang, J. (2015). Bioinformatic identification and expression analysis of banana microRNAs and their targets. *PLoS ONE* 10:e0123083. doi: 10.1371/journal.pone.0123083
- Changan, S. S., Ali, K., Kumar, V., Garg, N. K., and Tyagi, A. (2018). Absciscic acid biosynthesis under water stress: anomalous behavior of the 9-cis-epoxycarotenoid dioxygenase1 (NCED1) gene in rice. *Biol. Plant.* 62, 663–670. doi: 10.1007/s10535-018-0807-2
- Corratge-Faillie, C., and Lacombe, B. (2017). Substrate (un) specificity of Arabidopsis NRT1/PTR FAMILY (NPF) proteins. *J. Exp. Bot.* 68, 3107–3113. doi: 10.1093/jxb/erw499
- Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of MIRNA genes. *Plant Cell* 23, 431–442. doi: 10.1105/tpc.110.082784
- Dai, X., and Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* 39, W155–W159. doi: 10.1093/nar/gky316
- Dezulian, T., Remmert, M., Palatnik, J. F., Weigel, D., and Huson, D. H. (2006). Identification of plant microRNA homologs. *Bioinformatics* 22, 359–360. doi: 10.1093/bioinformatics/bti802
- Ding, D., Zhang, L., Wang, H., Liu, Z., Zhang, Z., and Zheng, Y. (2009). Differential expression of miRNAs in response to salt stress in maize roots. *Ann. Bot.* 103, 29–38. doi: 10.1093/aob/mcn205
- Ding, Y., Tao, Y., and Zhu, C. (2013). Emerging roles of microRNAs in the mediation of drought stress response in plants. *J. Exp. Bot.* 64, 3077–3086. doi: 10.1093/jxb/ert164
- Floyd, S. K., and Bowman, J. L. (2004). Gene regulation: ancient microRNA target sequences in plants. *Nature* 428, 485–486. doi: 10.1038/428485a
- Ford, K. L., Cassin, A., and Bacic, A. F. (2011). Quantitative proteomic analysis of wheat cultivars with differing drought stress tolerance. *Front. Plant Sci.* 2:44. doi: 10.3389/fpls.2011.00044
- Giacomelli, J. I., Weigel, D., Chan, R. L., and Manavella, P. A. (2012). Role of recently evolved miRNA regulation of sunflower HaWRKY6 in response to temperature damage. *New Phytol.* 195, 766–773. doi: 10.1111/j.1469-8137.2012.04259.x
- Gkirtzou, K., Tsamardinos, I., Tsakalides, P., and Poirazi, P. (2010). MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors. *PLoS ONE* 5:e11843. doi: 10.1371/journal.pone.0011843
- Gollack, D., Lüking, I., and Yang, O. (2011). Plant tolerance to drought and salinity: stress regulating transcription factors and their functional significance in the cellular transcriptional network. *Plant Cell Rep.* 30, 1383–1391. doi: 10.1007/s00299-011-1068-0
- Good, A. G., and Zaplachinski, S. T. (1994). The effects of drought stress on free amino acid accumulation and protein synthesis in *Brassica napus*. *Physiol. Plant.* 90, 9–14. doi: 10.1111/j.1399-3054.1994.tb02185.x
- Hajieghrari, B., Farrokhi, N., Goliaei, B., and Kavousi, K. (2015). Computational identification, characterization and analysis of conserved miRNAs and their targets in *Amborella Trichopoda*. *J. Data Mining Genom. Proteom.* 6:168. doi: 10.4172/2153-0602.1000168
- Hamza, N. B., Sharma, N., Tripathi, A., and Mishra, N. S. (2016). MicroRNA expression profiles in response to drought stress in *Sorghum bicolor*. *Gene Exp. Patterns* 20, 88–98. doi: 10.1016/j.gexp.2016.01.001
- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Chem. Monthly* 125, 167–188. doi: 10.1007/BF00818163
- Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., et al. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 8:R183. doi: 10.1186/gb-2007-8-9-r183
- Jaiswal, S., Antala, T. J., Mandavia, M. K., Chopra, M., Jasrotia, R. S., Tomar, R. S., et al. (2018). Transcriptomic signature of drought response in pearl millet (*Pennisetum glaucum* L.) and development of web-genomic resources. *Sci. Rep.* 8:3382. doi: 10.1038/s41598-018-21560-1

- Jaworski, K., Grzegorzewska, W., Swiezawska, B., and Szmidt-Jaworska, A. (2010). Participation of second messengers in plant responses to abiotic stress. *Postepy Biol. Komorki* 37, 847–868.
- Jia, J., Zhao, P., Cheng, L., Yuan, G., Yang, W., et al. (2018). MADS-box family genes in sheepgrass and their involvement in abiotic stress responses. *BMC Plant Biol.* 18:42. doi: 10.1186/s12870-018-1259-8
- Jia, Y., Gu, H., Wang, X., Chen, Q., Shi, S., Zhang, J., et al. (2012). Molecular cloning and characterization of an F-box family gene CarF-box1 from chickpea (*Cicer arietinum* L.). *Mol. Biol. Rep.* 39, 2337–2345. doi: 10.1007/s11033-011-0984-y
- Jones-Rhoades, M. W., and Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* 14, 787–799. doi: 10.1016/j.molcel.2004.05.027
- Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.* 57, 19–53. doi: 10.1146/annurev.arplant.57.032905.105218
- Kang, C., He, S., Zhai, H., Li, R., Zhao, N., and Liu, Q. (2018). A sweet potato auxin response factor gene (IbARF5) is involved in carotenoid biosynthesis and salt and drought tolerance in transgenic Arabidopsis. *Front. Plant Sci.* 9:1307. doi: 10.3389/fpls.2018.01307
- Kantar, M., Lucas, S. J., and Budak, H. (2011). miRNA expression patterns of *Triticum dicoccoides* in response to shock drought stress. *Planta* 233, 471–484. doi: 10.1007/s00425-010-1309-4
- Katiyar, A., Smita, S., Muthusamy, S. K., Chinnusamy, V., Pandey, D. M., and Bansal, K. C. (2015). Identification of novel drought-responsive microRNAs and trans-acting siRNAs from *Sorghum bicolor* (L.) Moench by high-throughput sequencing analysis. *Front. Plant Sci.* 6:506. doi: 10.3389/fpls.2015.00506
- Koia, J., Moyle, R., Hendry, C., Lim, L., and Botella, J. R. (2013). Pineapple translation factor SU11 and ribosomal protein L36 promoters drive constitutive transgene expression patterns in *Arabidopsis thaliana*. *Plant Mol. Biol.* 81, 327–336. doi: 10.1007/s11103-012-0002-3
- Kole, C., Muthamilarasan, M., Henry, R., Edwards, D., Sharma, R., Abberton, M., et al. (2015). Application of genomics-assisted breeding for generation of climate resilient crops: progress and prospects. *Front. Plant Sci.* 6:563. doi: 10.3389/fpls.2015.00563
- Kumar, A., Kumar, S., Fougat, R. S., and Zala, H. N. (2018). *In-silico* identification and validation of miRNAs in pearl millet (*Pennisetum glaucum* L.). *Curr. Plant Biol.* 14, 41–49. doi: 10.1016/j.cpb.2018.09.007
- Leclercq, M., Diallo, A. B., and Blanchette, M. (2013). Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucleic Acids Res.* 41, 7200–7211. doi: 10.1093/nar/gkt466
- Li, W. X., Oono, Y., Zhu, J., He, X. J., Wu, J. M., Iida, K., et al. (2008). The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and post transcriptionally to promote drought resistance. *Plant Cell* 20, 2238–2251. doi: 10.1105/tpc.108.059444
- Liu, W., Tai, H., Li, S., Gao, W., Zhao, M., Xie, C., et al. (2014). bHLH 122 is important for drought and osmotic stress resistance in Arabidopsis and in the repression of ABA catabolism. *New Phytol.* 201, 1192–1204. doi: 10.1111/nph.12607
- Lorenz, R., Bernhart, S. H., Honer zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., et al. (2011). ViennaRNA package 2.0. algorithms. *Mol. Biol.* 6:26. doi: 10.1186/1748-7188-6-26
- Mathews, D. H., Moss, W. N., and Turner, D. H. (2010). Folding and finding RNA secondary structure. *Cold Spring Harbor Perspect. Biol.* 2:a003665. doi: 10.1101/cshperspect.a003665
- Mittal, S., Arora, K., Rao, A. R., Mallikarjuna, M. G., Gupta, H. S., and Nepolean, T. (2017a). Genomic selection for drought tolerance using genome-wide SNPs in maize. *Front. Plant Sci.* 8:550. doi: 10.3389/fpls.2017.00550
- Mittal, S., Banduni, P., Mallikarjuna, M. G., Rao, A. R., Jain, P. A., Dash, P. K., et al. (2018). Structural, functional, and evolutionary characterization of major drought transcription factors families in maize. *Front. Chem.* 6:177. doi: 10.3389/fchem.2018.00177
- Mittal, S., Mallikarjuna, M. G., Rao, A. R., Jain, P. A., Dash, P. K., and Nepolean, T. (2017b). Comparative analysis of CDPK family in maize, arabidopsis, rice, and sorghum revealed potential targets for drought tolerance improvement. *Front. Chem.* 5:115. doi: 10.3389/fchem.2017.00115
- Molnar, A., Schwach, F., Studholme, D. J., Thuenemann, E. C., and Baulcombe, D. C. (2007). miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature* 447, 1126–1129. doi: 10.1038/nature05903
- Nepolean, T., Hossain, F., Arora, K., Sharma, R., Kaliyugam, S., Swati, M., et al. (2014). Functional mechanisms of drought tolerance identified in subtropical maize using genome-wide association mapping. *BMC Genomics* 15:1182. doi: 10.1186/1471-2164-15-1182
- Nozawa, M., Miura, S., and Nei, M. (2012). Origins and evolution of microRNA genes in plant species. *Genome Biol. Evol.* 4, 230–239. doi: 10.1093/gbe/evs002
- Patanun, O., Lertpanyasampatha, M., Sojikul, P., Viboonjun, U., and Narangajavana, J. (2013). Computational identification of microRNAs and their targets in cassava (*Manihot esculenta* Crantz.). *Mol. Biotechnol.* 53, 257–269. doi: 10.1007/s12033-012-9521-z
- Qin, Z., Li, C., Mao, L., and Wu, L. (2014). Novel insights from non-conserved microRNAs in plants. *Front. Plant Sci.* 5:586. doi: 10.3389/fpls.2014.00586
- Razmjoo, K., Heydarizadeh, P., and Sabzalain, M. R. (2008). Effect of salinity and drought stresses on growth parameters and essential oil content of Matricaria chamomile. *Int. J. Agri. Biol.* 10, 451–454.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Singh, B. P., Singh, B., Kumar, V., Singh, P. K., Jayaswal, P. K., Mishra, S., et al. (2015). Haplotype diversity and association analysis of SNAC1 gene in wild rice germplasm. *Indian J. Genet. Plant Breed.* 75, 157–166. doi: 10.5958/0975-6906.2015.00025.5
- Singh, R. K., Muthamilarasan, M., and Prasad, M. (2017). “Foxtail millet: an introduction,” in *The Foxtail Millet Genome*, ed. Prasad, M (Cham: Springer), p. 1–9. doi: 10.1007/978-3-319-65617-5_1
- Sun, G. (2012). MicroRNAs and their diverse functions in plants. *Plant Mol. Biol.* 80, 17–36. doi: 10.1007/s11103-011-9817-6
- Sun, S. J., Guo, S. Q., Yang, X., Bao, Y. M., Tang, H. J., Sun, H., et al. (2010). Functional analysis of a novel Cys2/His2-type zinc finger protein involved in salt tolerance in rice. *J. Exp. Bot.* 61, 2807–2818. doi: 10.1093/jxb/erq120
- Sunkar, R., Maheswari, M., and Chakraborty, S. (2017). Small RNAs: regulators of plant development and climate resilience. *Indian J. Plant Physiol.* 22, 369–370. doi: 10.1007/s40502-01-0349-1
- Tang, J., and Chu, C. (2017). MicroRNAs in crop improvement: fine-tuners for complex traits. *Nat. Plants* 3:17077. doi: 10.1038/nplants.2017.77
- Tanzer, A., Amemiya, C. T., Kim, C. B., and Stadler, P. F. (2005). Evolution of microRNAs located within Hox gene clusters. *J. Exp. Zool. Mol. Dev. Evol.* 304, 75–85. doi: 10.1002/jez.b.21021
- Tian, Y., Tian, Y., Luo, X., Zhou, T., Huang, Z., Liu, Y., et al. (2014). Identification and characterization of microRNAs related to salt stress in broccoli, using high-throughput sequencing and bioinformatics analysis. *BMC Plant Biol.* 14:226. doi: 10.1186/s12870-014-0226-2
- Vetriventhan, M., Azevedo, V. C., Upadhyaya, H. D., Nirmalakumari, A., Kane-Potaka, J., Anitha, S., et al. (2020). Genetic and genomic resources, and breeding for accelerating improvement of small millets: current status and future interventions. *Nucleus* 1–23. doi: 10.1007/s13237-020-00322-3
- Wang, B., Sun, Y. F., Song, N., Wei, J. P., Wang, X. J., Feng, H., et al. (2014). MicroRNAs involving in cold, wounding and salt stresses in *Triticum aestivum* L. *Plant Physiol. Biochem.* 80, 90–96. doi: 10.1016/j.plaphy.2014.03.020
- Wang, H., Wang, H., Shao, H., and Tang, X. (2016). Recent advances in utilizing transcription factors to improve plant abiotic stress tolerance by transgenic technology. *Front. Plant Sci.* 7:67. doi: 10.3389/fpls.2016.00067
- Wang, W., Vinocur, B., Shoseyov, O., and Altman, A. (2004). Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* 9, 244–252. doi: 10.1016/j.tplants.2004.03.006
- Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., et al. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics* 21, 3610–3614. doi: 10.1093/bioinformatics/bti562
- Yadav, A., Khan, Y., and Prasad, M. (2019). Dehydration-responsive miRNAs in foxtail millet: genome-wide identification, characterization and expression profiling. *Planta* 243, 749–766. doi: 10.1007/s00425-015-2437-7
- Yi, F., Xie, S., Liu, Y., Qi, X., and Yu, J. (2013). Genome-wide characterization of microRNA in foxtail millet (*Setaria italica*). *BMC Plant Biol.* 13:212. doi: 10.1186/1471-2229-13-212

- Yi, X., Zhang, Z., Ling, Y., Xu, W., and Su, Z. (2015). PNRD: a plant non-coding RNA database. *Nucleic Acids Res.* 43, D982–D989. doi: 10.1093/nar/gku1162
- You, C., Cui, J., Wang, H., Qi, X., Kuo, L. Y., Ma, H., et al. (2017). Conservation and divergence of small RNA pathways and microRNAs in land plants. *Genome Biol.* 18, 1–19. doi: 10.1186/s13059-017-1291-2
- Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006). Conservation and divergence of plant microRNA genes. *Plant J.* 46, 243–259. doi: 10.1111/j.1365-313X.2006.02697.x
- Zhang, N., Yang, J., Wang, Z., Wen, Y., Wang, J., et al. (2014). Identification of novel and conserved microRNAs related to drought stress in potato by deep sequencing. *PLoS ONE* 9:e95489. doi: 10.1371/journal.pone.0095489
- Zhang, W., Luo, Y., Gong, X., Zeng, W., and Li, S. (2009). Computational identification of 48 potato microRNAs and their targets. *Comput. Biol. Chem.* 33, 84–93. doi: 10.1016/j.compbiolchem.2008.07.006
- Zhang, X., Zou, Z., Gong, P., Zhang, J., Ziaf, K., Li, H., et al. (2011). Over-expression of microRNA169 confers enhanced drought tolerance to tomato. *Biotechnol. Lett.* 33, 403–409. doi: 10.1007/s10529-010-0436-0
- Zhao, C., and Beers, E. P. (2013). Alternative splicing of Myb-related genes MYR1 and MYR2 may modulate activities through changes in dimerization, localization, or protein folding. *Plant Signal. Behav.* 8:e27325. doi: 10.4161/psb.27325
- Zhao, P. X., Miao, Z. Q., Zhang, J., Chen, S. Y., Liu, Q. Q., and Xiang, C. B. (2020). Arabidopsis MADS-box factor AGL16 negatively regulates drought resistance via stomatal density and stomatal movement. *J. Exp. Bot.* 71, 6092–6106. doi: 10.1093/jxb/eraa303

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Chakraborty, Viswanath, Malipatil, Rathore and Thirunavukkarasu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Bioinformatics Analysis of Actin Molecules: Why Quantity Does Not Translate Into Quality?

Anna V. Glyakina^{1,2} and Oxana V. Galzitskaya^{1,3*}

¹ Institute of Protein Research, Russian Academy of Sciences, Pushchino, Russia, ² Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Pushchino, Russia, ³ Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Russia

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Anastasia Anashkina,
Engelhardt Institute of Molecular
Biology (RAS), Russia
Nikolay Alemasov,
Russian Academy of Sciences, Russia

*Correspondence:

Oxana V. Galzitskaya
ogalzit@vega.protres.ru

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 15 October 2020

Accepted: 20 November 2020

Published: 10 December 2020

Citation:

Glyakina AV and Galzitskaya OV
(2020) Bioinformatics Analysis of Actin
Molecules: Why Quantity Does Not
Translate Into Quality?
Front. Genet. 11:617763.
doi: 10.3389/fgene.2020.617763

It is time to review all the available data and find the distinctive characteristics of actin that make it such an important cell molecule. The presented double-stranded organization of filamentous actin cannot explain the strong polymorphism of actin fibrils. In this work, we performed bioinformatics analysis of a set of 296 amino acid actin sequences from representatives of different classes of the Chordate type. Based on the results of the analysis, the degree of conservatism of the primary structure of this protein in representatives of the Chordate type was determined. In addition, 155 structures of rabbit actin obtained using X-ray diffraction analysis and electron microscopy have been analyzed over the past 30 years. From pairwise alignments and the calculation of root-mean-square deviations (RMSDs) for these structures, it follows that they are very similar to each other without correlation with the structure resolution and the reconstruction method: the RMSDs for 11,781 pairs did not exceed 3 Å. It turned out that in rabbit actin most of the charged amino acid residues are located inside the protein, which is not typical for the protein structure. We found that two of six exon regions correspond to structural subdomains. To test the double-stranded organization of the actin structure, it is necessary to use new approaches and new techniques, taking into account our new data obtained from the structural analysis of actin.

Keywords: monomer, filamentous, polymorphism, actin, resolution, charge

INTRODUCTION

Actin was discovered in 1948 by the Hungarian biochemist Bruno Straub. This protein was named for its ability to activate (hence actin) ATP hydrolysis catalyzed by myosin. Actin is a muscle tissue protein, the polymerized form of which (F-actin) forms microfilaments—one of the main components of the cytoskeleton of eukaryotic cells. Actin makes up 5–15% of the total cellular protein and is the most important protein in eukaryotic cells (Lodish et al., 2000). Actin analogs have also been found in bacteria (Popp et al., 2008, 2010; Galkin et al., 2009) and archaea (Izore et al., 2014; Braun et al., 2015). Actin monomer (G-actin) is a water-soluble globular structural protein with a molecular weight of 42 kDa, consisting of 375 or 374 amino acid residues. Differences in the amino acid sequences, both within the same species and between species, are extremely insignificant, no more than 25 amino acid substitutions. In vertebrates, depending on the isoelectric point, three actin isoforms are distinguished, α , β , and γ (Vandekerckhove and Weber, 1978).

α -actin is mainly characteristic of muscle cells, while β - and γ -actin are characteristic of non-muscle cells. α -actin, in turn, is divided into three types: smooth muscle α -actin, α -actin of skeletal muscle and α -actin of cardiac muscle (Vandekerckhove and Weber, 1978; Gunning et al., 1983; Erba et al., 1988; Miwa et al., 1991).

Domains have historically been divided into large and small, although their sizes are almost the same. The N- and C-termini of the polypeptide chain are located in a small domain. Each of the domains has two subdomains. By definition, subdomain 1 (residues 1–32, 70–144, and 338–372) and subdomain 2 (residues 33–69) are part of a small domain. The large domain consists of subdomain 3 (residues 145–180 and 270–337) and subdomain 4 (residues 181–269) (Selby and Bear, 1956). The domains are separated by a deep cleft (**Figure 1A**). The actin monomer is rather flat and fits into a “parallelepiped” with dimensions of $55 \text{ \AA} \times 55 \text{ \AA} \times 35 \text{ \AA}$.

Actin is one of the main components of myofibrils and, together with myosin and titin, provides muscle contraction. In other types of cells, actin forms a system of microfilaments and, together with other filamentous structures (microtubules and intermediate filaments), makes up the cytoskeleton and performs various functions (movement, cell reshaping, cytokinesis, exo- and endocytosis, redistribution of surface receptors, and other processes) (Carlier et al., 1997; Silacci et al., 2004; Kovar et al., 2006; Ferron et al., 2007).

The actin polymerization and depolymerization process is regulated by special proteins. For example, profilin, forming a complex with globular actin, prevents actin polymerization. Cytochalasin D binds to actin and forms a kind of “cap” at one end of the polymerizing actin, thereby regulating the polymerization process. There are proteins (latrunculin A) that prevent the polymerization of globular actin and proteins that “cut” actin filaments into short fragments. Conversely, there are proteins that “cross-link” already formed actin filaments, thus forming ordered rigid bundles of actin filaments or flexible coarse networks (Carlier et al., 1997; Silacci et al., 2004; Kovar et al., 2006; Ferron et al., 2007).

Actin monomers can interact with each other to form F-actin. The polymerization process can be initiated by increasing the concentration of cations or by adding special proteins. The polymerization process becomes possible because actin monomers can recognize each other and form intermolecular contacts. *In vitro*, at physiological salt concentrations, G-actin polymerizes into filamentous F-actin (Page et al., 1998; Oda et al., 2009; Dominguez and Holmes, 2011; Thomasson and Macnaughtan, 2013).

Historically examining all the structural work of actin organization to find direct evidence of double-stranded actin organization, we have concluded that there is no direct evidence for the existence of a double helix. The authors of the work on X-ray diffraction could not give an unambiguous answer about the double helix organization of actin (Selby and Bear, 1956), and only the authors of the work (Hanson and Lowy, 1963), based on electron microscopy data, taking into account the data of the previous X-ray work, said that actin most likely has a double helix organization. Apparently, such a conclusion

was nevertheless dictated by the recent discovery of the DNA double helix at that time. Over time, “the highly-likely model” became generally accepted and entered all textbooks (Jegou and Romet-Lemonne, 2020). A lot of data have been accumulated that is not suitable for the proposed organization. This model, on the one hand, cannot explain the strong polymorphism of filamentous actin, but, on the other hand, does not contradict the picture of interaction with partners, since the interaction sites are located on one side of the actin molecule. The aim of this work is to obtain more structural information about such important molecule as actin, to validate the structure of F-actin.

MATERIALS AND METHODS

Databases

A set of 296 amino acid sequences of actin from representatives of different classes of the Chordate type: mammals (*Homo sapiens*, *Bos taurus*, *Mus musculus*, *Rattus norvegicus*); aves (*Gallus gallus*, *Anas platyrhynchos*, *Meleagris gallopavo*); Reptiles (*Chelonia agassizi*, *Pelodiscus sinensis*, *Anolis carolinensis*); amphibian (*Xenopus tropicalis*); fish (*Danio rerio*, *Tetraodon nigroviridis*, *Oryzias latipes*) were taken from the UniProt database.

A list of 155 protein structures of actin was taken from the UniProtKB database, record number P68135, gene ACTA1, wild rabbit species (*Oryctolagus cuniculus*). These structures were deposited in the Protein Data Bank between 1991 and early 2020. Half of these structures (72) are actin monomeric structures, the other half (83) are structures containing two or more actin monomers.

The canonical reviewed protein sequences (20,364 fasta records) were extracted from the human reference proteome (uniprot request reviewed:yes AND organism: “Homo sapiens (Human) [9606]” AND proteome:up000005640).

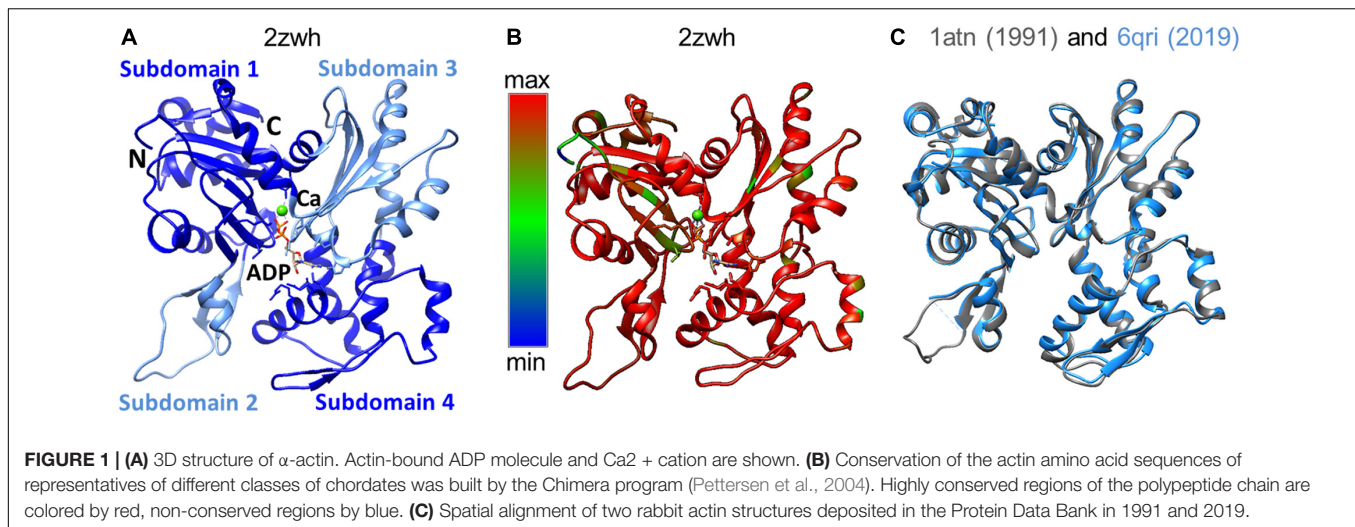
Structural Characteristics

Spatial alignment of 155 actin structures, calculation of the root-mean-square deviation (RMSD) of Ca atoms for each pair of superimposed structures, and calculation of the accessible surface area (ASA) for each amino acid residue in actin structures were performed using the YASARA program (Krieger et al., 2002). If pdb file had several actin structures, then only one (first) structure was taken for spatial alignment and RMSD calculation. ASA was defined as the surface area over which a water ball with a radius of $1.4 \text{ \AA}$ rolls. A residue was called external if its ASA was more than 50% of the maximum ASA observed in the Protein Data Bank for each type of amino acid residue.

RESULTS AND DISCUSSION

Conservation and Splicing Sites

The amino acid sequence of skeletal and cardiac muscle actin consists of 375 amino acid residues, including one unusual amino acid residue, 3-methylhistidine, which is formed post-translationally. The N-terminal amino acid of actin is acetylated.



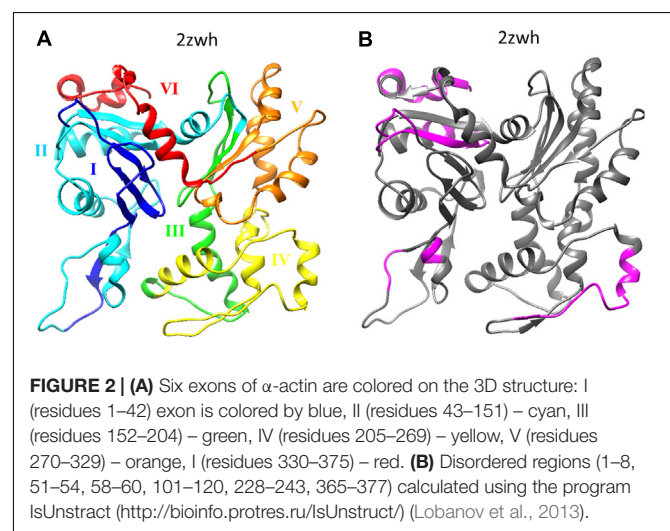
According to the results of the alignment of the actin amino acid sequences, a high degree of conservatism of the primary structure of actin was observed (**Figure 1B**). The spatial alignment of the two rabbit actin structures resolved in 1991 and 2019 is presented in **Figure 1C**. The RMSD for this pair is 0.6 Å, where the percent of amino acid residues aligned is 94%.

Cytoplasmic actins differ from vertebrate skeletal muscle actin only by 25 substitutions. It is essential that the region of the polypeptide chain containing residues 18–75 is stable, while regions 2–18 and 259–298 contain many substitutions. The high conservatism of the primary structure of actin, apparently, is a consequence of its high functional activity, which requires the preservation of the centers of interaction with both other actin molecules and actin-binding proteins. It should be noted that regions 18–25 and 259–298 do not belong to the F-actin core (Glyakina et al., 2020).

The data on the primary structure of actin in higher plant obtained on the basis of the analysis of nucleotide sequences of actin genes indicate that the variability of plant actin is much higher than that of animal actin. In particular, soybean isoactins contain 35–45 substitutions. In general, plant actin differs from animal actin by 55–65 amino acid residues. Actin substitutions in plants include significant number of charged residues; therefore, their isoelectric point can differ by almost one unit (pH 5.1–5.8) (Meagher and McLean, 1990).

Although more than 95% of the known protein sequences are derived from DNA translation, there is no single reference nucleic acid sequence for the given UniProtKB/Swiss-Prot protein sequence. To obtain splicing sites, we aligned the nucleotide sequence of human actin to the corresponding gene. There are only three substitutions in amino acid sequences between the human and rabbit actin sequences. Therefore, the splicing sites for the human gene will coincide with the splicing sites of the rabbit actin gene (**Supplementary Figure S1**).

If we compare the 3D actin structure with the splicing sites, we can see that exon IV (residues 205–269) is included in subdomain



4 (residues 181–269), and exon V (residues 270–329) is part of subdomain 3 (residues 270–337) (see **Figures 1A, 2A**).

Amino Acid Composition

To assess the increased and decreased content of one or another amino acid in the composition of α -actin from the skeletal muscles of rabbit, a comparison with the mean proteomic values of amino acid composition was carried out.

There is an increased content of amino acids such as Ile, Met, Thr, Tyr, and a decreased content of Cys, Leu, Gln compared to the average proteomic values for human, which are taken as a unit (**Supplementary Figure S2**).

Rabbit actin contains five cysteine residues (Cys10, Cys217, Cys257, Cys285, and Cys374). Smooth muscle actin contains another cysteine residue at position 17. Non-muscle actin contains two additional cysteine residues, Cys17 and Cys272. However, Cys10 is replaced by Val10 in actin, and the total number of cysteine residues is six. Only one of these residues, Cys374, is exposed on the surface of an intact ATP containing

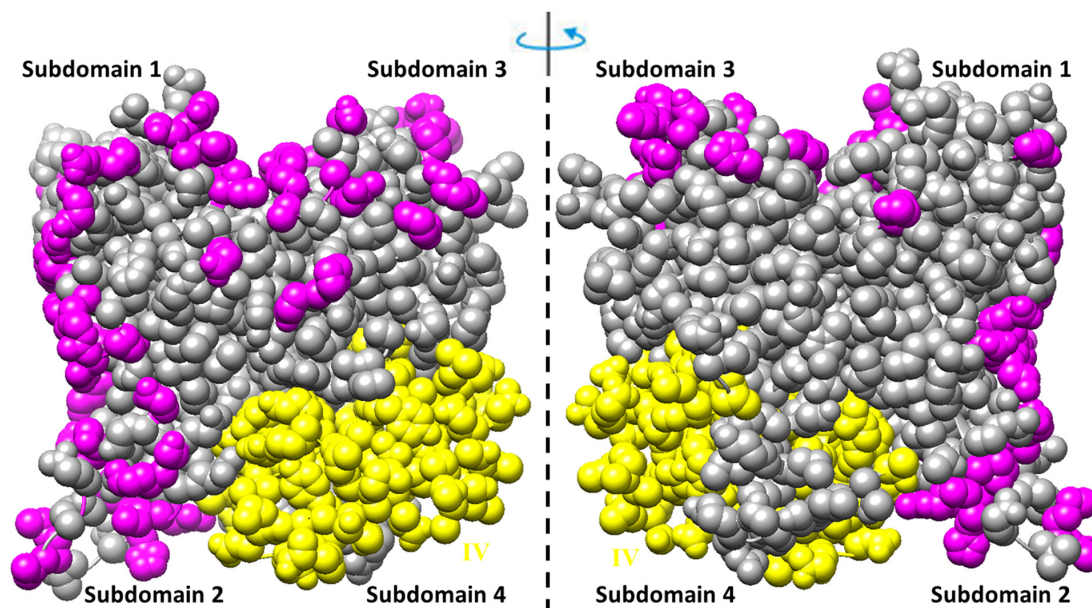


FIGURE 3 | Amino acid residues in the rabbit actin structure (2zwh) interacting with more than one of the 11 ligands (actin, profilin, gelsolin, DBP, cofilin, DNaseI, myosin, leiomodin, fimbrin, vinculin, tropomyosin) (Tikhomirova et al., 2018) are colored by magenta. Exon IV is colored by yellow.

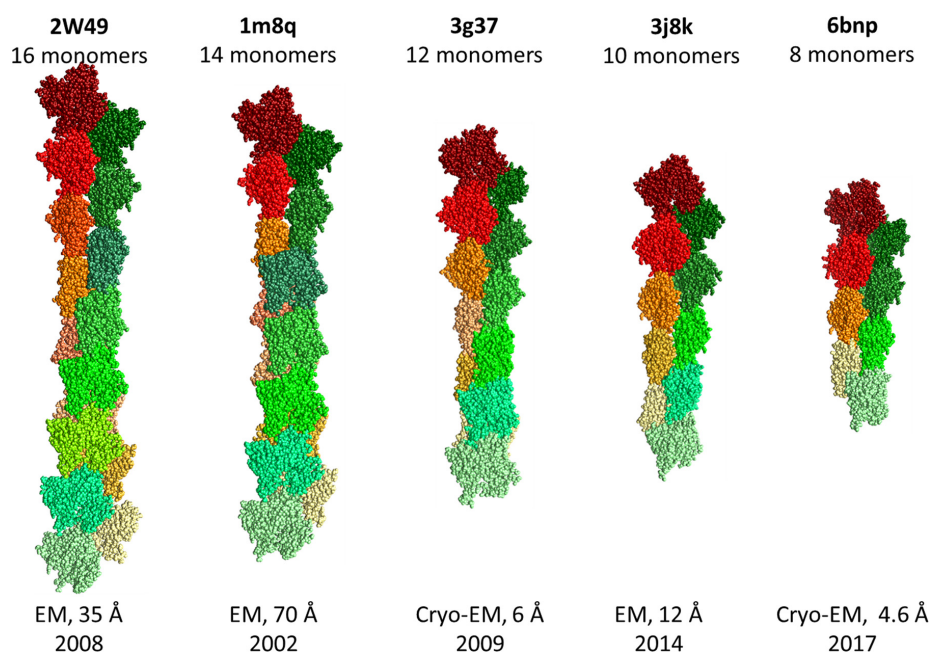


FIGURE 4 | Structures of F-actin from the Protein Data Bank.

G-actin molecule. The availability of other cysteine residues is determined by the degree of nativeness of the molecule and the type of nucleotide. In a solution with a high concentration of ADP, in addition to Cys374, another cysteine residue, apparently, Cys10, becomes available for SH-reagents. However, there are no disulfide bonds in the structure of rabbit actin.

Characteristics of Actin Surface: Distribution of Charges and Ligand Sites

The distribution of charged amino acid residues in rabbit actin is shown in **Supplementary Figure S3**. The charged amino acid residues make up 24%: the number of Arg is 18, Lys – 19, Glu – 28, Asp – 22. Therefore, the charge of the G-actin

molecule is negative (−13). The N-terminal segment of α -actin contains four acidic amino acid residues Asp-Glu-Asp-Glu, in the N-terminal segment of β - and γ -isoactins there are only three Asp and three Glu in β - and γ -isoactin, respectively. Substitutions of amino acid residues in the N-terminal segment of the polypeptide chain significantly affect the total charge of the molecule, changing the actin isoelectric point in the pH range 5.4–5.5. When analyzing the amino acid sequence of actin, attention is drawn to a large number of negatively charged groups, especially at the N-terminus of the chain. So, out of five N-terminal amino acid residues, four contain carboxyl groups in the side chains, and among the first 25 amino acid residues, seven are negatively charged.

Oztug Durer et al. (2010) try to investigate the constraints in D-loop (residues 39–50) plasticity as determined by its interactions with other dynamic elements of actin, including the C-terminus, the W-loop (residues 165–172), and the H-loop (residues 264–273). The involvement of these structural loops in contacts between monomers in the actin filament was predicted by the models of double-stranded organization of F-actin. The authors showed that introduction of disulfide bonds between residues 45, 47, 50 (D-loop), and residue 169 (W-loop) or 265 (H-loop) leads to the disruption of F-actin structure, which is expressed in the appearance of amorphous aggregates in the electron microscopy images (Oztug Durer et al., 2010). Consequently, the double-stranded model of F-actin maybe not the only one.

The fraction of charged amino acid residues of each type (Arg, Lys, Glu, Asp) in rabbit actin and bovine p450 (comparable to the size of actin, 475 amino acid residues) is shown in **Supplementary Table S1**. The fractions of charged amino acid residues in these proteins are practically the same and are close to the proteomic values (**Supplementary Table S1** and **Supplementary Figure S2**). However, most of the charged amino acid residues in rabbit actin are found within the protein as compared to bovine p450.

Actin is a unique building material widely used by the cell to construct various elements of the cytoskeleton and contractile apparatus. This is due to the fact that the processes of actin polymerization and depolymerization can be easily regulated using special proteins that bind to actin. Thus, actin is involved in many protein–protein interactions. The number of interaction partners for yeast actin is 222 according to the STRING database (Szklarczyk et al., 2015) (**Supplementary Figure S4**). There are no interaction partners for human actin in this database.

The residues involved in the interactions with more than one of the ligands (actin, profilin, gelsolin, DBP, cofilin, DNaseI, myosin, leiomodin, fimbrin, vinculin, tropomyosin) (Tikhomirova et al., 2018) are shown in **Figure 3**. Interestingly, that such residues are located predominantly on one side of the molecule and are not located in subdomain 4, which includes exon IV (residues 205–269) (see **Figures 2A, 3** and **Supplementary Figure S1**).

Structural Alignments

Of the 155 three-dimensional (3D) structures of rabbit actin, 72 are monomeric and 83 are oligomeric. 100 structures

were obtained by X-ray diffraction analysis with a resolution of 1.29–7.88 Å, 52 structures by cryo-electron microscopy with a resolution of 3.6–70 Å, and one structure by fiber diffraction with a resolution of 3.3 Å, and two model structures (1ALM and 1UY5).

We performed pairwise spatial alignments of 154 actin structures and calculated the RMSDs between the C α atoms. The RMSDs for 11,781 pairs do not exceed 3 Å: 0–1 Å for 5461 pairs, 1–2 Å for 5522 pairs, and 2–3 Å for 798 pairs (**Supplementary Figures S5A,C**). The fraction of aligned amino acid residues in each pair is more than 50% (see **Supplementary Figure S5B**). It should be noted that the 3D structures of the actin monomer are very similar. It turns out that for almost 30 years the quality of the obtained actin structures has not improved, despite new technologies.

It should be noted that the RMSD between monomers in the structures of filamentous actin does not exceed 1.3 Å (**Figure 4**). It is very strange that the RMSD is zero for monomeric structures 2W49 and 1M8Q, since the actin structure contains flexible/disordered regions that will add polymorphism to the structural organization of actin (**Figure 2B**) (Lobanov et al., 2013).

CONCLUSION

Bioinformatics analysis of actin showed that:

- (1) The amino acid sequences of actin in representatives of different classes of chordates are highly conservative;
- (2) Analysis of exons showed that exon IV (residues 205–269) corresponds to subdomain 4 (residues 181–269) and exon V (residues 270–329) corresponds to subdomain 3 (residues 270–337);
- (3) The 3D actin rabbit monomer structures resolved from 1991 to 2020 (during 30 years) are very similar: the RMSD for 11,781 pairs does not exceed 3 Å, the RMSD is about zero for monomeric structures in the filamentous actin (2W49 and 1M8Q);
- (4) Most of the charged amino acid residues are located within actin structure, which is unusual for a protein structure.

Due to the high polymorphism, it has not yet been possible to obtain the structure of filamentous actin using X-ray diffraction analysis. Thus, all hope for obtaining this structure is for new methods such as XFEL and cryo-electron microscopy. For further reconstruction of filamentous actin using cryo-electron microscopy, it is necessary to take into account the adjustment of monomers in the organization of filament, rather than simple copying of the “building block,” which gives a close to zero RMSD between monomers, which is observed for some filament structures from the Protein Data Bank.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

OG conceived and supervised the study. AG implemented the calculations, drew figures, and funded the manuscript. AG and OG analyzed the data and wrote the manuscript. Both authors read and approved the final version of the manuscript.

FUNDING

AG was funded by the Russian Science Foundation (Grant No. 19-74-10051).

REFERENCES

- Braun, T., Orlova, A., Valegård, K., Lindås, A.-C., Schröder, G. F., and Egelman, E. H. (2015). Archaeal actin from a hyperthermophile forms a single-stranded filament. *Proc. Natl. Acad. Sci. U.S.A.* 112, 9340–9345. doi: 10.1073/pnas.1509069112
- Carlier, M.-F., Laurent, V., Santolini, J., Melki, R., Didry, D., Xia, G.-X., et al. (1997). Actin depolymerizing factor (ADF/Cofilin) enhances the rate of filament turnover: implication in actin-based motility. *J. Cell Biol.* 136, 1307–1322. doi: 10.1083/jcb.136.6.1307
- Dominguez, R., and Holmes, K. C. (2011). Actin structure and function. *Annu. Rev. Biophys.* 40, 169–186. doi: 10.1146/annurev-biophys-042910-155359
- Erba, H. P., Eddy, R., Shows, T., Kedes, L., and Gunning, P. (1988). Structure, chromosome location, and expression of the human gamma-actin gene: differential evolution, location, and expression of the cytoskeletal beta- and gamma-actin genes. *Mol. Cell. Biol.* 8, 1775–1789. doi: 10.1128/mcb.8.4.1775
- Ferron, F., Rebowski, G., Lee, S. H., and Dominguez, R. (2007). Structural basis for the recruitment of profilin-actin complexes during filament elongation by Ena/VASP. *EMBO J.* 26, 4597–4606. doi: 10.1038/sj.emboj.7601874
- Galkin, V. E., Orlova, A., Rivera, C., Mullins, R. D., and Egelman, E. H. (2009). Structural polymorphism of the ParM filament and dynamic instability. *Struct. Lond. Engl.* 1993, 1253–1264. doi: 10.1016/j.str.2009.07.008
- Glyakina, A. V., Surin, A. K., Grishin, S. Yu., Selivanova, O. M., Suvorina Yu, M., Bobyleva, L. G., et al. (2020). New model for stacking monomers in filamentous actin from skeletal muscles of *Oryctolagus cuniculus*. *Int. J. Mol. Sci.* 21:8319. doi: 10.3390/ijms21218319
- Gunning, P., Ponte, P., Okayama, H., Engel, J., Blau, H., and Kedes, L. (1983). Isolation and characterization of full-length cDNA clones for human alpha-, beta-, and gamma-actin mRNAs: skeletal but not cytoplasmic actins have an amino-terminal cysteine that is subsequently removed. *Mol. Cell. Biol.* 3, 787–795. doi: 10.1128/mcb.3.5.787
- Hanson, J., and Lowy, J. (1963). The structure of F-actin and of actin filaments isolated from muscle. *J. Mol. Biol.* 6, 46–IN5. doi: 10.1016/S0022-2836(63)80081-9
- Izoré, T., Duman, R., Kureisaite-Ciziene, D., and Löwe, J. (2014). Crenactin from *Pyrobaculum calidifontis* is closely related to actin in structure and forms steep helical filaments. *FEBS Lett.* 588, 776–782. doi: 10.1016/j.febslet.2014.01.029
- Jegou, A., and Romet-Lemonne, G. (2020). The many implications of actin filament helicity. *Semin. Cell Dev. Biol.* 102, 65–72. doi: 10.1016/j.semcdb.2019.10.018
- Kovar, D. R., Harris, E. S., Mahaffy, R., Higgs, H. N., and Pollard, T. D. (2006). Control of the assembly of ATP- and ADP-actin by formins and profilin. *Cell* 124, 423–435. doi: 10.1016/j.cell.2005.11.038

ACKNOWLEDGMENTS

We are grateful to Roman Ievlev and Nikita Dovidchenko for assistance in preparation of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.617763/full#supplementary-material>

- Krieger, E., Koraimann, G., and Vriend, G. (2002). Increasing the precision of comparative models with YASARA NOVA—a self-parameterizing force field. *Proteins* 47, 393–402. doi: 10.1002/prot.10104
- Lobanov, M. Y., Sokolovskiy, I. V., and Galzitskaya, O. V. (2013). IsUnstruct: prediction of the residue status to be ordered or disordered in the protein chain by a method based on the Ising model. *J. Biomol. Struct. Dyn.* 31, 1034–1043. doi: 10.1080/07391102.2012.718529
- Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). *Molecular Cell Biology*, 4th Edn. New York, NY: W. H. Freeman.
- Meagher, R. B., and McLean, B. G. (1990). Diversity of plant actins. *Cell Motil.* 16, 164–166. doi: 10.1002/cm.970160303
- Miwa, T., Manabe, Y., Kurokawa, K., Kamada, S., Kanda, N., Bruns, G., et al. (1991). Structure, chromosome location, and expression of the human smooth muscle (enteric type) gamma-actin gene: evolution of six human actin genes. *Mol. Cell. Biol.* 11, 3296–3306. doi: 10.1128/mcb.11.6.3296
- Oda, T., Iwasa, M., Aihara, T., Maéda, Y., and Narita, A. (2009). The nature of the globular- to fibrous-actin transition. *Nature* 457, 441–445. doi: 10.1038/nature07685
- Oztug Durer, Z. A., Diraviyam, K., Sept, D., Kudryashov, D. S., and Reisler, E. (2010). F-actin structure destabilization and DNase I binding loop fluctuations. *J. Mol. Biol.* 395, 544–557. doi: 10.1016/j.jmb.2009.11.001
- Page, R., Lindberg, U., and Schutt, C. E. (1998). Domain motions in actin11Edited by D. Rees. *J. Mol. Biol.* 280, 463–474. doi: 10.1006/jmbi.1998.1879
- Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et al. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605–1612. doi: 10.1002/jcc.20084
- Popp, D., Narita, A., Ghoshdastider, U., Maeda, K., Maéda, Y., Oda, T., et al. (2010). Polymeric structures and dynamic properties of the bacterial actin Alfa. *J. Mol. Biol.* 397, 1031–1041. doi: 10.1016/j.jmb.2010.02.010
- Popp, D., Narita, A., Oda, T., Fujisawa, T., Matsuo, H., Nitanai, Y., et al. (2008). Molecular structure of the ParM polymer and the mechanism leading to its nucleotide-driven dynamic instability. *EMBO J.* 27, 570–579. doi: 10.1038/sj.emboj.7601978
- Selby, C. C., and Bear, R. S. (1956). The structure of actin-rich filaments of muscles according to x-ray diffraction. *J. Biophys. Biochem. Cytol.* 2, 71–85. doi: 10.1083/jcb.2.1.71
- Silacci, P., Mazzolai, L., Gauci, C., Stergiopoulos, N., Yin, H. L., and Hayoz, D. (2004). Gelsolin superfamily proteins: key regulators of cellular functions. *Cell. Mol. Life Sci. CMLS* 61, 2614–2623. doi: 10.1007/s00018-004-4225-6
- Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., et al. (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 43, D447–D452. doi: 10.1093/nar/gku1003

- Thomasson, M. S., and Macnaughtan, M. A. (2013). Microscopy basics and the study of actin–actin-binding protein interactions. *Anal. Biochem.* 443, 156–165. doi: 10.1016/j.ab.2013.09.008
- Tikhomirova, T. S., Ievlev, R. S., Suvorina, M. Yu, Bobyleva, L. G., Vikhlyantsev, I. M., Surin, A. K., et al. (2018). Search for functionally significant motifs and amino acid residues of actin. *Mol. Biol.* 52, 118–135. doi: 10.1134/S0026893318010193
- Vandekerckhove, J., and Weber, K. (1978). At least six different actins are expressed in a higher mammal: an analysis based on the amino acid sequence of the amino-terminal tryptic peptide. *J. Mol. Biol.* 126, 783–802. doi: 10.1016/0022-2836(78)90020-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Glyakina and Galzitskaya. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



NETO2 Is Deregulated in Breast, Prostate, and Colorectal Cancer and Participates in Cellular Signaling

Maria S. Fedorova^{1†}, Anastasiya V. Snezhkina^{1*†}, Anastasiya V. Lipatova¹, Vladislav S. Pavlov¹, Anastasiya A. Kobelyatskaya¹, Zulfiya G. Guvatova¹, Elena A. Pudova¹, Maria V. Savvateeva¹, Irina A. Ishina¹, Tatiana B. Demidova², Nadezhda N. Volchenko³, Dmitry Y. Trofimov⁴, Gennady T. Sukhikh⁴, George S. Krasnov¹ and Anna V. Kudryavtseva^{1*}

¹ Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, ² A. N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow, Russia, ³ National Medical Research Radiological Center, Ministry of Health of the Russian Federation, Moscow, Russia, ⁴ National Medical Research Center for Obstetrics, Gynecology and Perinatology Named After Academician V.I. Kulakov, Ministry of Health of the Russian Federation, Moscow, Russia

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Stanislav Rybtsov,
University of Edinburgh,
United Kingdom
Elvira Galieva,
Novosibirsk State University, Russia

*Correspondence:

Anastasiya V. Snezhkina
leftger@rambler.ru
Anna V. Kudryavtseva
rhizamoeba@mail.ru

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 14 August 2020

Accepted: 19 November 2020

Published: 10 December 2020

Citation:

Fedorova MS, Snezhkina AV,
Lipatova AV, Pavlov VS,
Kobelyatskaya AA, Guvatova ZG,
Pudova EA, Savvateeva MV, Ishina IA,
Demidova TB, Volchenko NN,
Trofimov DY, Sukhikh GT, Krasnov GS
and Kudryavtseva AV (2020) NETO2
Is Deregulated in Breast, Prostate,
and Colorectal Cancer
and Participates in Cellular Signaling.
Front. Genet. 11:594933.
doi: 10.3389/fgene.2020.594933

The *NETO2* gene (neuropilin and tolloid-like 2) encodes a protein that acts as an accessory subunit of kainate receptors and is predominantly expressed in the brain. Upregulation of *NETO2* has been observed in several tumors; however, its role in tumorigenesis remains unclear. In this study, we investigated *NETO2* expression in breast, prostate, and colorectal cancer using quantitative PCR (qPCR), as well as the effect of shRNA-mediated *NETO2* silencing on transcriptome changes in colorectal cancer cells. In the investigated tumors, we observed both increased and decreased *NETO2* mRNA levels, presenting no correlation with the main clinicopathological characteristics. In HCT116 cells, *NETO2* knockdown resulted in the differential expression of 17 genes and 2 long non-coding RNAs (lncRNAs), associated with the upregulation of circadian rhythm and downregulation of several cancer-associated pathways, including Wnt, transforming growth factor (TGF)- β , Janus kinase (JAK)-signal transducer and activator of transcription (STAT), mitogen-activated protein kinase (MAPK), and phosphatidylinositol 3-kinase (PI3K)/protein kinase B (AKT) pathways. Furthermore, we demonstrated the possibility to utilize a novel model organism, short-lived fish *Nothobranchius furzeri*, for evaluating *NETO2* functions. The ortholog *neto2b* in *N. furzeri* demonstrated a high similarity in nucleotide and amino acid sequences with human *NETO2*, as well as was characterized by stable expression in various fish tissues. Collectively, our findings demonstrate the deregulation of *NETO2* in the breast, prostate, and colorectal cancer and its participation in the tumor development primarily through cellular signaling.

Keywords: *NETO2*, breast cancer, prostate cancer, colorectal cancer, fish model

INTRODUCTION

NETO2 (neuropilin and tolloid-like 2) was first described in 2001 as a gene demonstrating significant similarity to *NETO1* (Stohr et al., 2002). Both genes encode putative transmembrane proteins containing two CUB domains, followed by a low-density lipoprotein class A (LDL_A) module in the extracellular region and conserved FXNPXY-like motif in the cytoplasmic region.

This structure of NETO proteins suggested their possible role in intracellular signaling pathways. Furthermore, a series of reports have shown the expression of *NETO2* in the brain, as well as its involvement in the modulation of most kainate receptors (KARs) (Straub et al., 2011) and N-methyl-D-aspartate (NMDA) receptors (Wyeth et al., 2014). Additionally, the interaction of NETO2 with other neuronal proteins, including glutamate receptor-interacting protein (GRIP) (Tang et al., 2012) and K(+)-Cl(−) cotransporter (KCC2), has been reported (Ivakine et al., 2013). Currently, NETO2 and its paralog NETO1 are widely recognized as the main auxiliary subunits of KARs.

More recently, *NETO2* was shown to be involved in carcinogenesis. Increased expression of the *NETO2* gene has been observed in proliferating infantile hemangiomas (Calicchio et al., 2009), hepatocellular (Villa et al., 2016) and nasopharyngeal carcinomas (He et al., 2019), as well as in renal (Snezhkina et al., 2018), lung (Oparina et al., 2012), colorectal (Hu et al., 2015; Fedorova et al., 2017), gastric (Liu et al., 2019), and pancreatic (Li et al., 2019) cancers. Hu et al. (2015) have revealed that *NETO2* overexpression significantly correlates with advanced tumor stage, invasion, and metastasis, as well as increases the risk of patient death in colorectal cancer. In patients with gastric and pancreatic cancers, a similar correlation was observed between *NETO2* expression and tumor progression and worse overall survival (Li et al., 2019; Liu et al., 2019). Reportedly, *NETO2* is related to five-gene transcriptomic signatures predicting rapidly growing tumors and survival in patients with hepatocellular carcinoma (Villa et al., 2016). Furthermore, *NETO2* has been observed in a macrophage-related gene signature predicting resistance to targeted therapeutics, as well as survival, in glioma patients (Sun et al., 2019). Additionally, *NETO2* expression is reportedly associated with the deletion of the *PTEN* (Phosphatase and tensin homolog) gene and high intratumor heterogeneity in prostate cancer (Yun et al., 2019).

In this study, we analyzed *NETO2* expression in prostate, breast, and colorectal cancer samples obtained from Russian patients. In these sample sets, both increased and decreased *NETO2* expression were observed, with no significant correlation between its mRNA levels and primary clinicopathological tumor characteristics. Additionally, we performed shRNA-mediated *NETO2* knockdown in the colorectal cancer cell line, HCT116, resulting in the differential expression of 17 genes and 2 long non-coding RNAs (lncRNAs), as well as deregulation of several cellular pathways, including tumor-associated Wnt, transforming growth factor (TGF)- β , Janus kinase (JAK)-signal transducer and activator of transcription (STAT), mitogen-activated protein kinase (MAPK), and phosphatidylinositol 3-kinase (PI3K)/protein kinase B (AKT) pathways. Moreover, we revealed a high similarity between human *NETO2* and its ortholog in fish *Nothobranchius furzeri*, which demonstrate the shortest captive lifespan for a vertebrate (3 months), a novel model for investigating aging and aging-related pathologies, including cancer (Terzibasi et al., 2007). Additionally, we evaluated gene expression in different tissues of *N. furzeri* using quantitative PCR (qPCR). These results can help to generate unique fishes with gene overexpression/downregulation in the target tissues to establish its role in tumorigenesis

[for example, using CRISPR activation (CRISPRa) or interference (CRISPRi) tools].

MATERIALS AND METHODS

Patients and Tumor Samples

In total, 74 colorectal, 40 prostate and 32 breast tumors, with paired adjacent normal tissues, were collected at the National Medical Research Radiological Center, Ministry of Health of the Russian Federation. No patient received chemotherapy, radiation, targeted therapy, and/or immunotherapy before surgery. Postoperative tumor and normal tissues were immediately frozen in liquid nitrogen and store at -80°C . This study was approved by the Ethics Committee of P.A. Hertsen Moscow Cancer Research Institute, Ministry of Health of the Russian Federation. All experiments were performed in strict accordance with the principles outlined in the Declaration of Helsinki (1964). The patients provided written informed consent to participate in this study. The clinicopathological characteristics of the patients and tumors are presented in **Supplementary Tables S1–S3**. Additionally, colorectal cancer samples were genetically characterized by mutations in *KRAS*, *NRAS*, and *BRAF*, as well as the microsatellite instability (MSI) status. Analysis of mutations in the “hot spots” of the *KRAS*, *NRAS*, and *BRAF* genes was performed in Evrogen (Russia) using the qPCR with following validation of the results with Sanger sequencing. The MSI detection was carried out in the same company using PCR amplification of microsatellite markers (*BAT25*, *BAT26*, *D2S123*, *D5S346*, and *D17S250*).

RNA Isolation and cDNA Synthesis

Tumor and normal tissues were homogenized in a lysis buffer using the MagNA Lyser Instrument (Roche, Switzerland). RNA was isolated from these tissues, as well as from cell cultures, using MagNA Pure Compact RNA Isolation Kit (Roche) on a MagNA Pure Compact System (Roche). The isolated RNA was quantified using a fluorometer, Qubit 2.0 (Thermo Fisher Scientific, United States). RNA quality was measured on an Agilent Bioanalyzer 2100 (Agilent Technologies, United States). Reverse transcription was performed from 1 mg of RNA using Mint Reagent Kit (Evrogen, Russia).

qPCR

Quantitative PCR was performed using TaqMan Gene Expression Assay (Thermo Fisher Scientific), primers and probes, for the *NETO2* gene (Hs00983152_m1). *GAPDH*, *GUSB*, and *B2M* were used as reference genes for prostate, colorectal, and breast cancer samples, respectively. Primers and probes for reference genes are shown in **Supplementary Table S4**. qPCR was performed on an Applied Biosystems 7500 Real-Time PCR System (Thermo Fisher Scientific) according to the scheme described in Krasnov et al. (2015). Each reaction was performed in triplicate. Data obtained by qPCR were analyzed using the ddCt method and the original ATG software (Kudryavtseva et al., 2018). *NETO2* expression was considered meaningful if at least a 2-fold

expression was determined owing to potential variations in the reference gene expression.

Cell Culture

The human colorectal cancer HCT116 cell line was cultured in DMEM (Dulbecco's Modified Eagle's medium; PanEco, Russia) supplemented with 10% FBS (fetal bovine serum; Thermo Fisher Scientific) and 100 units of penicillin/streptomycin (Thermo Fisher Scientific) in a humidified incubator at 37°C and 5% CO₂.

Cell Transfection

shRNA targeting the nucleotide residues 509–529 of the human *NETO2* gene (NCBI Gene ID: 81831, location 16q12.1) protein-coding region was designed and synthesized as follows: Sense, 5'-GATCCGCGCCAAATTATCCTGACTCATCACGTGATGAGTCAGGATAATTTGGCGTTTTT-3', and antisense, 5'-AATTC AAAACGCCAAATTATCCTGACTCATCACGTGATGAGTCAGGATAATTTGGCGCG-3'. This pair of oligonucleotides and non-specific control scrambled (SCR) shRNA were cloned into BamH and EcoR sites of pLSLP plasmid to establish the pLSLP-anti-*NETO2* lentiviral vector. Lentiviral constructs were transfected into 293T cells using Lipofectamine LTX reagent (Thermo Fisher Scientific). Virus-containing supernatants were collected 24 h after the transfection and were used to infect target HCT116 cells in triple repeats with the addition of 5 µg/mL polybrene (Sigma, United States). Infected cells were selected using the regular growth medium containing 1 µg/mL puromycin (Sigma).

Western Blot Analysis

Western blot analysis was performed as previously described (Kudryavtseva et al., 2016). For *NETO2* detection, the primary recombinant antibody EPR3497 (Abcam, United States) was used.

Transcriptome Sequencing and Analysis

For cDNA library preparation, we utilized RNA isolated from cell cultures with RIN (RNA integrity number) values ≥ 8 . Libraries were prepared with TruSeq Stranded mRNA Library Prep Kit (Illumina, United States) according to the manufacturer's recommendations. Transcriptome sequencing was performed on a NextSeq 500 System (Illumina) using 76 base pair single-end reads. At least 30M reads were obtained for each sample. The primary analysis of raw sequences was performed as previously described (Pudova et al., 2019). Then, data as transcripts per million (TPM) were imported in the R environment. Differential gene expression was analyzed using the DESeq2 package of Love et al. (2014). The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was used for pathway enrichment analysis.

Analysis of Orthologs

Nucleotide and amino acid sequences of *NETO2* in human samples and fishes were aligned using Blastn and Blastp, respectively. Gene expression data were obtained from the Bgee database¹.

¹<https://bgee.org/>

Neto2b Expression Analysis in *N. furzeri*

The eggs of fish *N. furzeri* GRZ were obtained from a commercial supplier and were bred in the Aquatic Housing System (Aquaneering, United States) at the Center for Precision Genome Editing and Genetic Technologies for Biomedicine at the Engelhardt Institute of Molecular Biology. The study was approved by the Ethics Committee of the A.N. Severtsov Institute of Ecology and Evolution Russian Academy of Sciences (approval no. 27, 9.10.2019).

Eight tissue samples (brain, intestines, heart, head kidney, liver, stomach, muscles, and skin) from a female *N. furzeri* were subjected to RNA isolation using QIAzol lysis reagent (Qiagen, Germany), subsequently treated with DNase I (Thermo Fisher Scientific). Quantification of the isolated RNA was performed using the NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific). Reverse transcription was carried out with 500 ng of RNA using RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific). A pair of primers (forward: CCACCCAACAAGGAGTGTGT and reverse: CCCGTGGAGGTAACAAGACC) was designed for *neto2b* gene detection by qPCR, performed on the Rotor-Gene Q 5 plex HRM (Qiagen) using the following scheme: 95°C for 10 min, 40 cycles of 95°C for 15 s, and 62°C for 60 s.

Statistical Analysis

The significance of differences observed between two groups (tumors/normal tissues and control cells/treated cells) was assessed using the non-parametric Mann-Whitney *U* test. Correlation analysis was performed using Spearman's rank correlation coefficient (r_s). For transcriptome analysis, the Benjamini-Hochberg method was used to calculate the adjusted *p*-values [the false discovery rate (FDR)]. Differences and correlations were considered significant at a *p*-value of < 0.05 .

RESULTS

NETO2 Expression in Breast, Prostate, and Colorectal Cancer

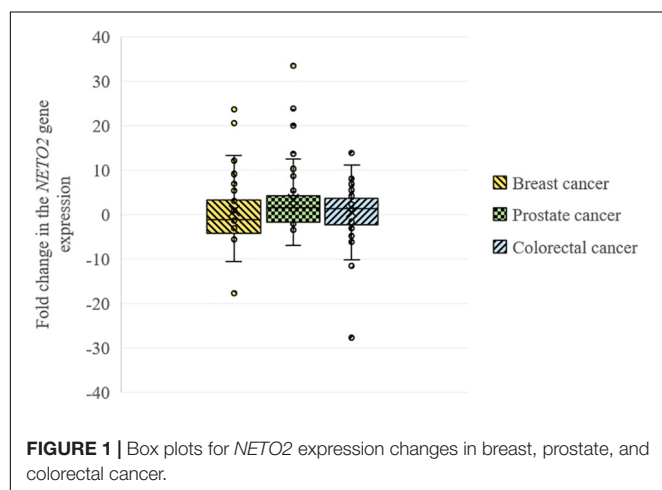
Using qPCR, we analyzed the relative *NETO2* mRNA level in sets of breast ($n = 32$), prostate ($n = 40$), and colorectal ($n = 74$) cancer samples. In breast cancer, *NETO2* gene expression was increased in 31% of cases (2–24-fold), with decreased mRNA levels detected in 44% of samples (2–18-fold) (Table 1). *NETO2* expression was characterized by a 2–33-fold average increase in 40% of prostate cancer samples and a 2–7-fold average decrease in 22.5% of cases. In colorectal cancer, *NETO2* mRNA levels were increased in 35% (2–14-fold) of investigated samples and decreased in 26% of cases (2–28-fold). The highest median value of altered gene expression was determined in prostate cancer (Table 1). In general, a similar trend of *NETO2* expression was observed in investigated cancer types (Figure 1).

Furthermore, we performed a correlation analysis between several clinicopathological characteristics of tumors (stage for all cancers, differentiation for prostate and colorectal cancer, mutation, and MSI status for colorectal cancer) and *NETO2*

TABLE 1 | Frequency of mRNA level changes in breast, prostate, and colorectal cancer.

Tumor	Frequency of mRNA level changes, % (n)*		Median of mRNA level changes, n-fold
	Increase	Decrease	
Breast cancer	31 (10/32)	44 (14/32)	1
Prostate cancer	40 (16/40)	22.5 (9/40)	1.8
Colorectal cancer	35 (26/74)	26 (19/74)	1.2

*p-value <0.05.

**FIGURE 1** | Box plots for *NETO2* expression changes in breast, prostate, and colorectal cancer.

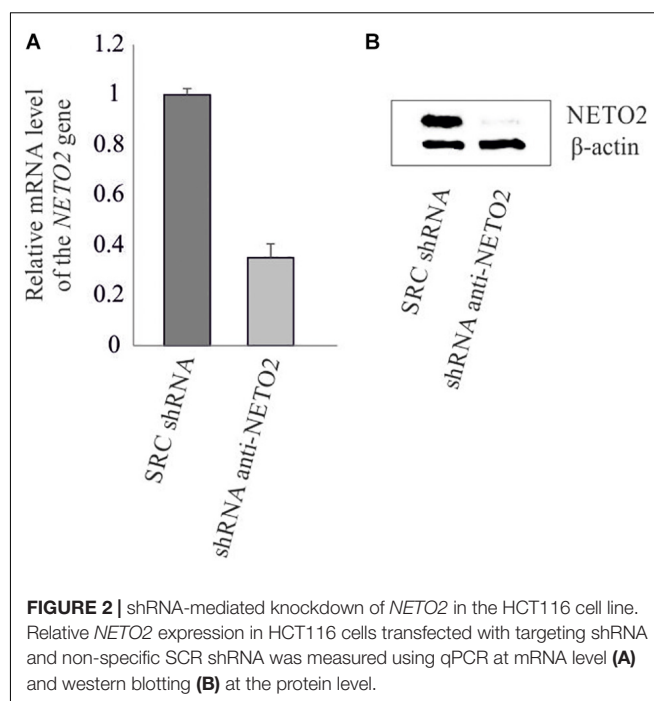
expression. The highest correlation coefficients were observed between *NETO2* expression and pathological stage for breast cancer ($r_s = 0.15$) and prostate cancer ($r_s = 0.25$), tumor differentiation for prostate cancer ($r_s = -0.16$), and *BRAF* mutation status ($r_s = 0.18$) for colorectal cancer. Correlation coefficients of less than 0.1 were observed between the *NETO2* mRNA level and other investigated clinicopathological characteristics.

Stable Knockdown of *NETO2* in HCT116

The pLSPV lentiviral vector expressing shRNA targeting *NETO2* (pLSPV-anti-*NETO2*) and the vector expressing non-targeting SCR shRNA were transfected into HCT116 in triplicate. Protein and mRNA levels of the *NETO2* gene were measured in experimental and control HCT116 cells using western blotting and qPCR (Figure 2). The results obtained confirmed the *NETO2* knockdown.

Effects of *NETO2* Knockdown on Gene Expression

Seventeen protein-coding differently expressed genes (DEGs) and two lncRNAs (DE lncRNAs) were revealed through the RNA-Seq analysis ($\text{Log}_2\text{FC} \geq 1$, $\text{Log}_2\text{FC} \leq -1$, $\text{CPM} \geq 1$, $\text{FDR} < 0.05$) in treated HCT116 cells when compared with the control group. We detected the upregulation of eight genes (*CYP2U1*, *NR1D1*, *ZNF804A*, *SEMA3C*, *DBP*, *NT5E*, *NAV3*, and *KDM5D*) and one lncRNA (*LINC02043*), as well as the

**FIGURE 2** | shRNA-mediated knockdown of *NETO2* in the HCT116 cell line. Relative *NETO2* expression in HCT116 cells transfected with targeting shRNA and non-specific SCR shRNA was measured using qPCR at mRNA level (A) and western blotting (B) at the protein level.

downregulation of nine genes (*MUC16*, *KRT6A*, *ACVRL1*, *PTP4A1*, *HYAL1*, *CLIC3*, *TEX19*, *DES*, and *KCNK3*) and one lncRNAs (*AL355075.4*) (Figure 3A).

Additionally, we performed KEGG pathway enrichment analysis using all DEGs with $\text{FDR} < 0.05$. Three significantly enriched pathways for downregulated genes and one pathway for upregulated genes were identified (Figure 3B). These included “Wnt signaling pathway,” “TGF-beta signaling pathway,” “signaling pathways regulating pluripotency of stem cells,” and “circadian rhythm.”

NETO2 Gene Ortholog in Fish *N. furzeri*

We analyzed the *NETO2* ortholog in the short-lived fish *N. furzeri*, termed *neto2b* (NCBI Gene ID: 107381994, location sgr07). Alignment of cDNA sequences demonstrated 69% of identity, whereas amino acid sequence alignment revealed higher similarity with 85% of conservative substitutions (positives) and only 2% of gaps (Supplementary Figure S1). In comparison, 71% of nucleotide sequences (cDNA) and 74% of amino acid sequence (6% of gaps) similarities were observed in *Danio rerio* (Supplementary Figure S2). Using qPCR, we measured *neto2b* expression in different tissues of *N. furzeri* (female) to estimate its potential use as a fish model presenting gene tissue-specific overexpression/downregulation. As observed in humans, *neto2b* demonstrated the highest mRNA levels in the fish brain. The expression of *neto2b* in other investigated tissues is presented in Figure 4.

DISCUSSION

A series of studies aimed at investigating the biological functions of *NETO2* has reported its crucial role in neural glutamate

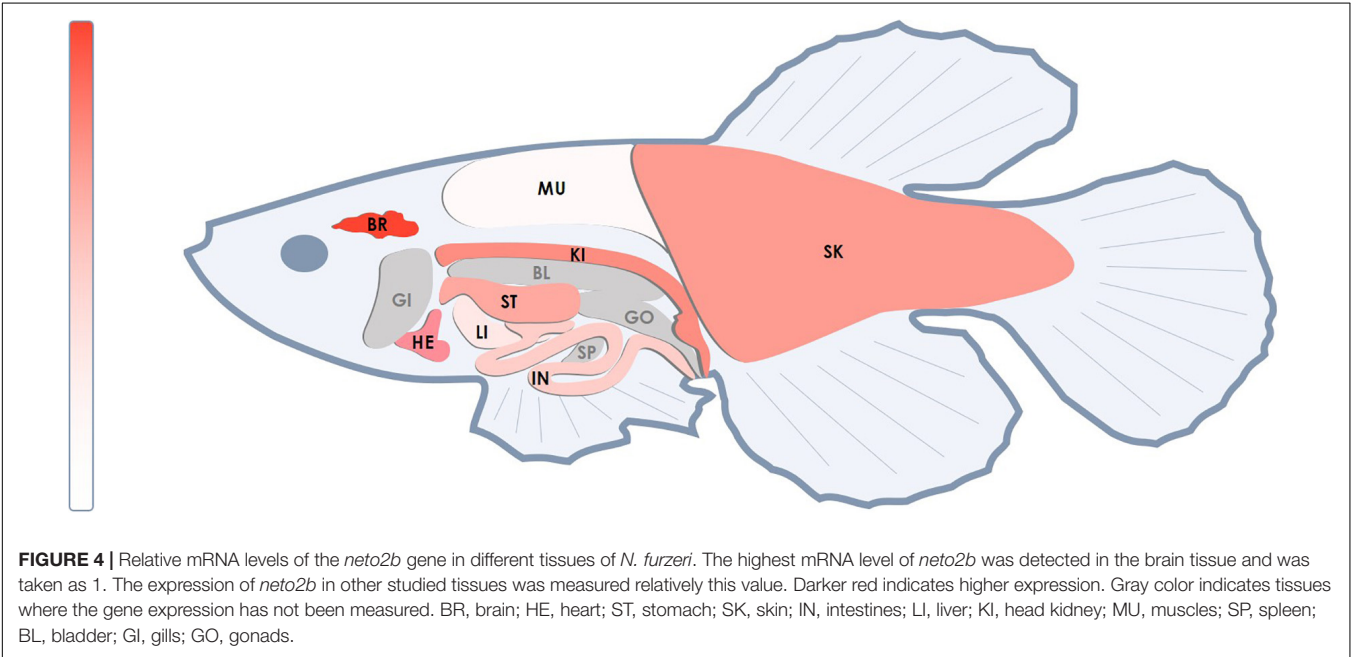
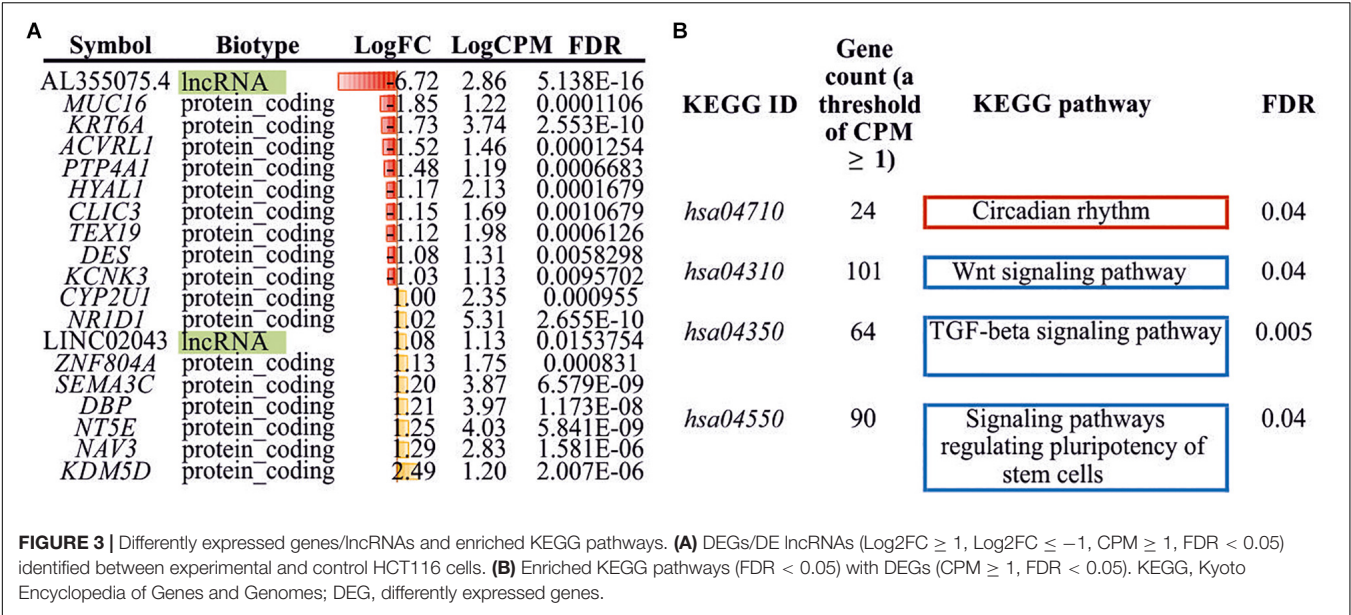


FIGURE 4 | Relative mRNA levels of the *neto2b* gene in different tissues of *N. furzeri*. The highest mRNA level of *neto2b* was detected in the brain tissue and was taken as 1. The expression of *neto2b* in other studied tissues was measured relatively this value. Darker red indicates higher expression. Gray color indicates tissues where the gene expression has not been measured. BR, brain; HE, heart; ST, stomach; SK, skin; IN, intestines; LI, liver; KI, head kidney; MU, muscles; SP, spleen; BL, bladder; GI, gills; GO, gonads.

signaling. However, *NETO2* was also found to be upregulated in several tumors. Our research group was one of the first to reveal *NETO2* overexpression in solid tumors (renal and lung cancer) (Oparina et al., 2012). Recently, it has been reported that elevated *NETO2* expression was associated with tumor progression, poor prognosis, and reduced survival in cancer patients (Hu et al., 2015; He et al., 2019). In this study, we first demonstrated the deregulation of *NETO2* expression in breast and prostate cancer. The *NETO2* mRNA level was predominantly downregulated in breast cancer (44%), with elevated gene expression observed in 31% of samples. In prostate cancer, *NETO2* mRNA levels were upregulated in 40% of cases, with a 1.8 median change. Negligible correlations were observed between the *NETO2* expression and stage of breast and prostate cancer, as well as differentiation of prostate tumors. For colorectal cancer, we expanded a set of samples from a previous study (Fedorova et al., 2017) and confirmed that *NETO2* expression was increased in about one-third of these tumors (41 and 35% in previous and current studies, respectively). Nevertheless, a study evaluating Chinese patients with colorectal cancer has revealed the upregulation of *NETO2* in 52.6% of cases, reporting an association between expression and advanced tumor stage and invasion, poor differentiation, lymph node metastasis, and unfavorable prognosis in patients (Hu et al., 2015). In contrast, we

failed to observe any significant correlations between the *NETO2* mRNA level and tumor stage, as well as differentiation that can be explained by the difference of the studied cohorts. Also, no association of *NETO2* expression with *KRAS*, *NRAS*, and *BRAF* mutations or MSI status were found.

Despite several studies indicating the involvement of *NETO2* in tumorigenesis, the underlying mechanism remains unclear. Li et al. (2019) have investigated *NETO2* functions in pancreatic cancer, and demonstrated that *NETO2* knockdown reduced the proliferative capacity of pancreatic cancer cells and suppressed tumor growth *in vivo*; *NETO2* overexpression conversely stimulated cell proliferation, invasion, and migration via the activation of the STAT3 pathway. Similar observations were reportedly documented following *NETO2* knockdown in nasopharyngeal carcinoma cells (He et al., 2019). The depletion of *NETO2* expression results in decreased proliferation, invasion, and migration of tumor cells and induced apoptosis via activating Caspase-3 signaling. Additionally, *NETO2* overexpression promotes the invasion and metastasis of gastric cancer cells by inducing epithelial-mesenchymal transition (EMT) by upregulating TNFRSF12A, which mediates the activation of the PI3K/AKT/NF- κ B/Snail axis (Liu et al., 2019). In the present work, we demonstrated that stable *NETO2* knockdown in HCT116 cells resulted in the downregulation of Wnt and TGF- β signaling pathways, as well as signaling pathways regulating the pluripotency of stem cells. In the KEGG database, the “signaling pathways regulating pluripotency of stem cells” group comprises the Jak-STAT, MAPK, PI3K-Akt, Wnt, and TGF- β signaling pathways. All these stem cell-related pathways are extensively implicated in tumorigenesis (Dreesen and Brivanlou, 2007). Thus, our results demonstrated the participation of *NETO2* in the deregulation of cell signaling in tumors and confirmed its relation to STAT and PI3K-Akt signaling as previously reported. Conversely, *NETO2* reduction leads to an upregulation of the circadian rhythm pathway in colorectal cancer cells. Circadian rhythm is closely associated with the cell cycle and has been implicated in DNA-damage response (Hunt and Sassone-Corsi, 2007), with the disruption of this pathway considered a risk factor for carcinogenesis (Straif et al., 2007). According to literature, *NETO2* is related to cell proliferation and apoptosis in tumor cells (He et al., 2019; Li et al., 2019). These processes could be connection points between the functions of *NETO2* and the circadian rhythm pathway.

Furthermore, we determined 17 genes presenting more than a 2-fold change in expression following *NETO2* knockdown. Increased expression was observed in genes involved in cell metabolism [*CYP2U1* (arachidonic acid metabolism) and *NT5E* (metabolism of nucleotides)], circadian rhythm (*NR1D1* and *DBP*), regulation of developmental processes (*SEMA3C* and *NAV3*), histone demethylation (*KDM5D*), and transcriptional regulation (*ZNF804A*). Decreased mRNA levels were noted for genes related to cell adhesion (*MUC16*), differentiation (*TEX19*), proliferation, and migration (*ACVRL1*, *PTP4A1*, and *HYAL1*), as well as cytoskeleton organization (*KRT6A* and *DES*) and ion transmembrane transport (*CLIC3* and *KCNK3*). Most genes (*CYP2U1*, *SEMA3C*, *NT5E*, *KRT6A*, *NAV3*, *ACVRL1*, *KCNK3*, *MUC16*, *PTP4A1*, and *HYAL1*) were previously found to be

implicated in the development of colorectal cancer. Interestingly, *NETO2* is reportedly associated with the neuron navigator 3 (*NAV3*) gene, which is also predominantly expressed in the nervous tissue. Copy number changes in the *NAV3* gene have been observed in colorectal cancer (Carlsson et al., 2012). Moreover, *NAV3* is reportedly involved in the p73-mediated tumor suppression, as well as in Jak-STAT and GnRH signaling in colorectal cancer cells (Carlsson et al., 2012; Uboveja et al., 2020). Furthermore, the participation of *NETO2* has been revealed in the Jak-STAT signaling pathway during cancer. Additionally, *NETO2* silencing resulted in the altered expression of two lncRNAs, LINC02043, and AL355075.4, for which the functions and related cellular pathways have not been established. However, lncRNA LINC02043 was previously reported as a prognostic marker of recurrence-free survival in hepatocellular carcinoma (Luo et al., 2020). Notably, regulation of *NETO2* expression with microRNAs and microRNA-lncRNA interaction has been recently shown in esophageal cancer. *NETO2* was determined as a target of miR-206 and miR-143-5p, and participation of *NETO2* tumor progression and angiogenesis through overexpression of lncRNA FAM225A absorbed miR-206 was revealed (Wada et al., 2020; Zhang et al., 2020). LncRNAs found in our study can also be potential regulators of *NETO2* expression in colorectal cancer.

Although a wide range of studies has proposed increased *NETO2* expression in several tumors, we revealed its deregulation in breast, prostate, and colorectal cancers (both decreased and increased mRNA levels). The function of *NETO2* in tumorigenesis remains unclear; however, our findings and those obtained in previous studies indicate its participation in cancer-related signaling pathways. Moreover, all identified cancer-related signaling pathways (Wnt, TGF- β , STAT, MAPK, and PI3K-Akt) have previously been shown to be associated with breast, prostate, and colorectal cancer (Dhillon et al., 2007; Massagué, 2008; Pencik et al., 2016; Zhan et al., 2016; Jiang et al., 2020).

Finally, we analyzed the ortholog of the human *NETO2* gene in the unique animal model, the short-lived fish *N. furzeri*. This fish is characterized by an extremely short captive lifespan of 3 months. The use of *N. furzeri* for genetic studies allows for the rapid generation of transgenic lines and short-term experiments. We showed that the ortholog *neto2b* had the high similarity with human *NETO2* in nucleotide (69%) and amino acid (85%) sequences (even more than those for *D. rerio*), as well as stable expression in the majority of the fish tissues. These results confirm that *NETO2* is a very conservative gene among vertebrates and indicate the possibility to use *N. furzeri* for the study of the gene function and its role in tumorigenesis at the organism level.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://www.ncbi.nlm.nih.gov/sra/PRJNA655834>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of P.A. Hertsen Moscow Cancer Research Institute, Ministry of Health of the Russian Federation. The patients/participants provided their written informed consent to participate in this study. The animal study was reviewed and approved by the Bioethics Commission from the A. N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences.

AUTHOR CONTRIBUTIONS

AS and AVK conceived and designed the work. NV clinically characterized tumor tissue samples. AL designed and maintained knockout cell line. MF, VP, and AAK performed the qPCR experiments. ZG, EP, MF, II, and MS performed the RNA isolation, library preparation, and western blot analysis. GK, DT, and VP carried out bioinformatics analysis. TD performed experimental procedures with the fish. AS, GS, and AVK wrote the manuscript. All authors read and approved the final manuscript.

REFERENCES

- Calicchio, M. L., Collins, T., and Kozakewich, H. P. (2009). Identification of signaling systems in proliferating and involuting phase infantile hemangiomas by genome-wide transcriptional profiling. *Am. J. Pathol.* 174, 1638–1649. doi: 10.2353/ajpath.2009.080517
- Carlsson, E., Ranki, A., Sipilä, L., Karenko, L., Abdel-Rahman, W. M., Ovaska, K., et al. (2012). Potential role of a navigator gene NAV3 in colorectal cancer. *Br. J. Cancer* 106, 517–524. doi: 10.1038/bjc.2011.553
- Dhillon, A. S., Hagan, S., Rath, O., and Kolch, W. (2007). MAP kinase signalling pathways in cancer. *Oncogene* 26, 3279–3290. doi: 10.1038/sj.onc.1210421
- Dreesen, O., and Brivanlou, A. H. (2007). Signaling pathways in cancer and embryonic stem cells. *Stem Cell Rev.* 3, 7–17.
- Fedorova, M. S., Snezhkina, A. V., Pudova, E. A., Abramov, I. S., Lipatova, A. V., Kharitonov, S. L., et al. (2017). Upregulation of NETO2 gene in colorectal cancer. *BMC Genet.* 18(Suppl. 1):117. doi: 10.1186/s12863-017-0581-8
- He, A. R., Zhu, Q., and Gao, S. (2019). Reducing NETO2 expression prevents human nasopharyngeal carcinoma (NPC) progression by suppressing metastasis and inducing apoptosis. *Biochem. Biophys. Res. Commun.* 513, 494–501. doi: 10.1016/j.bbrc.2019.03.061
- Hu, L., Chen, H. Y., Cai, J., Yang, G. Z., Feng, D., Zhai, Y. X., et al. (2015). Upregulation of NETO2 expression correlates with tumor progression and poor prognosis in colorectal carcinoma. *BMC Cancer* 15:1006. doi: 10.1186/s12885-015-2018-y
- Hunt, T., and Sassone-Corsi, P. (2007). Riding tandem: circadian clocks and the cell cycle. *Cell* 129, 461–464. doi: 10.1016/j.cell.2007.04.015
- Ivakine, E. A., Acton, B. A., Mahadevan, V., Ormond, J., Tang, M., Pressey, J. C., et al. (2013). Neto2 is a KCC2 interacting protein required for neuronal Cl⁻ regulation in hippocampal neurons. *Proc. Natl. Acad. Sci. U.S.A.* 110, 3561–3566. doi: 10.1073/pnas.1212907110
- Jiang, N., Dai, Q., Su, X., Fu, J., Feng, X., and Peng, J. (2020). Role of PI3K/AKT pathway in cancer: the framework of malignant behavior. *Mol. Biol. Rep.* 47, 4587–4629. doi: 10.1007/s11033-020-05435-1
- Krasnov, G. S., Dmitriev, A. A., Sadritdinova, A. F., Fedorova, M. S., Snezhkina, A. V., Melnikova, N. V., et al. (2015). [Evaluation of gene expression of

FUNDING

This work was supported by Grant 075-15-2019-1660 from the Ministry of Science and Higher Education of the Russian Federation.

ACKNOWLEDGMENTS

We thank the National Medical Research Radiological Center, Ministry of Health of the Russian Federation for tissue samples, National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I. Kulakov, Ministry of Health of the Russian Federation, and A. N. Severtsov Institute of Ecology and Evolution for assistance in the data analysis. This work was performed using the equipment(s) at EIMB RAS “Genome” center (http://www.eimb.ru/rus/ckp/ccu_genome_c.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.594933/full#supplementary-material>

- Hexokinases in colorectal cancer with the use of bioinformatics methods]. *Biofizika* 60, 1050–1056.
- Kudryavtseva, A. V., Fedorova, M. S., Zhavoronkov, A., Moskalev, A. A., Zasedatelev, A. S., Dmitriev, A. A., et al. (2016). Effect of lentivirus-mediated shRNA inactivation of HK1, HK2, and HK3 genes in colorectal cancer and melanoma cells. *BMC Genet.* 17(Suppl. 3):156. doi: 10.1186/s12863-016-0459-1
- Kudryavtseva, A. V., Nyushko, K. M., Zaretsky, A. R., Shagin, D. A., Sadritdinova, A. F., Fedorova, M. S., et al. (2018). [Suppression of NR0B2 gene in clear cell renal cell carcinoma is associated with Hypermethylation of Its promoter]. *Mol. Biol.* 52, 482–488. doi: 10.7868/S0026898418030114
- Li, Y., Zhang, Y., and Liu, J. (2019). NETO2 promotes pancreatic cancer cell proliferation, invasion and migration via activation of the STAT3 signaling pathway. *Cancer Manag. Res.* 11, 5147–5156. doi: 10.2147/CMAR.S204260
- Liu, J. Y., Jiang, L., He, T., Liu, J. J., Fan, J. Y., Xu, X. H., et al. (2019). NETO2 promotes invasion and metastasis of gastric cancer cells via activation of PI3K/Akt/NF-kappaB/Snail axis and predicts outcome of the patients. *Cell Death Dis.* 10:162. doi: 10.1038/s41419-019-1388-5
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8
- Luo, Y., Ye, J., Wei, J., Zhang, J., and Li, Y. (2020). Long noncoding RNAbased risk scoring system predicts prognosis of alcoholrelated hepatocellular carcinoma. *Mol. Med. Rep.* 22, 997–1007. doi: 10.3892/mmr.2020.11179
- Massagué, J. (2008). TGFβ in cancer. *Cell* 134, 215–230. doi: 10.1016/j.cell.2008.07.001
- Oparina, N. Y., Sadritdinova, A. F., Snezhkina, A. V., Dmitriev, A. A., Krasnov, G. S., Senchenko, V. N., et al. (2012). Increase in NETO2 gene expression is a potential molecular genetic marker in renal and lung cancers. *Russ. J. Genet.* 48, 506–512. doi: 10.1134/S1022795412050171
- Pencik, J., Pham, H. T. T., Schmoeller, J., Javaheri, T., Schleder, M., Culig, Z., et al. (2016). JAK-STAT signaling in cancer: from cytokines to non-coding genome. *Cytokine* 87, 26–36. doi: 10.1016/j.cyt.2016.06.017
- Pudova, E. A., Lukyanova, E. N., Nyushko, K. M., Mikhaylenko, D. S., Zaretsky, A. R., Snezhkina, A. V., et al. (2019). Differentially expressed genes associated

- with prognosis in locally advanced lymph node-negative prostate cancer. *Front. Genet.* 10:730. doi: 10.3389/fgene.2019.00730
- Snezhkina, A. V., Nyushko, K. M., Zaretsky, A. R., Shagin, D. A., Sadritdinova, A. F., Fedorova, M. S., et al. (2018). Transcription factor SAP30 is involved in the activation of *NETO2* gene expression in clear cell renal cell carcinoma. *Mol. Biol.* 52, 385–392. doi: 10.7868/S0026898418030072
- Stohr, H., Berger, C., Frohlich, S., and Weber, B. H. (2002). A novel gene encoding a putative transmembrane protein with two extracellular CUB domains and a low-density lipoprotein class A module: isolation of alternatively spliced isoforms in retina and brain. *Gene* 286, 223–231. doi: 10.1016/S0378-1119(02)00438-9
- Straif, K., Baan, R., Grosse, Y., Secretan, B., El Ghissassi, F., Bouvard, V., et al. (2007). Carcinogenicity of shift-work, painting, and fire-fighting. *Lancet Oncol.* 8, 1065–1066. doi: 10.1016/S1470-2045(07)70373-X
- Straub, C., Zhang, W., and Howe, J. R. (2011). *Neto2* modulation of kainate receptors with different subunit compositions. *J. Neurosci.* 31, 8078–8082. doi: 10.1523/JNEUROSCI.0024-11.2011
- Sun, X., Liu, X., Xia, M., Shao, Y., and Zhang, X. D. (2019). Multicellular gene network analysis identifies a macrophage-related gene signature predictive of therapeutic response and prognosis of gliomas. *J. Transl. Med.* 17:159. doi: 10.1186/s12967-019-1908-1
- Tang, M., Ivakine, E., Mahadevan, V., Salter, M. W., and McInnes, R. R. (2012). *Neto2* interacts with the scaffolding protein GRIP and regulates synaptic abundance of kainate receptors. *PLoS One* 7:e51433. doi: 10.1371/journal.pone.0051433
- Terzibasi, E., Valenzano, D. R., and Cellerino, A. (2007). The short-lived fish *Nothobranchius furzeri* as a new model system for aging studies. *Exp. Gerontol.* 42, 81–89. doi: 10.1016/j.exger.2006.06.039
- Uboveja, A., Satija, Y. K., Siraj, F., Sharma, I., and Saluja, D. (2020). p73 - NAV3 axis plays a critical role in suppression of colon cancer metastasis. *Oncogenesis* 9:12. doi: 10.1038/s41389-020-0193-4
- Villa, E., Critelli, R., Lei, B., Marzocchi, G., Camma, C., Giannelli, G., et al. (2016). Neoangiogenesis-related genes are hallmarks of fast-growing hepatocellular carcinomas and worst survival. Results from a prospective study. *Gut* 65, 861–869. doi: 10.1136/gutjnl-2014-308483
- Wada, M., Goto, Y., Tanaka, T., Okada, R., Moriya, S., Idichi, T., et al. (2020). RNA sequencing-based microRNA expression signature in esophageal squamous cell carcinoma: oncogenic targets by antitumor miR-143-5p and miR-143-3p regulation. *J. Hum. Genet.* 65, 1019–1034. doi: 10.1038/s10038-020-0795-x
- Wyeth, M. S., Pelkey, K. A., Petralia, R. S., Salter, M. W., McInnes, R. R., and McBain, C. J. (2014). *Neto* auxiliary protein interactions regulate kainate and NMDA receptor subunit localization at mossy fiber-CA3 pyramidal cell synapses. *J. Neurosci.* 34, 622–628. doi: 10.1523/JNEUROSCI.3098-13.2014
- Yun, J. W., Lee, S., Ryu, D., Park, S., Park, W. Y., Joung, J. G., et al. (2019). Biomarkers associated with tumor heterogeneity in prostate cancer. *Transl. Oncol.* 12, 43–48. doi: 10.1016/j.tranon.2018.09.003
- Zhan, T., Rindtorff, N., and Boutros, M. (2016). Wnt signaling in cancer. *Oncogene* 36, 1461–1473. doi: 10.1038/ncr.2016.304
- Zhang, C., Luo, Y., Cao, J., Wang, X., Miao, Z., and Shao, G. (2020). Exosomal lncRNA FAM225A accelerates esophageal squamous cell carcinoma progression and angiogenesis via sponging miR-206 to upregulate *NETO2* and *FOXP1* expression. *Cancer Med.* 9, 8600–8611. doi: 10.1002/cam4.3463

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Fedorova, Snezhkina, Lipatova, Pavlov, Kobelyatskaya, Guvatova, Pudova, Savvateeva, Ishina, Demidova, Volchenko, Trofimov, Sukhikh, Krasnov and Kudryavtseva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Characterization of Genome-Wide DNA Methylation and Hydroxymethylation in Mouse Arcuate Nucleus of Hypothalamus During Puberty Process

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Julia Fedotova,
Russian Academy of Sciences, Russia
Jun Wu,
East China Normal University, China
Wen-Lian Chen,
Shanghai University of Traditional
Chinese Medicine, China

*Correspondence:

Jielin Sun
jlsun@sjtu.edu.cn
Hua Li
kaikaixin@sjtu.edu.cn

†These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 06 November 2020

Accepted: 24 November 2020

Published: 14 December 2020

Citation:

Shen Y, Zhou S, Zhao X, Li H and
Sun J (2020) Characterization of
Genome-Wide DNA Methylation and
Hydroxymethylation in Mouse Arcuate
Nucleus of Hypothalamus During
Puberty Process.
Front. Genet. 11:626536.
doi: 10.3389/fgene.2020.626536

Yihang Shen^{1†}, Shasha Zhou^{2†}, Xiaodong Zhao¹, Hua Li^{3*} and Jielin Sun^{1*}

¹ Shanghai Center for Systems Biomedicine, Ministry of Education Key Laboratory for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China, ² Department of Endocrinology, Shanghai Children's Hospital, Shanghai Jiao Tong University, Shanghai, China, ³ Bio-ID Center, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China

Background: Pulsatile pituitary gonadotropin secretion governed by hypothalamic gonadotropin-releasing hormone (GnRH) is essential for the pubertal onset. The epigenetic mechanism underlying the activation of GnRH-dependent regulatory axis in hypothalamus remains elusive. This study aims to explore the potential correlation between the signature of DNA (hydroxyl)methylation and pubertal process.

Methods: Hypothalamic arcuate nucleus (ARC) of mouse at early (4-weeks) and late pubertal (8-weeks) stages underwent RNA-, RRBS-, and RRHP-seq to investigate the genome-wide profiles of transcriptome, differential DNA methylation and hydroxymethylation.

Results: A series of differential expressed genes (DEGs) involved in sexual development could be separated into three subgroups with the significant difference of DNA methylation or hydroxymethylation or both in promoter regions. Compared to DNA methylation, DNA hydroxymethylation partook in more signaling pathways including synapse morphology, channel activity and glial development, which could enhance transsynaptic change and glia-to-neuron communication to facilitate GnRH release. The correlation between transcription and these epigenetic modifications indicated that DNA hydroxymethylation impacted with gene transcription independently of DNA methylation spanning puberty.

Conclusion: Our results characterized the hydroxymethylation pattern and provided an insight into the novel epigenetic regulation on gene expression during pubertal process.

Keywords: puberty onset, DNA methylation, DNA hydroxymethylation, GnRH, ARC

INTRODUCTION

Pubertal development is a multi-factorial process accompanied by maturity of skeletal height, growth spurt, and a myriad of hormonal changes involving genetic, nutritional, socioeconomic, and environmental factors in a systematic manner leading to reproductive maturation. Pubertal development is governed by the hypothalamic–pituitary–gonadal (HPG) axis, and begins with hypothalamic gonadotropin-releasing hormone (GnRH) neurons. Usually, the dormant HPG axis presents the silencing GnRH, extremely low levels of luteinizing hormone (LH) and follicle stimulating hormone (FSH), and estrogen or testosterone until ~8–9 years of age in human beings (Lee and Houk, 2006, 2008). The pulsatile secretion of GnRH from the hypothalamus stimulates on the gonadotroph cells of the pituitary gland to secrete gonadotropins, LH and FSH, and the gonadotropins then stimulate the production of estrogen from the ovaries in females, and testosterone from the testes in males (Chulani and Gordon, 2014). In turn, the secretion of GnRH in hypothalamus is majorly regulated by KiSS-1 metastasis suppressor (Kiss1) and Kiss1 receptor (Kiss1r, also known as GPR54). Kiss1 neurons of the arcuate nucleus (ARC) in the hypothalamus seem to be essential for pulsatile GnRH release in both sexes. Transcriptional activation of these genes was considered as a core mechanism underlying the puberty initiation, which was precipitated by epigenetic cues (Ojeda and Lomniczi, 2014).

Previous studies have indicated that hypothalamic DNA methylation is strongly implicated in the onset of puberty in mammals (Lomniczi et al., 2015; Yuan et al., 2019). Loss of DNA methylation or demethylation has been observed in specific contexts through active or passive mechanisms (Wu and Zhang, 2010). Active DNA demethylation is the enzymatic process that leads to the removal of the methyl group from 5-methylcytosine (5mC) via successive oxidation [5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), 5-carboxylcytosine (5caC)] catalyzed by ten-eleven translocation (TET) family (Kohli and Zhang, 2013). TET2 has been determined to promote transcription and peptide release of GnRH, and consequently maintain reproductive function in *in vitro* and *in vivo* (Kurian et al., 2016). However, the roles of active DNA demethylation in transcriptional regulation in puberty onset is never elucidated.

Female mice have been widely used in multiple studies on pubertal development as they present the similar molecular behaviors in HPG axis and stable cycles of menstrual calendar like human (Pohl et al., 2007). Hypothalamic ARC underwent a huge epigenetic and genetic reprogramming to adapt to the response and feedback on sexual hormones during the stages of early pubertal (2–5-weeks of age) and late puberty (5–8-weeks of age). Here, we harvested 4- and 8-weeks hypothalamic ARC and employed RNA-seq, reduced representation bisulfite sequencing (RRBS) and hydroxymethylation profiling (RRHP) on a genome-wide scale. Given a large number of differential expressed genes (DEGs) and differential 5(h)mC signals across the whole genome, we discovered novel connections between DNA (hydroxyl)methylated modification and gene expression, emphasizing the importance

of epigenetic alterations in regulating transcription during pubertal process.

MATERIALS AND METHODS

Experimental Animals

C57BL/6 female mice purchased from Shanghai SLAC Laboratory Animal Co., Ltd. (Shanghai, China) were housed in clean cages and maintained at $22 \pm 2^\circ\text{C}$ with a constant 12-h light/dark schedule. The animals were allowed free access to food and water. 4- and 8-weeks-old mice ($n = 10$ per group) were used in this study. Initially, preliminary experiment for dye injection was used to target the location of ARC using initial orientation (0.4 mm lateral, 1.60 mm posterior to bregma, 7.40 mm below the surface of the dura) as previously described (Greenwood et al., 2014; Hu et al., 2015). Mice were sacrificed via cervical dislocation, and the whole brains were isolated immediately. The hypothalamic ARC tissues in each group were harvested and gathered for the consequent experiments according to the previous dye staining (**Supplementary Figure 1A**). All the procedures were followed by the Institutional Animal Care and Use Committee of Shanghai Jiao Tong University.

RNA-seq Library Construction and Data Analysis

ARC tissues were stored in 1 ml TRIZOL (Thermo Fisher Scientific, Waltham, MA, USA) and grinded in liquid nitrogen, and were added 100 μl chloroform and fully mixed, then centrifuged with highest speed at 4°C for 10 min. The supernatant was moved into a new tube, and added the isopropanol with same volume, and centrifuged with highest speed at 4°C for 10 min. The precipitate was washed by 75% cold ethanol, and dissolved by appropriate DEPC water. The concentration and quality of RNA was measured by Nanodrop 2000 (Thermo Fisher Scientific) and Agilent bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). 4 μg of RNA in each group were used for library preparation by NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB, Ipswich, MA, USA) following manufacturer's instructions and were sequenced on an Illumina HiSeq platform.

The raw data was trimmed adaptors and filter out low quality reads using Trimmomatic (non-default parameters: SLIDINGWINDOW:4:15 LEADING:10 TRAILING:10 MINLEN:35) (Bolger et al., 2014), and checked the quality of clean reads using Fastqc (Andrews, 2013). Next, clean reads were aligned to the latest mouse genome assembly mm10 using Hisat2 v2.0.5 (non-default parameters: $-\text{rna-strandness RF -dta}$) (Kim et al., 2015). The transcripts were assembled and the expression levels were estimated with FPKM values using the StringTie algorithm (non-default parameters: $-\text{rf}$) (Pertea et al., 2015). Differential mRNA and lncRNA expression among the groups were evaluated using an R package Ballgown (Frazee et al., 2015), and the significance of differences by the Benjamini & Hochberg (BH) p -value adjustment method were computed. Gene annotation was described by Ensembl genome browser database (<http://www.ensembl.org/index.html>). The R package ClusterProfiler was used to annotate the differential genes with

gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Yu et al., 2012).

RRBS and RRHP Library Construction and Data Analysis

Genomic DNA of ARC in two groups were extracted using the QIAquick Gel Extraction Kit (Qiagen, Hilden, Germany). The 200 ng high-quality DNA was then digested by restriction endonucleases MspI (NEB) and subjected to 3'-end blunting and single nucleotide (A) addition and adaptor ligation. For RRHP, 5hmC positions at the adaptor junctions were modified by T4 phage β -glucosyltransferase (NEB), and non-glucosyl-5hmCs were removed by another round of MspI digestion. The 250–500 bp fragments were then selected and treated with bisulfite conversion using Epitect Bisulfite Kit (Qiagen) according to the manufacturer's instructions. Converted DNA were eluted and performed PCR amplification to enrich for fragments with adaptors on both ends. The constructed libraries were quantified using Agilent Bioanalyzer 2100 (Agilent Technologies, Carlsbad, CA, USA) and subjected to high-throughput sequencing using the Illumina HiSeq 2500 platform with paired-end 50 bp sequencing (PE50).

For RRBS, Trim Galore v0.5.0 (non-default parameters: `-max-n 0 -length 35 -rrbs`) were used to filter adapters, short reads (length < 35 bp) and low quality reads. For RRHP, Cutadapt v1.18 (non-default parameters: `-max-n 0 -minimum-length 35`), and Trimmomatic v0.38 (non-default parameters: `SLIDINGWINDOW:4:15 LEADING:10 TRAILING:10 MINLEN:35`) were used to filter adapters, short reads (length < 35 bp) and low quality reads. FastQC (with default parameters) was used to ensure high reads quality. Trimmed reads of RRBS data were aligned to reference genome (assembly GRCm38) using Bismark v0.7.0 (with default parameters) and analyzed DNA methylation profiles using methylKit package (Akalin et al., 2012). DMRs were selected by false discovery rate (FDR) < 0.05 and methylation percentage change between control and test groups are > 10%. For RRHP, clean reads were mapped to the mouse genome (assembly GRCm38) using the Bowtie2 v2.3.4.1 (with default parameters) software. Aligned reads with CCGG tag at 5' end were counted. Differentially hydroxymethylated regions (DHMRs) were determined using the diffReps software. DHMRs were analyzed by log₂ fold change (FC) > 1 or < -1, FDR < 10⁻⁴.

Data Deposits

The raw sequencing data was deposited to ArrayExpress assigned with the accession number E-MTAB-9420 and E-MTAB-9421.

RESULTS

The Differential Expressed Genes During Pubertal Process

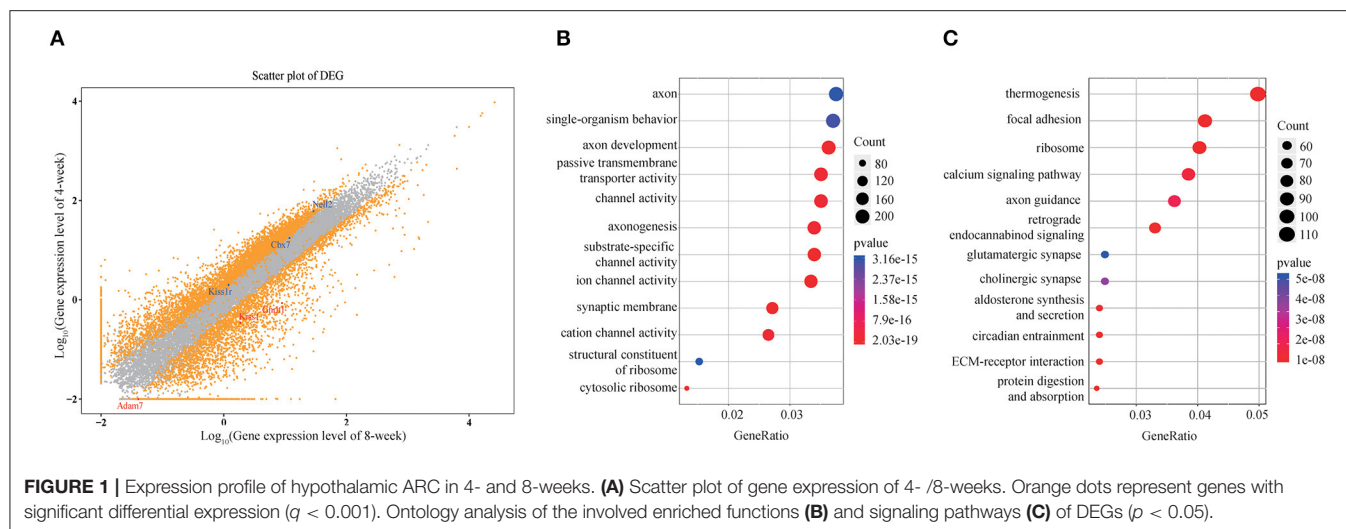
To investigate the changes of epigenome and transcriptome of hypothalamic ARC during puberty progression, we conducted 18 libraries for RNA, RRBS, RRHP-seq derived from hypothalamic ARC of C57BL/6 mice in two sexual developmental stages of early and late puberty. The alignment of data and the

correlations within duplications in each group were summarized in **Supplementary Table 1** and **Supplementary Figure 2**, which indicated a good quality of biological materials in this study. As previously described (Li et al., 2017), developmental stages of C57BL/6 mouse were roughly divided into prepuberty and early puberty periods (2–5-weeks of age), late puberty (5–8-weeks of age), and young adulthood (8–12-weeks of age). Here, we detected the vulva morphology and the changes of LH and FSH in sera of 4- and 8-weeks mice, and validated that 8-weeks mice displayed a phenotype of higher hormone levels and gonadal activation compared with 4-weeks mice (**Supplementary Figures 1B,C**).

A total of 5,778 DEGs were obtained among which 1,787 protein coding genes were up-regulated while 3,991 were down-regulated in 8-weeks group compared with 4-weeks one (log₂FC > 1 or < -1, FDR < 10⁻³). Take an example of the well-acknowledged puberty associated genes, in contrast to the high expression of Kiss1, GnRH, and Adam7 in early pubertal stage of ARC, the presence of substantially decreased Cbx7, Kiss1r, and Nell2 was observed during pubertal process (**Figure 1A**). Moreover, the functional and signaling pathway enrichment analysis showed that DEGs majorly involved in neurodevelopment, synaptic behavior and transmembrane and extracellular signal transduction (**Figures 1B,C**), indicating that the specific functional neurons in ARC underwent a complicated process of signals communication and stimulation for maturation. We also observed a high correlation of glutamatergic synapse with puberty ($p = 5.24 \times 10^{-8}$) and the high expression of glutamate metabolism associated genes such as Grik, Grin and Adcy families in 4-weeks compared with 8-weeks ARC. Previous study indicates that the coordinated activity of glutamatergic neurons and GnRH neurons facilitates the sensitivity of GnRH secretion (Parent et al., 2005). Besides that, choline, aldosterone and endocannabinoid did synthesize and secrete, which likely had profound effects on puberty initiation (Wenger et al., 2002; Biasi, 2010; Genovesi et al., 2018). Taken together, these observations validated the known genetic signatures and indicated multiple activated signaling pathways during pubertal process.

Genome-Wide Landscape of DNA Methylation and Hydroxymethylation in ARC

To further investigate the regulatory machinery underlying transcription, RRBS and RRHP were employed to detect the altered genome-wide distribution of 5mC and 5hmC in 4- and 8-weeks old hypothalamic ARC. Our data showed that DMRs and DHMRs (RRBS: FC > or < 10%, FDR < 0.05; RRHP: log₂FC > 1 or < -1, FDR < 10⁻⁴) majorly occurred at the promoter, intron and intergenic regions (**Figures 2A,B**). Here, we focused on the CpG loci located at promoter region of DEGs and observed that these genes whose promoter showed the significantly differential DNA methylation levels between 4- and 8-weeks groups were functionally enriched in signaling pathways closely connected with sexual development and hormone secretion (**Figures 2C,D**). Likewise, DEGs whose CpG loci at promoter had the remarkable



increasing or declining hydroxymethylation levels were majorly associated with the function of synapse morphology and channel activity including glutamatergic synapse (**Figures 2E,F**). In addition, we found that most of functions and pathways (86.8%) enriched by DHMRs overlapped with those of DEGs compared with DNA methylation (**Figures 2G,H**). Previous studies have determined that transsynaptic change and glial-neuronal connection attribute to GnRH neuron activation (Ojeda et al., 2010). Current data showed that DHMRs involved in more genes enriching the functions of glutamate and gamma-aminobutyric acid (GABA) neuron as well as glia development rather than DNA methylation. Our results indicated that DNA methylation might impact with the more intuitive phenotype of hormone secretion while the functions related DNA hydroxymethylation were associated with various auxiliary neurons which were not the most obvious characteristics of sexual development, but were used to trigger and maintain the pubertal initiation. Taken together, the given 5(h)mC patterns in ARC suggested that DNA (hydroxyl)methylation was closely connected with gene expression in puberty onset.

Independent Roles of DNA Hydroxymethylation and Methylation in Regulating Gene Expression

As well-acknowledgment that DNA methylation in promoters usually negatively correlated with transcription, while DNA hydroxymethylation in promoters displayed a positive correlation with gene expression, we observed the consistent epigenetic regulation of DNA hydroxyl(methylation) on DEGs in our system. Given the overlap of enriched functions between DNA (hydroxy)methylation and DEGs, we grouped three clusters of gene which are negatively correlated with DNA methylation, positively correlated with DNA hydroxymethylation as well as both (**Figure 3A**). The presence of nine genes including *Bhlha15*, *Ins15*, *Msmg*, *Plcb2*, *Slc17a8*, *Sox3*, *Tnfrsf2*, *Uck2*, and *Ypel2* in first cluster, 3,277 genes including

Epb41, *Pebp1* in second cluster, as well as 38 genes including *Areg* and *Nr5a1* in third cluster was observed (**Figure 3B**). During pubertal development, the connection between DNA methylation and hydroxymethylation was seemingly not closed, maybe only in 38 genes of the third cluster.

Moreover, we obtained the FCs of transcription, 5mC and 5hmC between 4- and 8-weeks, and compared the FC of gene expression normalized by the FC of 5(h)mC, indicating that DNA hydroxymethylation could impact gene expression more powerful than DNA methylation (**Figure 3C**). To further investigate the relationship between differential 5mC and 5hmC from the overlapped D(H)MRs, we calculated correlation using Pearson's chi-squared test. Unexpectedly, we failed to observe any significant relativity between each other ($r = 0.0543$, $p > 0.05$) (**Figure 3D**). Although in consideration of 5hmC as an intermediate of demethylation, however, the results above suggested that DNA hydroxymethylation played a regulatory role in transcription independent of DNA methylation although they had the closely chemical connection.

DISCUSSION

Although mutations in multiple genes such as kisspeptin system, *MKRN3*, *DLK1* have been identified in sporadic and familial cases of central precocious puberty (CPP), many factors involved in pubertal initiation and transition remain poorly understood (Aguirre and Eugster, 2018). Our data shows that the DEGs of hypothalamic ARC between 4- and 8-weeks include a number of well-acknowledged pubertal associated genes. However, we have meanwhile detected ARC of adult rat, and fail to observe any significant change of *Kiss1* and *GnRH* compared with 4- or 8-weeks (data not shown), which suggests that the expressions of *GnRH* and other pubertal genes are fluctuant due to the pulsatile release of *GnRH* and periodic estrus cycle in adult individuals, but their expression changing law during the stages of puberty onset seems more stable than adult stage.

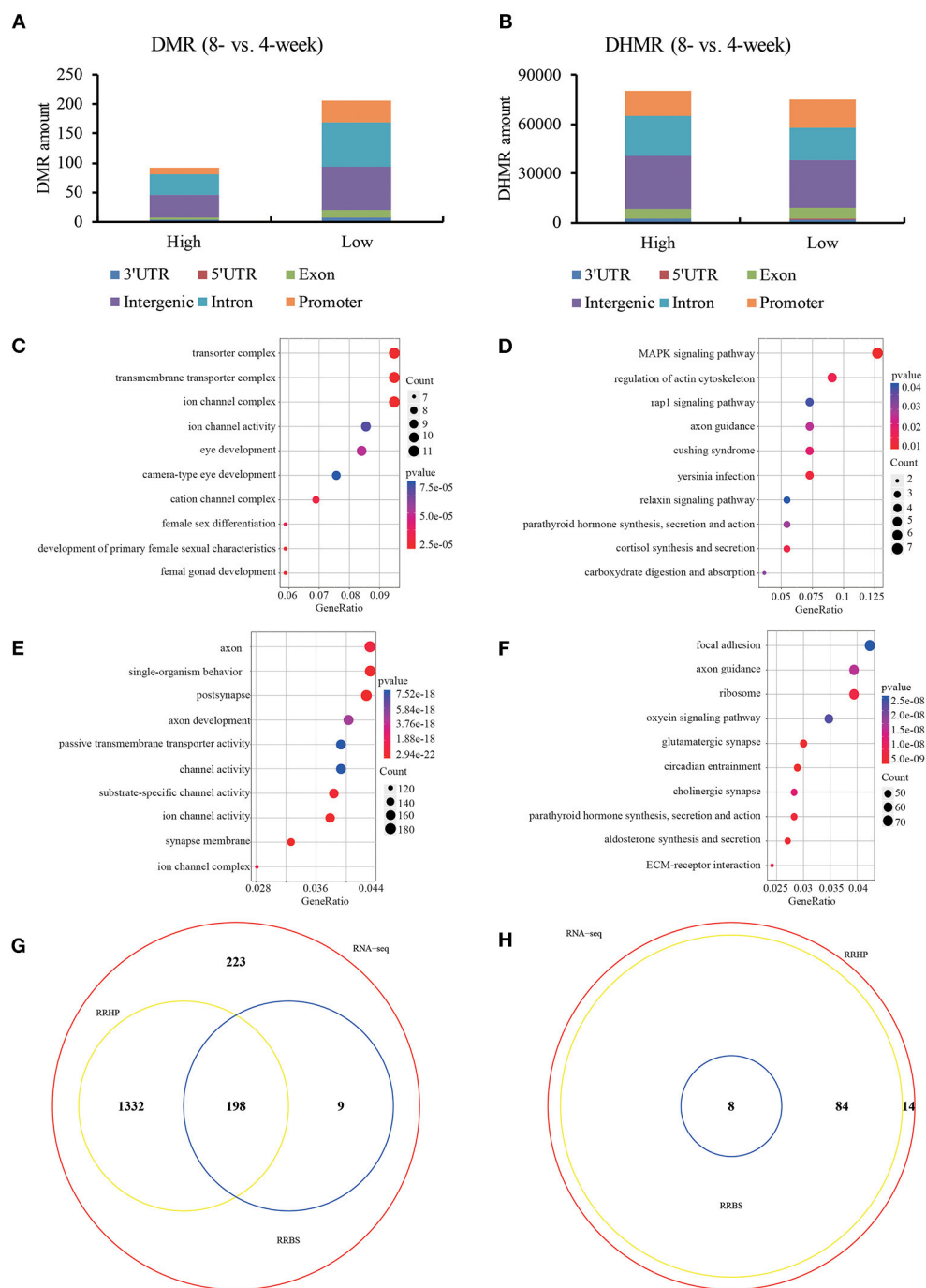
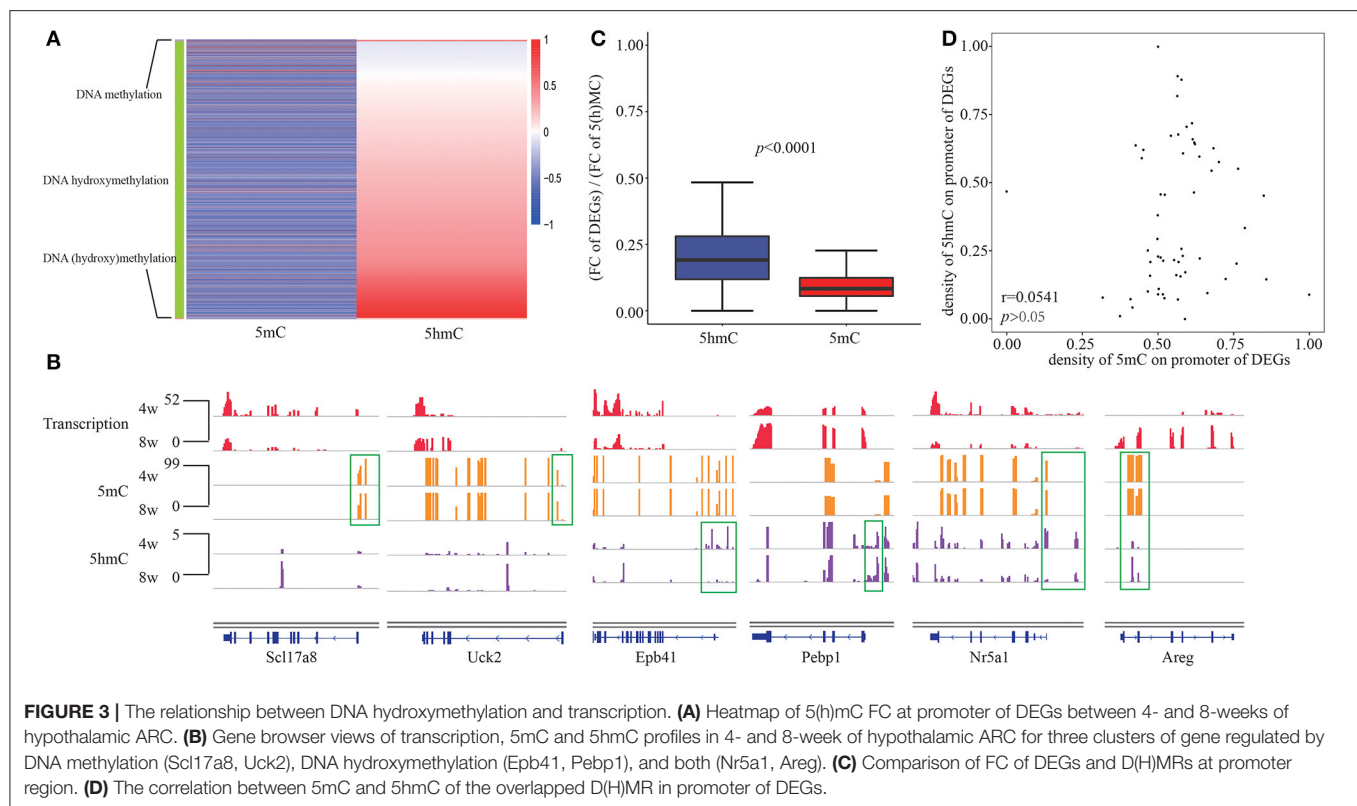


FIGURE 2 | DNA (hydroxy)methylation patterns of hypothalamic ARC in 4- and 8-weeks. The distribution of DMR (A) and DHMR (B) regions in genomic contexts of DEGs. DMR: differential methylated regions, FC > or < 20%, $q < 0.05$; DHMR: differential hydroxymethylated regions, $\log_2FC > 1$ or < -1 , $q < 1e-4$. Ontology analysis of the involved enriched functions (C) and signaling pathways (D) of differential methylated DEGs ($p < 0.05$). Ontology analysis of the involved enriched functions (E) and signaling pathways (F) of differential hydroxymethylated DEGs ($p < 0.05$). The Venn diagram view of enriched functions (G) and signaling pathways (H) among DEGs with differential (hydroxy)methylation.

A large number of recent studies have suggested that CpG methylation changes are likely to show a crucial regulatory in controlling the transcription of the well-acknowledged pubertal

genes related to GnRH and estrogen signaling pathways in mammalian hypothalamus (Mellen et al., 2012; Alves et al., 2017; Thompson et al., 2018). Although our omics data indicates that



DNA methylation change is associated with the expression of a small proportion of genes on gonad development from early to later stages of puberty onset, however, the expressions of more genes are actually impacted by DNA hydroxymethylation at promoter including Kiss1 and Kiss1r. The secretion of GnRH in hypothalamus is majorly regulated by KiSS-1 metastasis suppressor (Kiss1) and Kiss1 receptor (Kiss1r, also known as GPR54). Kiss1 neurons of the arcuate nucleus (ARC) in the hypothalamus seem to be essential for pulsatile GnRH release in both sexes. We speculate that an accompanied genome-wide demethylation processes to facilitate synapse organization to accommodate to regulating the high-level hormone secretion in hypothalamus. 5hmC as a hallmark and an intermediate of demethylation process is determined to play an essential role in normal sexual development in central nervous system in this study. Furthermore, the differential DNA hydroxymethylated genes and their involved functions and pathways have no more than 15% overlap with the ones of DNA methylation, and differences of 5hmC at the promoter of DEGs affect transcription more robustly than 5mC even at the same “CG” loci, which implies that DNA hydroxymethylation exerts an epigenetic regulation independent of DNA methylation although they tightly connect with each other from the chemical basis. 5hmC, which is more than the intermediate of demethylation process *per se*, is likely to reverse the traditional repressive functions of MeCP2 (Mellen et al., 2012), and recruit multiple transcription factors to create an environment to facilitate gene transcription (Ichiyama et al., 2015).

Additionally, we focus on the associated regulatory networks of Kiss1 and GnRH, which seems to be the most important gene for puberty onset. We find that the enriched functions involving Kiss1 but not GnRH from DHMR associated genes and DEGs completely overlapped. We speculated that DNA hydroxymethylation is likely to govern the upstream GnRH regulatory axis in the early stage of puberty initiation. Our data reveals the dynamic DNA (hydroxy)methylation changes of genome of ARC during puberty process, not only indicates the gene expression regulation, but also provides the potential therapeutic targets by epigenetic drugs for puberty associated diseases treatment.

CONCLUSION

Overall, our data shows the dynamic change of genome-wide methylation and hydroxymethylation in hypothalamic ARC, and uncovers a novel characterization of DNA hydroxymethylation for regulating transcription during pubertal process. The outcomes advance the understanding on a novel mechanism of epigenetic regulation on gene, and contributes to improving therapeutic strategy for disorders of sex development.

DATA AVAILABILITY STATEMENT

The raw sequencing data was deposited to ArrayExpress assigned with the accession numbers E-MTAB-9420 and E-MTAB-9421.

ETHICS STATEMENT

The animal study was reviewed and approved by Shanghai Children's Hospital, Shanghai Jiao Tong University.

AUTHOR CONTRIBUTIONS

YS analyzed the data, prepared for the figures, and drafted the manuscript. SZ performed the animal experiments and harvested the samples. JS and HL designed the overall project, instructed experiments, and revised the manuscript. XZ helped design the overall project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Major projects of Science and Technology Commission of Shanghai (No. 17JC1400800), Interdisciplinary Program of Medical Engineering Cross Fund

REFERENCES

- Aguirre, R. S., and Eugster, E. A. (2018). Central precocious puberty: from genetics to treatment. *Best Pract. Res. Clin. Endocrinol. Metab.* 32, 343–354. doi: 10.1016/j.beem.2018.05.008
- Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F. E., Figueroa, M. E., Melnick, A., et al. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* 13:R87. doi: 10.1186/gb-2012-13-10-r87
- Alves, B. R., Cardoso, R. C., Doan, R., Zhang, Y., Dindot, S. V., Williams, G. L., et al. (2017). Nutritional programming of accelerated puberty in heifers: alterations in DNA methylation in the arcuate nucleus. *Biol. Reprod.* 96, 174–184. doi: 10.1095/biolreprod.116.144741
- Andrews, S. (2013). *FastQC A Quality Control Tool for High Throughput Sequence Data*. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Biasi, E. (2010). Effects of postnatal dietary choline manipulation against MK-801 neurotoxicity in pre- and postadolescent rats. *Brain Res.* 1362, 117–132. doi: 10.1016/j.brainres.2010.09.012
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Chulani, V. L., and Gordon, L. P. (2014). Adolescent growth and development. *Prim. Care* 41, 465–487. doi: 10.1016/j.pop.2014.05.002
- Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., and Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* 33, 243–246. doi: 10.1038/nbt.3172
- Genovesi, S., Antolini, L., Orlando, A., Tassistro, E., Giussani, M., Nava, E., et al. (2018). Aldosterone-to-renin ratio depends on age and sex in children attending a clinic for cardiovascular risk assessment. *J. Hypertens.* 36, 344–352. doi: 10.1097/HJH.0000000000001571
- Greenwood, M., Bordieri, L., Greenwood, M. P., Rosso Melo, M., Colombari, D. S., Colombari, E., et al. (2014). Transcription factor CREB3L1 regulates vasopressin gene expression in the rat hypothalamus. *J. Neurosci.* 34, 3810–3820. doi: 10.1523/JNEUROSCI.4343-13.2014
- Hu, M. H., Li, X. F., McCausland, B., Li, S. Y., Gresham, R., Kinsey-Jones, J. S., et al. (2015). Relative importance of the arcuate and anteroventral periventricular kisspeptin neurons in control of puberty and reproductive function in female rats. *Endocrinology* 156, 2619–2631. doi: 10.1210/en.2014-1655
- Ichiyama, K., Chen, T., Wang, X., Yan, X., Kim, B. S., Tanaka, S., et al. (2015). The methylcytosine dioxygenase Tet2 promotes DNA demethylation and activation of cytokine gene expression in T cells. *Immunity* 42, 613–626. doi: 10.1016/j.immuni.2015.03.005
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Kohli, R. M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* 502, 472–479. doi: 10.1038/nature12750
- Kurian, J. R., Louis, S., Keen, K. L., Wolfe, A., Terasawa, E., and Levine, J. E. (2016). The Methylcytosine dioxygenase ten-eleven translocase-2 (tet2) enables elevated GnRH gene expression and maintenance of male reproductive function. *Endocrinology* 157, 3588–3603. doi: 10.1210/en.2016-1087
- Lee, P. A., and Houk, C. P. (2006). Gonadotropin-releasing hormone analog therapy for central precocious puberty and other childhood disorders affecting growth and puberty. *Treat. Endocrinol.* 5, 287–296. doi: 10.2165/00024677-200605050-00003
- Lee, P. A., and Houk, C. P. (2008). Physical and psychosexual pubertal development: physiology and pathophysiology of puberty with lessons learned from those with disorders of sex development. *Pediatr. Endocrinol. Rev.* 6, 248–256.
- Li, C., Chai, Y., Wang, L., Gao, B., Chen, H., Gao, P., et al. (2017). Programmed cell senescence in skeleton during late puberty. *Nat. Commun.* 8:1312. doi: 10.1038/s41467-017-01509-0
- Lomniczi, A., Wright, H., and Ojeda, S. R. (2015). Epigenetic regulation of female puberty. *Front. Neuroendocrinol.* 36:3. doi: 10.1016/j.yfrne.2014.08.003
- Mellen, M., Ayata, P., Dewell, S., Kriaucionis, S., and Heintz, N. (2012). MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. *Cell* 151, 1417–1430. doi: 10.1016/j.cell.2012.11.022
- Ojeda, S. R., and Lomniczi, A. (2014). Puberty in 2013, Unravelling the mystery of puberty. *Nat. Rev. Endocrinol.* 10, 67–69. doi: 10.1038/nrendo.2013.233
- Ojeda, S. R., Lomniczi, A., Loche, A., Matagne, V., Kaidar, G., Sandau, U. S., et al. (2010). The transcriptional control of female puberty. *Brain Res.* 1364, 164–174. doi: 10.1016/j.brainres.2010.09.039
- Parent, A. S., Matagne, V., and Bourguignon, J. P. (2005). Control of puberty by excitatory amino acid neurotransmitters and its clinical implications. *Endocrine* 28, 281–286. doi: 10.1385/ENDO:28:3:281
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122
- Pohl, J., Olmstead, M. C., Wynne-Edwards, K. E., Harkness, K., and Menard, J. L. (2007). Repeated exposure to stress across the childhood-adolescent period alters rats' anxiety- and depression-like behaviors in adulthood: the

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.626536/full#supplementary-material>

Supplementary Figure 1 | The phenotype and morphology of pubertal female mice. **(A)** Dye injection for ARC location. Left: the ventral view of hypothalamus; right: sectional view of coronal suture of hypothalamus. Red pentagon indicates the location of ARC. **(B)** Comparison of vulva morphology between 4- and 8-weeks mice highlighted by the red arrows. The serum levels of LH **(C)** and FSH **(D)** between 4- and 8-weeks mice. “***” represent $p < 0.01$.

Supplementary Figure 2 | The repeatability and reliability of omics data. Correlation analysis of replicate data, including RNA-seq, RRBS-seq, and RRHP-seq for 4- and 8-weeks ARC samples. Correlations were calculated using whole genome data.

- importance of stressor type and gender. *Behav Neurosci.* 121, 462–474. doi: 10.1037/0735-7044.121.3.462
- Thompson, E. E., Nicodemus-Johnson, J., Kim, K. W., Gern, J. E., Jackson, D. J., Lemanske, R. F., et al. (2018). Global DNA methylation changes spanning puberty are near predicted estrogen-responsive genes and enriched for genes involved in endocrine and immune processes. *Clin Epigenetics* 10:62. doi: 10.1186/s13148-018-0491-2
- Wenger, T., Gerendai, I., Fezza, F., Gonzalez, S., Bisogno, T., Fernandez-Ruiz J, et al. (2002). The hypothalamic levels of the endocannabinoid, anandamide, peak immediately before the onset of puberty in female rats. *Life Sci.* 70, 1407–1414. doi: 10.1016/S0024-3205(01)01516-8
- Wu, S. C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. *Nat. Rev. Mol. Cell Biol.* 11, 607–620. doi: 10.1038/nrm2950
- Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16, 284–287. doi: 10.1089/omi.2011.0118
- Yuan, X., Zhou, X., Chen, Z., He, Y., Kong, Y., Ye, S., et al. (2019). Genome-Wide DNA methylation analysis of hypothalamus during the onset of puberty in gilts. *Front Genet.* 10:228. doi: 10.3389/fgene.2019.00228

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Shen, Zhou, Zhao, Li and Sun. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



ChIAMM: A Mixture Model for Statistical Analysis of Long-Range Chromatin Interactions From ChIA-PET Experiments

Yibeltal Arega¹, Hao Jiang¹, Shuangqi Wang², Jingwen Zhang², Xiaohui Niu¹ and Guoliang Li^{1,2*}

¹ Agricultural Bioinformatics Key Laboratory of Hubei Province, Hubei Engineering Technology Research Center of Agricultural Big Data, College of Informatics, Huazhong Agricultural University, Wuhan, China, ² National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Veniamin Fishman,
Institute of Cytology and Genetics,
Russian Academy of Sciences, Russia
Zhihua Zhang,
Beijing Institute of Genomics, Chinese
Academy of Sciences, China

*Correspondence:

Guoliang Li
guoliang.li@mail.hzau.edu.cn

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 11 October 2020

Accepted: 11 November 2020

Published: 14 December 2020

Citation:

Arega Y, Jiang H, Wang S,
Zhang J, Niu X and Li G (2020)
ChIAMM: A Mixture Model
for Statistical Analysis of Long-Range
Chromatin Interactions From
ChIA-PET Experiments.
Front. Genet. 11:616160.
doi: 10.3389/fgene.2020.616160

Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) is an important experimental method for detecting specific protein-mediated chromatin loops genome-wide at high resolution. Here, we proposed a new statistical approach with a mixture model, chromatin interaction analysis using mixture model (ChIAMM), to detect significant chromatin interactions from ChIA-PET data. The statistical model is cast into a Bayesian framework to consider more systematic biases: the genomic distance, local enrichment, mappability, and GC content. Using different ChIA-PET datasets, we evaluated the performance of ChIAMM and compared it with the existing methods, including ChIA-PET Tool, ChiaSig, Mango, ChIA-PET2, and ChIA-PoP. The result showed that the new approach performed better than most top existing methods in detecting significant chromatin interactions in ChIA-PET experiments.

Keywords: ChIA-PET, chromatin interactions, genome-wide, mixture model, bayesian framework

INTRODUCTION

Diverse high-throughput methods have been developed to detect genome-wide chromatin interactions, including chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and high-throughput chromosome conformation capture (Hi-C) (Fullwood et al., 2009; Lieberman-Aiden et al., 2009). ChIA-PET was first introduced in 2009 as an essential experimental method for studying genome-wide chromatin interactions mediated by a specific protein of interest. It can discover many chromatin interactions at a higher resolution that are needed for studying gene transcription regulation. It has been widely used to study various proteins such as estrogen receptor alpha, RNA polymerase II (RNAPII), CCCTC binding factor (CTCF) in human and mouse genome (Fullwood et al., 2009; Handoko et al., 2011; Li et al., 2012; Tang et al., 2015), and H3K4me3, H3K9me2, and RNAPII in rice and maize (Peng et al., 2019; Zhao et al., 2019).

The processing of raw ChIA-PET data is not easy. ChIA-PET experiment will generate tens of millions of paired reads containing a tag and linker sequence (barcode). The tag can be short (generated by the original protocol, and it is about 20 base pairs) or long (generated by the improved protocol, and it is about 150–250 base pairs) (Li et al., 2017). The steps to process raw ChIA-PET data include linker trimming, read alignment, paired-end tag (PET) filtering, PCR duplicate

removal, peak calling, and chromatin interaction calling. In ChIA-PET data, similar to other high-throughput sequencing data, there is a mixture of signals (fragment pairs from real chromatin interactions, termed as true pairs) and noise (fragment pairs from random ligation, termed as false pairs). Distinguishing the true interaction pairs from the random noise is not a simple task, and complicated computational tools are needed (He et al., 2016). Up to now, there are several published tools, and ChIA-PET Tool (Li et al., 2010), ChiaSig (Paulsen et al., 2014), Mango (Phanstiel et al., 2015), ChIA-PET2 (Li et al., 2016), and ChIA-PoP (Huang et al., 2019) are the representative ones.

The ChIA-PET Tool is the first software package for the automatic processing of ChIA-PET sequence data, which uses hypergeometric distribution (HG) as the statistical method and accounts for the sequencing depth bias. It fails to correct the major source of bias (He et al., 2015; Phanstiel et al., 2015), such as the genomic distance between the interacting regions. ChiaSig (Paulsen et al., 2014) advanced the ChIA-PET Tool by incorporating genomic distance between interacting anchors. It uses non-central HG distribution for modeling the frequency of chromatin interactions, and the model considers the non-specific ligations that exist because of genomic distance proximity. As a limitation, ChiaSig has a high false-negative rate (He et al., 2015), it executes the final step in ChIA-PET Tool data analysis, and users are expected to write their programs (Phanstiel et al., 2015). Similar to ChiaSig, Mango (Phanstiel et al., 2015) is designed for correcting the primary source of biases from genomic proximity using the binomial model. As a limitation, Mango does not model the interactions between different chromosomes. Besides, it is too conservative at the significant loop calling step, just reporting a small number of interactions, which led to a high false-negative rate (Li et al., 2016). ChIA-PET2 (Li et al., 2016) is a complete analysis pipeline that uses a Bayesian mixture model to process both bridge and half-linker ChIA-PET data from raw sequencing reads to significant chromatin loop calls. As a limitation, it gives slightly different results for the same input (Huang et al., 2019). ChIA-PoP (Huang et al., 2019) was proposed using zero truncated Poisson distribution for accounting for the genomic distance and sequence biases. It is designed for short-read ChIA-PET datasets only. ChIA-PoP considers intra- and interchromosomal interaction as a separate model. Recently, ChIA-PIPE (Lee et al., 2020) was proposed by integrating the special functions related to the experiment types, data processing, and structural interpretation. ChIA-PIPE used ChiaSig (Paulsen et al., 2014) to calculate the statistical significance of interactions.

All the above existing tools considered only the genomic distance or anchor depth as biases. But in different studies, the GC content and mappability score are listed as systematic sources of biases (Yaffe and Tanay, 2011; Hu et al., 2012; Imakaev et al., 2012). Hence, the existing tools failed to address it. Besides, from the existing tools, except for ChIA-PET Tool V3 (Li et al., 2019), ChIA-PET2 (Li et al., 2016), and ChIA-PIPE (Lee et al., 2020), others are designed exclusively for short-read ChIA-PET data analysis.

Here, we present a new statistical method called chromatin interaction analysis using mixture model (ChIAMM) to

distinguish signals from noise in ChIA-PET data. It considers the genomic distance between anchors, sequence depth, GC content, and mappability as systematic sources of bias. The model was tested on both RNAPII and CTCF ChIA-PET data from human K562 and MCF7 and RNAPII and H3K9me2 ChIA-PET data from rice MH63. The performance of the proposed method was evaluated with the aggregate peak analysis (APA) plot, CTCF coverage of anchors, and CTCF motif orientation analysis. The results showed that the new method performed better with the most top existing tools.

MATERIALS AND METHODS

Public Datasets Used

In this study, MCF7 and K562 RNAPII data in Li et al. (2012), MCF7 and K562 CTCF data in GEO with accession numbers GSM970215 and GSM970216, respectively, and MH63 RNAPII and H3K9me2 data in Zhao et al. (2019) were processed. For the CTCF enrichment and motif orientation analyses, the CTCF peak regions from ENCODE ChIP-Seq datasets ENCFF990LUT and ENCFF720OXG for MCF7, and ENCFF559HEE and ENCFF681OMH for K562 datasets were used.

Systematic Biases Considered in the Study

In this study, we used genomic distance, GC content, mappability, and enrichment as systematic biases of the ChIA-PET experiment. We used ChIA-PET Tool version 3 (V3) (Li et al., 2019) as the primary processing pipeline to find the anchor sites, genomic distance, interaction frequency, type of interaction, marginal count, and self-ligation PETs. It is known that regions close together along the genomic sequence will have a higher chance of forming random contacts. Thus, it is essential to integrate the genomic distance into the model (Paulsen et al., 2014), and we primarily considered the genomic distance as a bias. The second bias is the GC content, defined as the percentage of cytosine (C) and guanine (G) bases in a given region. In different studies, GC content has been reported as a systematic bias in next-generation sequencing (NGS) applications (Yaffe and Tanay, 2011; Hu et al., 2012), and the GC content of each anchor is calculated using bedtools nuc (Quinlan and Hall, 2010) function. The third bias is the mappability score, which is defined as the mappability of all possible k-mers in a given anchor site. The mappability track is downloaded from the UCSC Genome Browser website (Derrien et al., 2012), and the overlap of the mappability track with anchors was performed using bedtools. The last systematic bias is the local enrichment in a given region. It is well known that the anchors with more enrichment have a higher probability of forming interligation PETs by random chance. Different studies have considered enrichment as systematic bias in their analysis (Li et al., 2010, 2016; Paulsen et al., 2014; He et al., 2015; Niu and Lin, 2015; Phanstiel et al., 2015; Huang et al., 2019). In this study, we measured the anchor enrichment using the number of self-ligation PETs found by ChIA-PET Tool (V3).

Statistical Mixture Model

In many situations, like the ChIA-PET experiment, due to the complex nature of the observed data, using single parametric distribution is insufficient for inference. Here, we used a mixture model. It offers a solution to this problem by assuming that the frequency of chromatin interactions can be represented by a weighted sum of distributions, with each distribution representing a proportion contribution to the data.

We used a mixture model for modeling the interaction frequency of the ChIA-PET experiment. Let $Y = \{y_i, i = 1, 2, \dots, n\}$ represent the interaction frequencies for each observed anchor pair i from n unique anchor pairs (say, anchor A_i and B_i). The interaction frequency, y_i , has a two-component mixture distribution, i.e., signal and noise. The mixture model integrates signal and noise interaction frequency as follows:

$$y_i \sim \sum_{j=0}^1 W_{ji} p(\cdot | \lambda_{ji}) \quad i = 1, 2, \dots, n$$

where W_{ji} is the mixing probability (i.e., W_{0i} and W_{1i} represent the probability of pair i being a false pair and true pair, respectively), and $W_{0i} + W_{1i} = 1$.

It is well known that Poisson distribution is the most popular distribution for modeling NGS count data, and in the above model, $p(\cdot | \lambda)$ is the $(k-1)$ truncated Poisson distribution. The model considers the interaction frequency, $y_i \geq k$ (where k is a cut-off point). The cut-off point is used to decide a pair that is kept in the analysis. Most of the time, it is determined by the researcher. In this study, the cut-off value is ≥ 2 , the same as in (Fullwood et al., 2009).

The probability mass function for Poisson distribution is written as:

$$p(Y = y | \lambda) = \frac{e^{-\lambda} \lambda^y}{y!}$$

and the probability mass function for $k-1$ truncated Poisson distribution is written as follow:

$$p(Y = y | y \geq k, \lambda) = \frac{\lambda^y}{y! \left\{ e^{-\lambda} - \left[1 + \lambda + \frac{\lambda^2}{2!} + \dots + \frac{\lambda^{k-1}}{(k-1)!} \right] \right\}}$$

$$\text{for } y = k, k+1, \dots$$

Therefore, for $k \geq 2$, $p(Y = y | \lambda)$ is written as:

$$p(Y = y | y > 1, \lambda) = \frac{\lambda^y}{y! [e^{-\lambda} - (1 + \lambda)]} \quad \text{for } y = 2, 3, \dots$$

In simplified form, we can express it using the cumulative distribution function (CDF) as follows:

$$p(Y = y | y > 1, \lambda) = \frac{p(y | \lambda)}{1 - F(1)}$$

For pair of i , $1 \leq i \leq n$, $p(Y = y | \lambda)$ will be $p(\cdot | \lambda_{0i})$ and $p(\cdot | \lambda_{1i})$, which model the interaction frequency conditional on it being noise and signal, respectively, and $F(1) = F(y \leq 1)$

represents the probability that the random variable takes a value ≤ 1 . Besides, from the biological perspectives, the signals have more intensity than the noises (Rousseau et al., 2011), and thus, we put the requirements $\lambda_{0i} < \lambda_{1i}$.

From the listed biases, genomic distance has no explicit rule to measure in interchromosomal interaction data. Hence, we model the intra- and interchromosomal interaction data separately and have different rate parameters (λ) and biases (x_i) as well. The rate parameters of intra- (λ) and inter chromosomal (λ') interactions are connected with the biases using the link function. The listed biases in this study are GC percentage (x_i^{gc} and $x_i'^{gc}$), mappability (x_i^{map} and $x_i'^{map}$) and enrichment (x_i^{enr} and $x_i'^{enr}$) for intra- and interchromosomal interaction, respectively. We considered the genomic distance only for intrachromosomal interactions. In the intrachromosomal analysis, we considered all the biases, but in the interchromosomal interaction analysis, we will remove out the distance from the statistical model.

The link functions of intra- and interchromosomal interaction are written as follows, respectively:

$$\log(\lambda_{0i}) = \beta_0 + \beta_1 \log(x_i^{enr}) + \beta_2 \log(x_i^{gc}) + \beta_3 \log(x_i^{map}) + \beta_4 \log(x_i^{dis})$$

$$\log(\lambda'_{0i}) = \beta'_0 + \beta'_1 \log(x_i'^{enr}) + \beta'_2 \log(x_i'^{gc}) + \beta'_3 \log(x_i'^{map})$$

In Bayesian inference, the prior distribution is a crucial part, representing the information about an uncertain parameter. The priors and model description of inter- and intrachromosomal interactions are similar. We used the prime symbol ($'$) for parameters in the interchromosomal interaction model. To simplify the next discussion, we will use the intrachromosomal interaction model parameters as an example.

A normal distribution is a natural prior choice for β_j . Therefore, the coefficients of the Poisson regression model, β_j , $j = 1, 2, 3, 4$ have normal prior with mean zero and reasonable variance to enable large enough deviations, $\beta_j \sim N(0, 3^2)$ (Carlin and Louis, 2008; Gelman et al., 2013; Halla-aho, 2015), and we declared $\lambda_{1i} = C + \lambda_{0i}$ to show that the frequency of signal is greater than the noise, where C is a positive number that follows zero truncated normal distribution with reasonable variance, $C \sim N(0, 3^2)$. In (Halla-aho, 2015), different C_i were considered, but the estimated C_i has very small variance. Therefore, the researcher recommended others to use the same C for next work. This help us in the side of reducing computational time. The statistical approach considers the correlation between common anchor pairs (Niu and Lin, 2015). The dependency incorporated in the weights of the mixture model, i.e., the weight changes from common to pair-specific values, $W_{1i} \sim \text{Beta}(mc_i, mc)$, where mc_i and mc is the marginal count of the i -th paired anchors and the mean of marginal count, respectively.

When we compute the marginal count, we considered the interaction frequency y_i two times; hence, we subtracted one y_i , i.e.,

$$mc_i = mc_{A_i} + mc_{B_i} - y_i$$

where mc_{A_i} and mc_{B_i} are the marginal count of anchor A_i and B_i , respectively, and y_i is the interaction count between anchors (Figure 1), and mc is the average of marginal counts and calculate as $mc = \frac{1}{n} \sum_{i=1}^n mc_i$.

Finally, we define the new latent variable Z_i , $i = 1, \dots, n$ that indicates the category of interaction groups, i.e., whether the interaction frequency is in the signal or noise group:

$$Z_i = \begin{cases} 1, & \text{the pair } i \text{ is a signal} \\ 0, & \text{the pair } i \text{ is a noise} \end{cases}$$

The indicator variable has two outcomes (0 and 1), and it follows the Bernoulli distribution, $Z_i \sim \text{Bernoulli}(W_{1i})$, for $i = 1, 2, \dots, n$, and it is concluded that pair i is signal pair whenever $P(Z_i = 1|Y)$ is bigger than a cut-off value, 0.5 (Niu and Lin, 2015).

Aggregate Peak Analysis

Aggregate peak analysis is the standard and recommended plot that measures the aggregate enrichment of putative peaks in a contact matrix. It plots the sum of a series of submatrices around the interaction anchors derived from the contact matrix. The matrix is created by summing together all submatrices around each putative individual peak. The resulting APA plot displays the total number of contacts that lie within the entire putative peak set at the center of the matrix. It is recommended to use peak to lower left (P2LL) value to compare the interactions from different methods. We generate an APA plot with 5-kb resolution contact matrices for significant chromatin interactions. The BEDPE files from the ChIA-PET data were used to build interaction matrices.

RESULTS

Chromatin interaction analysis using mixture model used a mixture model to distinguish signals from noise in the ChIA-PET experiment using the Bayesian approach. To evaluate and compare the performance of ChIAMM with the top existing methods, we used four short and two long-read ChIA-PET datasets. The short reads are RNAPII- and CTCF-associated datasets from human K562 and MCF7 cells, and the long reads are RNAPII- and H3K9me2-associated datasets from rice

Minghui 63 (MH63). We used human genome hg19 for K562 and MCF7 datasets and RS1 reference genome for rice datasets.

Convergence Diagnostics and Posterior Prediction

We used Stan statistical package (rstan) and checked the convergence of the algorithm with the trace plot and Rhat. The rstan package allows us to conveniently fit different models and access the outputs, including posterior inferences. In Bayesian inference, MCMC algorithms will draw a sample from the target posterior distribution after it has converged to equilibrium. However, there is no guarantee about whether it is converged or is close enough to the posterior distribution. Therefore, we have to check its convergence using a trace plot and Rhat. It is well known that trace plots are an essential tool for assessing the mixing of a chain. Trace plot is a time series plot of the Markov chains that shows the evolution of parameter vector over the iterations of one or many Markov chains. The Rhat produces the convergence diagnostic that compares the between- and within-chain estimates for model parameters. It is recommended to run at least four chains by default and use the sample if Rhat is <1.05 (Stan Development Team, 2016). The trace plot of intra- (β_j , λ_{0i} , W_{1i} , and C) and inter- (β_j , λ'_{0i} , W'_{1i} , and C') chromosomal interaction model parameters were checked. As we specified in the methodology, the parameters λ_{0i} , W_{1i} , λ'_{0i} , and W'_{1i} are pair specific. The convergence was checked on the random taken values. Here, as an example, we tested the convergence diagnostic and posterior prediction on MH63 RNAPII datasets. **Supplementary Figure 1** and **Supplementary Table 1** show the trace plot and Rhat value of the model parameters in the given datasets. The Rhat value of all parameters is 1, and chains are mixed well. Therefore, these results proved to us the convergence of the MCMC algorithm.

Posterior prediction is used to assess the fit between a model and the data. The fitted model has been validated using posterior predictive checks (PPCs) through simulating data from the model using parameters drawn from the posterior. The posterior prediction analysis was checked using a graphical prior and PPC plot. The PPC plot gives the graphical display that compares the observed data to the simulated data from the posterior predictive distribution. In **Supplementary Figure 2**, the dark line shows the distribution of the observed outcomes, and the lighter line shows

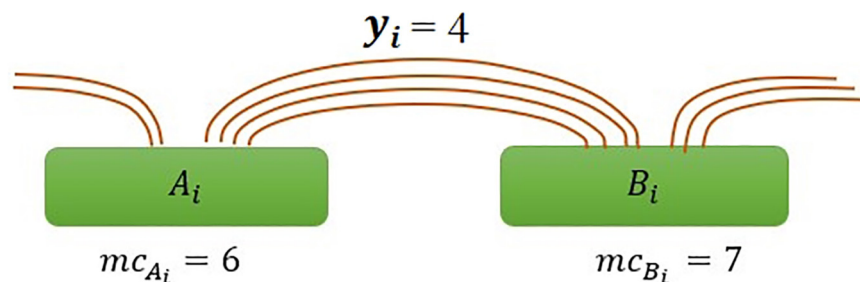


FIGURE 1 | Illustration of interaction frequency in the ChIAMM model. A_i and B_i represent anchor regions with marginal PET counts mc_{A_i} and mc_{B_i} , respectively, and y_i is the number of inter-ligation PETs between specified anchors A_i and B_i .

the first 100 kernel density estimate from the posterior predictive distribution in the MH63 RNAPII dataset. From the plot, the simulated data is overlapped with the actual data, or we assured that the fitted model recovered the data.

Comparing the Interactions of Short-Read Data From Different Methods

In this study, the ChIAMM found significant interactions using the value of W_{1i} (the probability of pair i being a true pair). The significant interactions from HG, ChiaSig, Mango, ChIA-PET2, and ChIA-PoP are found using the ChIA-PET Tool (V3), ChiaSig, Mango, ChIA-PET2, and ChIA-PoP pipelines, respectively. In all methods, we used the same cut-off of interaction frequency ≥ 3 . ChIAMM detected 1,465 and 3,679 potential pairs in MCF7 and K562 RNAPII datasets, respectively. These significant pairs are more than those identified by ChiaSig (828 in MCF7 and 1,828 in K562) and Mango (1,385 in MCF7 and 1,676 in K562). For CTCF-associated datasets, ChIAMM detected 719 and 2,085 significant pairs in the MCF7 and K562 datasets, respectively, which are more than those identified by ChiaSig (434 in MCF7 and 923 in K562). In contrast, some methods reported more interaction pairs than ChIAMM (Figure 2).

Supplementary Figure 3 shows the overlapped results between ChIAMM and other existing tools. As an example, in the MCF7 RNAPII dataset, we found higher overlapped interactions with HG (1,465), ChiaSig (1,334), and ChIA-PoP (1,113). Similarly, in the K562 CTCF dataset, it shows higher overlapped interactions with HG (2,084), ChIA-PoP (1,852), and ChiaSig (1,886). Besides, we found 257, 381, 387, and 1,047 overlapped significant interaction pairs among the six tools in MCF7 RNAPII, K562 RNAPII, MCF7 CTCF, and K562 CTCF datasets, respectively.

Aggregate Peak Analysis of the Interactions Between Different Methods

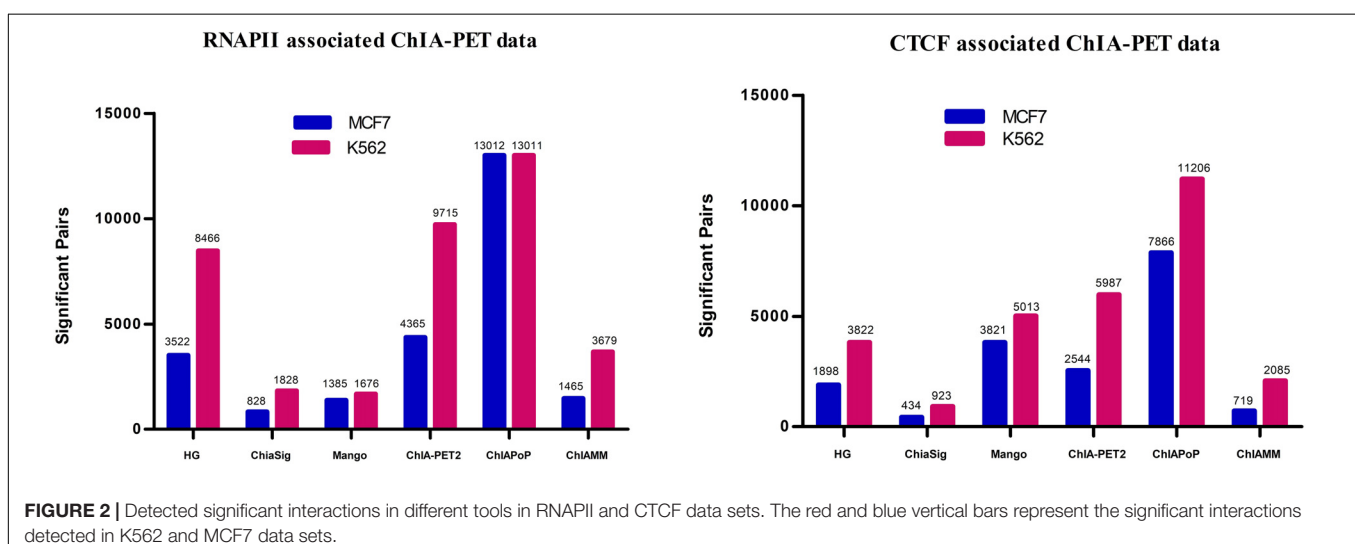
We used the APA plots to compare interactions from ChIAMM and other existing methods. To generate APA plots, we built

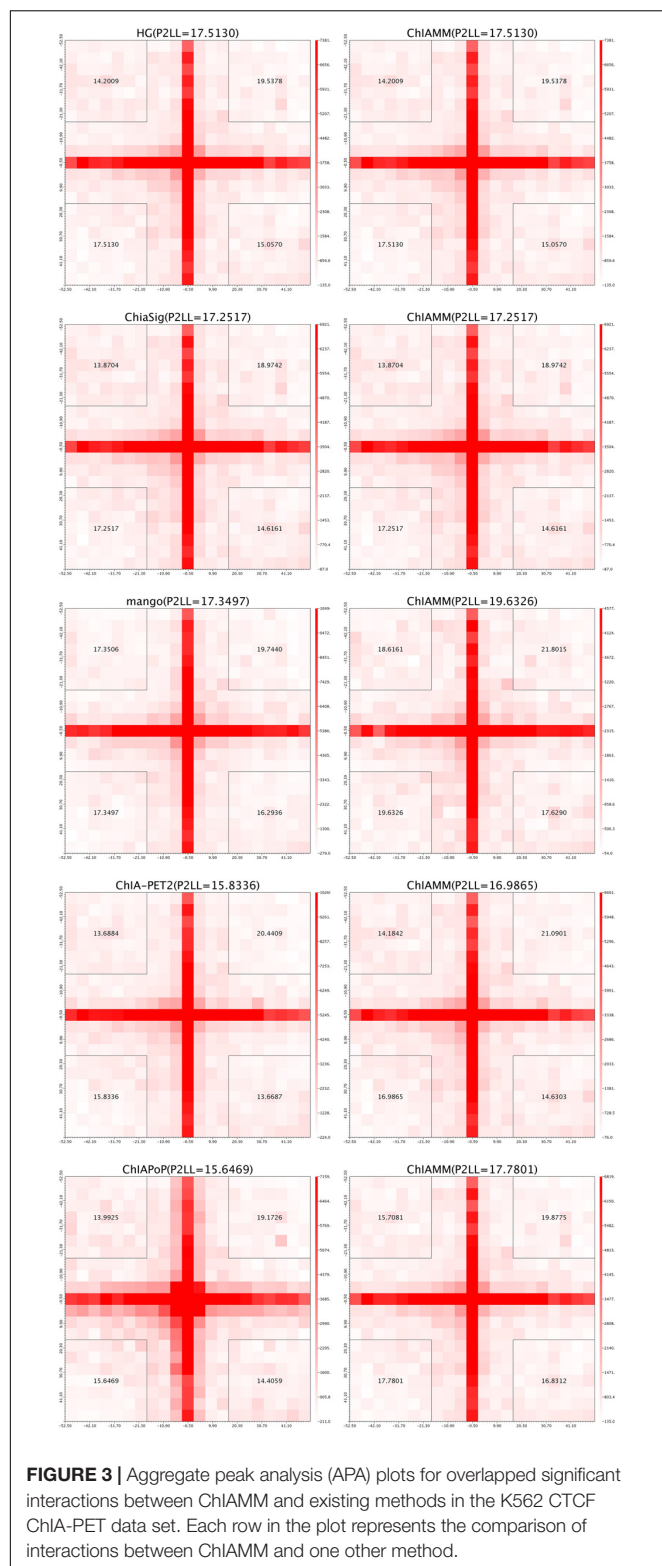
interaction matrices from BEDPE files, and the interaction counts were summed for all pairs of loci in 5-kb bins (Servant et al., 2015). Then, the APA score can quantify the level of a different set of interactions. In the APA plot, it is recommended to use P2LL value for comparison. P2LL is calculated as the ratio of the central pixel to the mean of the pixels in the lower-left corner of the interaction matrices. Higher scores indicate higher enrichment of interaction, and it is always good to find methods with higher P2LL value (Rao et al., 2014). For a fair comparison, in all methods, we considered the significant chromatin interactions with ≥ 3 supportive PETs. Then, we found the overlapped and unique significant interactions between ChIAMM and other existing tools.

For each dataset, we plotted five pairs of APA plot for overlapped interactions and four pairs of APA plot for unique interactions (no unique interactions found between ChIAMM and HG). In all datasets, in the overlapped interactions, ChIAMM has shown higher P2LL values with other tools. As expected, ChIAMM shows similar P2LL values with HG and ChiaSig tools (Figure 3 and Supplementary Figure 4). Besides, for unique interactions, ChIAMM has shown better pair ranking with other existing methods, with some exceptions, except Mango in K562 RNAPII, Mango in MCF7 and K562 in CTCF, ChiaSig in K562 RNAPII, and ChIA-PoP in MCF7 CTCF ChIA-PET datasets (Figure 4 and Supplementary Figure 5).

Comparison of CTCF Enrichment for Overlapped and Unique Interactions

In different studies, CTCF is a ubiquitously expressed and essential protein, and the DNA interactions are directly related to this protein (Ohlsson et al., 2010). For comparing enrichment of proteins in anchors, we used different CTCF peak files, i.e., the CTCF-peak regions from ENCODE ChIP-Seq datasets ENCFF720OXG and ENCFF990LUT for MCF7, and ENCFF681OMH and ENCFF559HEE for K562 cell line. For the CTCF coverage computation, we considered the overlapped and unique interactions between ChIAMM and other





existing methods with chromatin interaction frequency ≥ 3 . A comparison of CTCF enrichment means how many anchors are covered with the peak file. For both overlapped and unique interactions, we found the anchors that covered with the CTCF



peak file. **Supplementary Figure 6** shows the percentage of CTCF enriched and non-enriched anchors of the overlapped and unique interactions between ChIAMM and other methods in CTCF associated datasets. In these figures, ChIAMM shows equal CTCF enrichment with HG and ChiaSig in the overlapped interactions and shows a minimal difference with others. To ensure that this difference is statistically significant or not, we computed the Fisher's exact test. According to the p -value, in all datasets, the proportion difference of enriched anchors is statistically insignificant, except for ChIA-PoP in the overlapped interactions.

Comparison of CTCF Motif Orientation for Overlapped and Unique Interactions

It is well known that CTCF is an essential architectural protein to mediate long-range interactions. Different studies have shown that CTCF motif orientations at chromatin loop anchor regions are expected to have more convergent orientation than in other orientations (Zhang et al., 2018). Here, we compared the CTCF motif orientation of significant interactions (intrachromosomal) of ChIAMM with the existing tools. If the interaction is a real signal, it is expected to have convergent orientations more often than in other orientations. For the motif orientation analysis, a webserver <https://ccg.epfl.ch/pwmscan/> was used for scanning the reference genome (hg19), and the predicted CTCF motif was filtered and kept only the overlap result with CTCF peak regions. The CTCF peak files are the same as that we used in the previous CTCF enrichment comparison. Then, we found the overlapped result between the filtered predicted CTCF motif and significant chromatin interactions that we found using different tools. After that, we counted the number of significant pairs with convergent and other motif orientations. **Figure 5** and **Supplementary Figure 7** show the CTCF motif orientation analyses results for the overlapped and unique interactions in K562 and MCF7 CTCF datasets. The red color represents convergent motif orientation, and the blue color represents the other motif orientation. Fisher's exact p -values are given at the top of each bar. The p -value shows the test of a proportion of convergent motif orientation between ChIAMM and other existing methods. For each dataset, we performed five and four pairs (no unique interaction between ChIAMM and HG) of CTCF motif orientation analysis for overlapped and unique interactions between ChIAMM and existing methods, respectively. From these plots, in all datasets, ChIAMM showed equal motif orientation with ChIA-Sig (only in overlapped interactions) and HG. Statistically, the proportional difference in convergent orientation between methods was tested. Based on the p -value, in the overlapped interactions, the proportion of ChIAMM

motif orientation is not significantly different from other existing approaches, except ChIA-PoP. Likewise, in the unique interactions, it is statistically insignificant from others, except for Mango and ChIA-PET2.

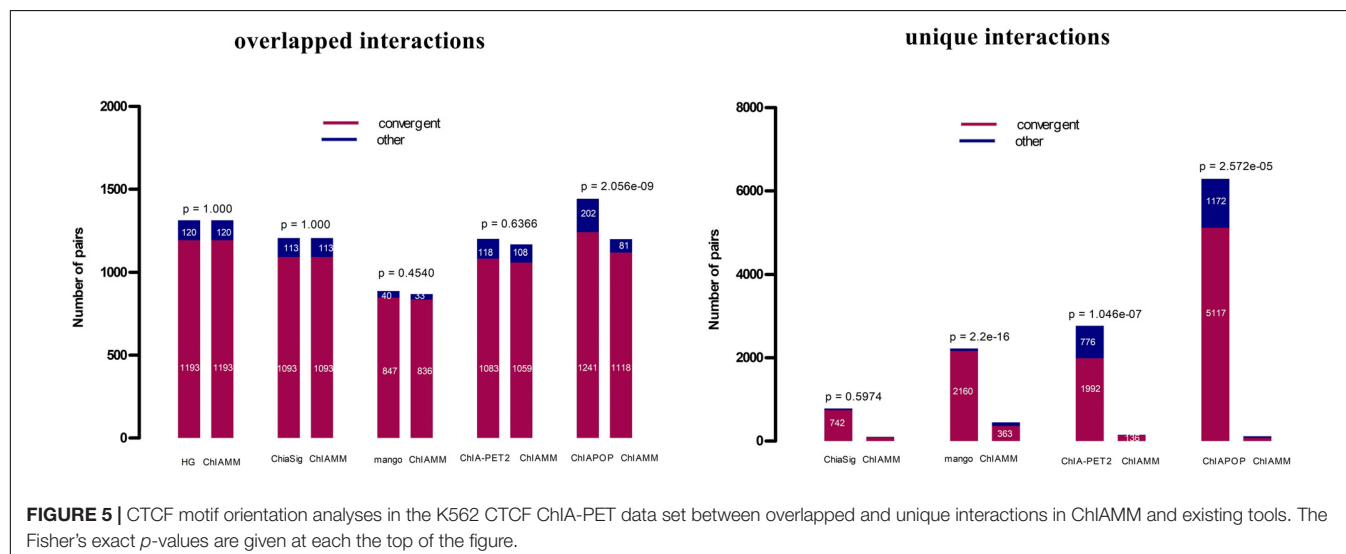
Comparing the Interactions of Long-Read Data From Different Methods

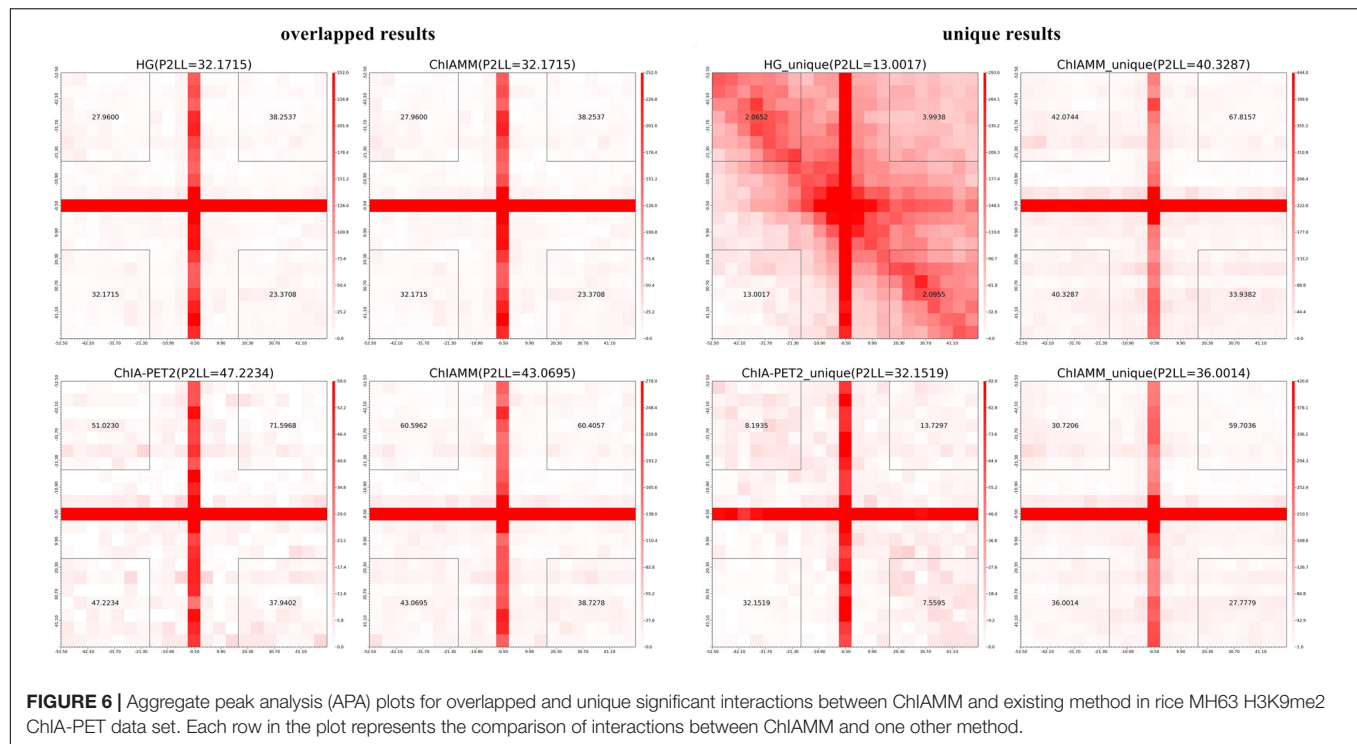
From the existing tools, only ChIA-PET Tool V3 and ChIA-PET2 can analyze long-read ChIA-PET data. Hence, we examined the result of ChIAMM with these two existing tools using the H3K9me2 and RNAPII datasets from rice MH63 variety. We used RS1 as the reference genome. In all methods, for a fair comparison, we considered the interaction frequency ≥ 3 . Similar to the short-read ChIA-PET datasets, we validated the interactions using the APA plot.

Chromatin interaction analysis using mixture model and other existing tools found the different amounts of significant chromatin interactions. **Supplementary Figure 8** shows the detected interactions in each tool; besides, it also shows the overlap interactions between ChIAMM and existing tools. HG found maximum significant chromatin interactions (63,745 and 6,242); ChIAMM found the next largest interactions (23,966 and 12,448); and ChIA-PET2 detected the smallest significant chromatin interactions (5,143 and 6,183) in MH63 RNAPII and H3K9me2 datasets, respectively. ChIAMM found maximum overlapped interactions with HG (23,821 and 2,903). The three tools found 2,744 and 969 overlapped significant chromatin interactions in MH63 RNAPII and MH63 H3K9me2 datasets.

Aggregate Peak Analysis of the Interactions Between Different Methods

To compare and evaluate ChIAMM in long-read ChIA-PET datasets, we generated the APA plot. Still, for the sake of fair comparison, we considered the chromatin interaction frequency ≥ 3 . We plotted the APA plots for overlapped and unique significant interactions between ChIAMM and other existing





tools. We plotted two pairs of APA plot for overlapped and unique interactions. **Figure 6** and **Supplementary Figure 9** show the APA plot for overlapped and unique interactions. In unique interactions, ChIAMM has shown higher P2LL values in both datasets. Besides, in the overlapped interactions, ChIAMM shows similar P2LL values with HG and lower P2LL values with ChIA-PET2 in H3K9me2 and RNAPII MH63 datasets.

DISCUSSION AND CONCLUSION

Chromatin interaction analysis by paired-end tag sequencing is a genome-wide, high-throughput, and high-resolution method to detect chromatin interactions associated with a specific protein of interest. Here, we described a new statistical approach called ChIAMM that corrects for non-specific interactions as a function of genomic distance, enrichment, GC content, and mappability score. It is designed for both short- and long-read ChIA-PET datasets. Using the RNAPII- and CTCF-associated data from human K562 and MCF7 cell and RNAPII- and H3K9me2-associated data from rice Minghui 63 (MH63), we demonstrated that our approach is better with the most effective top existing tools.

In various studies, enrichment, genomic distance, GC content, and mappability score were listed as systematic sources of bias. All the preexisting ChIA-PET tools considered only the genomic distance or enrichment as systematic biases. Therefore, all tools failed to address the possible biases in their study. Some are designed exclusively for short-read and only for intrachromosomal interaction ChIA-PET datasets. In this study,

we filled all the above gaps using the Poisson regression model. We considered the genomic distance, enrichment, GC content, and mappability score in the model, and we noticed its effect on the interaction frequency. **Supplementary Table 1** shows the estimated Poisson regression coefficients of biases in the MH63 RNAPII dataset. Each bias coefficient has a different sign and magnitude that tells the relationship type (positive or negative) and the degree of its effect, respectively. Enrichment and GC content, and mappability and genomic distance have a positive and negative effect, respectively. Besides, in the intrachromosomal interaction dataset, mappability and enrichment, and in the interchromosomal dataset, the GC content show a higher effect on loop detection.

Furthermore, some tools like Mango examined only intrachromosomal interaction. They removed all interchromosomal interactions in their model because they thought that interchromosomal interactions are the source of biases; besides, they could not find a technique that measures the genomic distance on different chromosomes. In this study, we dealt with these challenges via modeling inter- and intrachromosomal interaction data separately. This technique considered all four biases in the intrachromosomal interaction model and the three biases (we left out the genomic distance) in the interchromosomal interactions model. Using this technique, we salvaged essential significant interchromosomal interactions data rather than removal. Thus, this technique is a novel idea to consider interchromosomal interaction data into the study instead of total eradication.

Supplementary Table S2 shows the significant intra- and interchromosomal interaction (≥ 3) in various tools. Except for Mango and ChiaSig, other tools detected different amounts

of significant interchromosomal interactions. Comparatively, ChIA-PoP found the largest interchromosomal interactions; ChIAMM found 24, 28, 24, and 11 significant interchromosomal interactions from MCF7 RNAPII, K562 RNAPII, MCF7 CTCF, and K562 CTCF datasets, respectively. Therefore, discarding all interchromosomal data from the model is not a proper technique. It is considered as removed potential chromatin interaction from the analysis.

We compared ChIAMM results with the other five top existing tools using APA plot, CTCF coverage of anchors, and CTCF motif orientation. In the APA plot, we showed the performance of ChIAMM using overlapped and unique interaction frequency data. In all datasets, ChIAMM showed the highest enrichment of interaction with other existing methods, except Mango, an exceptionally conservative method, and it reports very few chromatin interactions. In the overlapped interactions, ChIAMM showed equal P2LL values with HG and ChiaSig, as expected, because ChIAMM and ChiaSig used ChIA-PET Tool as a primary processing pipeline, and this is also true for CTCF coverage and CTCF motif orientation analysis results. In CTCF coverage and motif orientation analysis, the new approach showed equal CTCF coverage and motif orientation with HG and ChiaSig in the overlapped interactions and relatively minimal differences with others. However, in almost all comparisons, the difference is statistically insignificant.

We compared the running time of ChIAMM with other preexisting methods. As an example, we analyzed the MCF7 CTCF ChIA-PET with threads, 12; RAM, 64 GB; cluster operating system, CentOS 6.6; central processing unit, Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50 GHz. ChIAMM took 48.1 min and showed better performance. ChIA-PET Tool, ChiaSig, Mango, ChIA-PET2, and ChIA-PoP took 17, 37, 36, 31, and 23 h, respectively. Overall, ChIAMM is the outperformed novel, fastest, and user-friendly tool than the most existing methods.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories

and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

YA and GL: conceptualization. YA: methodology and writing original draft preparation. YA and HJ: software. YA, SW, and JZ: data analysis. YA, GL, SW, and JZ: review and editing. GL and XN: supervision. GL: funding acquisition. All authors contributed to the article and approved the submitted version.

SOURCE CODE AND DOCUMENTATION

The source code and documentation for the ChIAMM is available at <http://www.guolianglab.org/subpages/RESOURCES/software.php> or <https://github.com/Yab29/ChIAMM>.

FUNDING

This research was supported by the National Natural Science Foundation of China (31771402 and 31970590) and the Fundamental Research Funds for the Central Universities (2662017PY116).

ACKNOWLEDGMENTS

We thank the group members from GL's Lab, especially Tongkai Sun and Stan online forum group, for helpful discussion and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.616160/full#supplementary-material>

REFERENCES

- Carlin, B. P., and Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. Boca Raton, FL: CRC.
- Derrien, T., Estellé, J., Sola, S. M., Knowles, D. G., Raineri, E., Guigó, R., et al. (2012). Fast computation and applications of genome mappability. *PLoS One* 7:e30377. doi: 10.1371/journal.pone.0030377
- Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., et al. (2009). An oestrogen-receptor- α -bound human chromatin interactome. *Nature* 462:58. doi: 10.1038/nature08497
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Boca Raton, FL: CRC.
- Halla-aho, V. (2015). *A Probabilistic Method for Quantifying Chromatin Interactions*. Helsinki: Aalto University.
- Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* 43:630. doi: 10.1038/ng.857
- He, C., Li, G., Nadhir, D. M., Chen, Y., Wang, X., and Zhang, M. Q. (2016). Advances in computational ChIA-PET data analysis. *Quant. Biol.* 4, 217–225. doi: 10.1007/s40484-016-0080-3
- He, C., Zhang, M. Q., and Wang, X. (2015). MICC: an R package for identifying chromatin interactions from ChIA-PET data. *Bioinformatics* 31, 3832–3834. doi: 10.1093/bioinformatics/btv445
- Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). HiCNorm: removing biases in Hi-C data via poisson regression. *Bioinformatics* 28, 3131–3133. doi: 10.1093/bioinformatics/bts570
- Huang, W., Medvedovic, M., Zhang, J., and Niu, L. (2019). ChIA-PoP: a new tool for ChIA-PET data analysis. *Nucleic Acids Res.* 47:e37. doi: 10.1093/nar/gkz062
- Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., et al. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* 9, 999–1003. doi: 10.1038/nmeth.2148
- Lee, B., Wang, J., Cai, L., Kim, M., Namburi, S., Tjong, H., et al. (2020). ChIA-PIPE: a fully automated pipeline for comprehensive ChIA-PET

- data analysis and visualization. *Sci. Adv.* 6:eaay2078. doi: 10.1126/sciadv.aay2078
- Li, G., Chen, Y., Snyder, M. P., and Zhang, M. Q. (2016). ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res.* 45:e4. doi: 10.1093/nar/gkw809
- Li, G., Fullwood, M. J., Xu, H., Mulawadi, F. H., Velkov, S., Vega, V., et al. (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.* 11:R22. doi: 10.1186/gb-2010-11-2-r22
- Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* 148, 84–98. doi: 10.1016/j.cell.2011.12.014
- Li, G., Sun, T., Chang, H., Cai, L., Hong, P., and Zhou, Q. (2019). Chromatin interaction analysis with updated ChIA-PET tool (V3). *Genes* 10:554. doi: 10.3390/genes10070554
- Li, X., Luo, O. J., Wang, P., Zheng, M., Wang, D., Piecuch, E., et al. (2017). Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat. Protoc.* 12:899. doi: 10.1038/nprot.2017.012
- Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369
- Niu, L., and Lin, S. (2015). A Bayesian mixture model for chromatin interaction data. *Stat. Appl. Genet. Mol. Biol.* 14, 53–64. doi: 10.1515/sagmb-2014-0029
- Ohlsson, R., Lobanenkov, V., and Klenova, E. (2010). Does CTCF mediate between nuclear organization and gene expression? *Bioessays* 32, 37–50. doi: 10.1002/bies.200900118
- Paulsen, J., Rodland, E. A., Holden, L., Holden, M., and Hovig, E. (2014). A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. *Nucleic Acids Res.* 42:e143. doi: 10.1093/nar/gku738
- Peng, Y., Xiong, D., Zhao, L., Ouyang, W., Wang, S., Sun, J., et al. (2019). Chromatin interaction maps reveal genetic regulation for quantitative traits in maize. *Nat. Commun.* 10:2632. doi: 10.1038/s41467-019-10602-5
- Phanstiel, D. H., Boyle, A. P., Heidari, N., and Snyder, M. P. (2015). Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics* 31, 3092–3098. doi: 10.1093/bioinformatics/btv336
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021
- Rousseau, M., Fraser, J., Ferraiuolo, M. A., Dostie, J., and Blanchette, M. (2011). Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics* 12:414. doi: 10.1186/1471-2105-12-414
- Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., et al. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16:259. doi: 10.1186/s13059-015-0831-x
- Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 163, 1611–1627. doi: 10.1016/j.cell.2015.11.024
- Stan Development Team (2016). Stan Modeling Language Users Guide and Reference Manual. Technical Report. Available online at: <https://mc-stan.org/users/documentation/>
- Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* 43:1059. doi: 10.1038/ng.947
- Zhang, R., Wang, Y., Yang, Y., Zhang, Y., and Ma, J. (2018). Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* 34, i133–i141. doi: 10.1093/bioinformatics/bty248
- Zhao, L., Wang, S., Cao, Z., Ouyang, W., Zhang, Q., Xie, L., et al. (2019). Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. *Nat. Commun.* 10:3640. doi: 10.1038/s41467-019-11535-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Arega, Jiang, Wang, Zhang, Niu and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Refinement of Draft Genome Assemblies of Pigeonpea (*Cajanus cajan*)

Soma S. Marla^{1*†}, Pallavi Mishra^{1†}, Ranjeet Maurya^{1†}, Mohar Singh¹,
Dhammaprakash Pandhari Wankhede¹, Anil Kumar², Mahesh C. Yadav¹, N. Subbarao³,
Sanjeev K. Singh¹ and Rajesh Kumar¹

¹ Indian Council for Agricultural Research (ICAR)-National Bureau of Plant Genetic Resources, New Delhi, India, ² Directorate of Education, Rani Lakshmi Bai Central Agricultural University, Jhansi, India, ³ School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Subhojit Datta,
Indian Council of Agricultural
Research, India
Vishal Acharya,
Institute of Himalayan Bioresource
Technology (CSIR), India

*Correspondence:

Soma S. Marla
soma.marla@icar.gov.in;
ssmarl@yahoo.com

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 17 September 2020

Accepted: 23 November 2020

Published: 15 December 2020

Citation:

Marla SS, Mishra P, Maurya R,
Singh M, Wankhede DP, Kumar A,
Yadav MC, Subbarao N, Singh SK
and Kumar R (2020) Refinement of
Draft Genome Assemblies
of Pigeonpea (*Cajanus cajan*).
Front. Genet. 11:607432.
doi: 10.3389/fgene.2020.607432

Genome assembly of short reads from large plant genomes remains a challenge in computational biology despite major developments in next generation sequencing. Of late several draft assemblies have been reported in sequenced plant genomes. The reported draft genome assemblies of *Cajanus cajan* have different levels of genome completeness, a large number of repeats, gaps, and segmental duplications. Draft assemblies with portions of genome missing are shorter than the referenced original genome. These assemblies come with low map accuracy affecting further functional annotation and the prediction of gene components as desired by crop researchers. Genome coverage, i.e., the number of sequenced raw reads mapped onto a certain location of the genome is an important quality indicator of completeness and assembly quality in draft assemblies. The present work aimed to improve the coverage in reported *de novo* sequenced draft genomes (GCA_000340665.1 and GCA_000230855.2) of pigeonpea, a legume widely cultivated in India. The two recently sequenced assemblies, A1 and A2 comprised 72% and 75% of the estimated coverage of the genome, respectively. We employed an assembly reconciliation approach to compare the draft assemblies and merge them, filling the gaps by employing an algorithm size sorting mate-pair library to generate a high quality and near complete assembly with enhanced contiguity. The majority of gaps present within scaffolds were filled with right-sized mate-pair reads. The improved assembly reduced the number of gaps than those reported in draft assemblies resulting in an improved genome coverage of 82.4%. Map accuracy of the improved assembly was evaluated using various quality metrics and for the presence of specific trait-related functional genes. Employed pair-end and mate-pair local libraries helped us to reduce gaps, repeats, and other sequence errors resulting in lengthier scaffolds compared to the two draft assemblies. We reported the prediction of putative host resistance genes against *Fusarium* wilt disease by their performance and evaluated them both in wet laboratory and field phenotypic conditions.

Keywords: assembly improvement, reconciliation, mate-pairs, disease resistance, pigeonpea genome

INTRODUCTION

Recent rapid developments in genome sequencing technologies have facilitated the generation of several draft assemblies in plants. These are valuable resources for elucidating genetic information and understanding the biology of the crop. However, each of these draft assemblies have strengths and weaknesses as they were sequenced and assembled based on different technologies and algorithms (Singh et al., 2012; Varshney et al., 2012). Draft assemblies differ depending on the sequencing technology and the assembly software employed. One assembly may be conservative in its selection of reads resulting in low genome coverage with many gaps. Another assembler may be vigorous, yielding more contigs but with many errors. Draft genomes are typically sets of a large contingent of assembled contigs and scaffolds that are often fragmented due to the presence of a large number of gaps interlaced by repetitive regions. Often in a misassembly different contigs are improperly joined. The contig mis-join problem arises due to inversions, relocation, or a translocation. Gaps arise also due to incorrect insertion or deletion of a particular sequenced read. These changes often result in the wrong placement of a contig onto a scaffold belonging to a different chromosome. Hence, the annotation of unfinished and partially assembled genomes creates ambiguities while accessing complete genetic information as desired by biologists.

Some reasons for incompleteness include: 1. gaps appearing due to polymorphisms in complex genomes where reads on either side of a gap represent two haplotypes that belong to two separate chromosomes, 2. an abundance of repeat elements that confuse the assembler and leave some gaps unfilled, and 3. lack of a sufficient number of reads to cover the part of the genome, requiring an additional library of reads to fill the gaps. Besides, in draft genome assembly base call errors, variations in read coverage depth also cause gaps and pose serious computational challenges while connecting nodes in a *De Bruijn* graph (Guizelini et al., 2016).

Complex eukaryotic genomes are known to contain a large volume of nearly identical copies of DNA repeats and fragments. Various types of repeats present in genomes of wheat, pigeonpea, maize, or potato include transposable elements, highly conserved gene clusters, and segmental duplications. The presence of identical DNA fragments further complicates computational assembly. During pre-assembly, short reads of equal sizes tend to be masked together and complicate the construction of a *De Bruijn* graph (Compeau et al., 2011). Recently introduced third generation single molecule real time technologies (Ardui et al., 2018) and Oxford nanopore technologies (Brown and Clarke, 2016) generate large sized reads which can readily be inserted to fill gaps caused by repetitive elements. Despite virtues, such as low levels of sensitivity and the high sequencing error rates of long read technologies, many plant researchers are opting to use short read sequencing technologies for financial reasons.

Two draft *de novo* genomes compared in the present study are short read assemblies generated from second generation sequencing technologies. Apart from assembly complexity due to smaller reads, repeat abundance also obviates gap closing and is

often responsible for the resulting low levels of genome coverage reported in draft assemblies. Modern sequencing platforms generate paired end or mate-pair read libraries. The mate-pair libraries are generated in different sizes and orientations (ranging from 3 to 5 bp and even up to 0.5 kb). They serve as potential inserts while filling gaps. Mate-pair libraries are recommended as a potential approach to mitigate repeats in computational genome assembly (Wetzel et al., 2011; Wang et al., 2012; Grau et al., 2019). In the present work, we demonstrated the application of mate pairs for gap closing during meta-assembly, that resulted in significant improvement of both the genome coverage and quality of the improved pigeonpea assembly.

Major techniques recommended for gaining contiguity and higher coverage in draft genomes broadly include, use of long inserts for gap filling (Wetzel et al., 2011), assembly reconciliation, hybrid assembly (Wang et al., 2012), filtering repeats, and iterative mapping using short reads to close the remaining gaps (Tarailo-Graovac and Chen, 2009; Tsai et al., 2010). The use of paired end or mate pairs for filling the gaps is a robust computational approach. The reconciliation approach (Alhakami et al., 2017) for closing gaps and correcting misassemblies involves comparing available data sets from different draft genomes of the same or related species, mapping their common reads, and finally merging them together to gain improved scaffold lengths with higher contiguity (Kumar et al., 2018; Mishra et al., 2019).

Pigeonpea [*Cajanus cajan* (L.) Millsp.] is a major food legume grown in India and is a diploid ($2n = 22$) with a genome size of 833.07 Mbp (Varshney et al., 2012). It is a widely cultivated pulse crop and a major source of dietary proteins in India with an annual production of 2.31 mt and productivity of 678 kg/ha (Pande et al., 2013). Prevailing low crop productivity may be attributed to the absence of high yielding cultivated varieties possessing resistance to various pests and diseases. In plants, resistance genes (*R* genes) play important roles in the recognition and protection from invading pests and pathogens. A few sources of resistance to biotic stresses can be found in available germplasm collections. Resistance genes are identified and found primarily organized in individual clusters that are strictly linked across the genome (David et al., 2009). Modern plant breeding techniques include marker-assisted breeding and genomic selection-accelerated development of superior crop varieties with the use of genomic resources and genetic information emitting from sequenced genome projects. The pigeonpea genome was *de novo* sequenced independently by two research groups (Singh et al., 2012; Varshney et al., 2012). These draft assemblies were made available in the public domain (GCA_000340665.1 and GCA_000230855.2), and are valuable resources for breeders. However, both the assemblies are incomplete with a sizable number of fragmented contigs and existing gaps. The lack of accurate genetic information is a major limitation for the prediction of gene complement components associated with desirable traits. Hence, the primary objective of the present work is to generate a more contiguous and complete assembly with improved genome coverage. We report an improved version of a improved assembly based on the genome reconciliation approach that first compares

the two available draft assemblies, and scores the matching blocks at each location followed by their merger. The meta-assembler tool employed in the present study detected a significant number of gaps and filled them iteratively using right-sized inserts from local pair-end and mate-pair libraries. The correctness of the mate pairs chosen by the meta-assembler during error correction was further validated by the mapping and alignment algorithm BIMA (Drucker et al., 2014). Completeness and map accuracy of the reconstructed assembly was verified for the presence of conserved plant resistance genes (*R* genes). Here we report the prediction of putative *R* genes, their isolation, and PCR screening of a known resistant cultivar against *Fusarium* wilt disease in both laboratory and field conditions.

RESULTS

Improvement of the Draft Genome Assemblies Employing the Reconciliation Algorithm

The reconciliation assembly approach was employed in the present work to refine the incomplete draft genome assemblies, A1 and A2. The assembly tool hybridSPAdes (Antipov et al., 2016) was employed for the selection of optimum k-mers, with evaluated combinations ranging from 21 to 55. We observed that k-mer sizes of 21, 33, 55, and 77 yielded superior assemblies with few fragmented sequences, a smaller number of contigs with high N50, and mean and median scaffold lengths in superior assemblies. The meta-assembler was employed for merging the two assemblies. Merged Illumina HiSeq sequences resulted in 46,979 reads with the N50 length of 24,087. The meta-assembler implemented the reconciliation algorithm to refine and obtain a reconstructed genome. In order to capture the suitable reference assembly set for alignment during the merger process, we examined the required order in which assemblies A1 and A2 were to be chosen as the master set (GCA_000230855.2) and slave sets for alignment with the former (GCA_000340665.1). We observed that choosing A1 as the master set with A2 as the slave set resulted in a highly contiguous superior assembly. The superiority of the merged meta-assembly was systematically evaluated with compression-expansion (CE) statistics. Gaps present in the scaffolds were closed using mate pairs. Gap sizes estimated by the LG_Gapcloser (Xu et al., 2019) were passed on to the next round of alignment. To locate suitably sized inserts, gaps were compared with mate-pair libraries employing BLAST (Altschul et al., 1990) and the single highest scoring mate-pair sequences were chosen. Gap closing mate pairs for gap closing ranged from 200 bp (lower side) to 1,350 bp with 500 bp as the mean size. Mate pairs used by the meta-assembler for gap closing were further validated by mapping and the alignment algorithm BIMA (Drucker et al., 2014).

The remaining gaps were filled by searching unique contig end sequences against unused reads. Analysis of the repeat composition and the identification of their size variations in turn

aided the significant reduction of gaps and contributed to the prediction of specific genes. The improved assembly had 46,979 contigs with a total size of 548.2 Mb covering 82.4% of the genome with high contiguity (Table 1).

Read Mapping

Read mapping increased in the improved meta-assembly from 75.6 to 72.7% in the two compared misassemblies to 82.4% in the improved meta-assembly. A higher number of reads were found to be mapped to the merged assembly compared to those in the A1 and A2 misassemblies. Mapping depth is a measure of the number of reads used for aligning the improved genome. It also helps to estimate the extent of similarity between the improved assembly and the compared misassemblies. Among the two draft assemblies, A2 was superior to A1 in the depth of read coverage. A relatively higher read depth in the A2 misassembly can be attributed to the high-identity Illumina reads used both in the initial assembly and in the later polishing steps. Our final assembly in terms of depth of coverage was superior to A2, with more gaps filled. In addition, the refined assembly had more GC-rich regions (Table 1) with improved gene component predictions. Many gap sequences with high AT composition were eliminated. The total GC content in the A1 and A2 assemblies were 37.2 and 32.8%, respectively and were enhanced to 45.5% in the meta-assembly reported in the present work. The improvement in GC rich fraction and of the N50 values in both contigs and scaffolds in the improved genome were achieved largely due to the application of mate-pair read sequence libraries for gap filling. High GC content is known to be associated with the concentration of coded genes in certain regions in a genome (Rao et al., 2013). In the present study, observed high GC content was obtained in the refined assembly A3 which appeared to be related to an increased number of predicted genes in the improved genome of pigeonpea.

Meta-Assembly, Annotation, and Quality Assessment

Two draft assemblies were compared, merged, and reassembled employing the two approaches as described above. We initially used unpaired reads for the assembly adopting an overlapping read approach. As no significant improvement was observed in read mapping depth, and eventual coverage, we resorted to available mate paired libraries to close the gaps. We used variable mate pairs during different alignment steps in the meta-assembly and succeeded in resolving repeat problems.

We wanted to ascertain which type of mate-pair libraries effectively resolved the repeat problem. In the assembly, we employed a meta-assembler (Wences and Schatz, 2015). In the first experiment, we only used a 648 Mb library and in the second experiment a 605 Mb and a 548 Mb library were taken together. Initially we used all the single paired read data sets available (minus two mate-pair data sets) of A1 along with all the data sets from A2. In the second treatment, we included two mate-pair data sets from A1 along with all the data available from A2. At the end, all

TABLE 1 | Genome assembly statistics of draft assemblies A1, A2 and improved A3 assembly.

Parameter	A1, assembly GenBank accession: GCA_000340665.1	A2, assembly GenBank accession: GCA_000230855.2	A3, improved assembly GenBank accession: GCA_015227855.1
Number of contigs	360,028	72,923	46,979
Contig N50	5,341	22,480	24,087
Contig L50	30,054	7,524	6,925
Number of scaffolds	NA	36,536	13,101
Scaffold N50	NA	555,764	574,622
Scaffold L50	NA	72	57
Total scaffold length	NA	592,970,700	548,600,000
Number of gaps	NA	72,774	36,561
Range of mate-pair sizes used	NA	NA	20–1,350 bp
Mean size of mate pairs used in gap closure	NA	NA	500 bp
Number of Ns	NA*	34,435,295	34,188,871
Genome coverage	199×	160×	174×
Percentage mapping	75.6%	72.7%	82.4%
GC content	37.2%	32.8%	45.5%
File size (Mb)	648 Mb	605 Mb	548 Mb

Data source: <https://www.ncbi.nlm.nih.gov>. *"N"s masked.

the output values and statistical metrics were collected for comparative performance analysis. We observed that all the available pigeonpea mate-pair libraries taken together resulted in the improvement of genome coverage. It is presumed that the incorporation of variable size mate-pair inserts helped in gap closing during the assembly.

In our final assembly, the contig N50 increased to 24,087, and scaffold N50 increased by 574,622. The total number of gaps decreased across the genome by 50.23% (Table 1). It was observed that the order in which the input draft assemblies were inputted into the meta-assembler drastically influenced the alignment quality and the resulting read coverage (Lindner et al., 2013). In the primary assembly, we treated assembly A1 as the master and aligned it with assembly A2. In the other variant, we used assembly A2 as the master and aligned it against assembly A1. The output resulted in a primary assembly that yielded us a scaffold length of 548,600,000.

Closure of Repeat-Derived Gaps

For each round of alignment undertaken between the A1 and A2 misassemblies, the meta-assembler built a graph, with vertices of the above alignments and edges joining the two alignments. If both had the same direction, they were readily rearranged into a single block thus providing contiguity. In this case, where the examined genomic segments from two misassemblies did not share the same direction, it indicated that there was an existing distance between them and the prevailing gaps needed to be filled. In such cases, variable size local pair-end and mate-pair libraries could offer right inserts to fill these gaps. While building the graph, the meta-assembler searched the mate-pair library for right sized inserts in order to complete the shortest path between any two of the contigs to fill a gap.

We evaluated the closure performances of the LR Gapcloser (Xu et al., 2019) and Gapfiller (Boetzer and Pirovano, 2012) tools

on the repeat-derived gaps. We first tested the performance of each tool using the raw mate-pair reads. Both the above tools used first raw pair-end and mate-pair libraries. We monitored the gap closure efficiency by evaluating the number of gaps closed applying indexed and hashed mate-pair libraries (Drucker et al., 2014). In the merged pigeonpea assembly, we estimated 37,145 repeat-derived gaps of which 584 gaps and 322,780 nucleotides out of a total 34,511,651 were closed. The gap sizes ranged from 20 to 15,510 bp. LR Gapcloser was more efficient in filling most of the gaps, achieving 82.4% and with low error rates.

We achieved improved contiguity by using long mate pairs to fill the gaps in the assembly and thereby achieved higher coverage in the improved assembly. Draft assembly A1 had 360,028 contigs with an N50 and L50 of 5,341 and 30,054, respectively. The reported genome coverage was 199×

Completeness of the Merged Assembly

BUSCO (Simao et al., 2015) was employed to evaluate the completeness of the conserved proteins in all three assemblies. The A3 assembly was found to be 94.02% complete. Of the total 1,440 BUSCO groups that were searched, the meta-assembly was found to contain 1,321 complete single-copy (S) BUSCOs, 33 complete duplicated (D) BUSCOs, 57 fragmented (F) BUSCOs, and 29 missing (M) BUSCOs. Whereas comparatively the A1 and A2 assemblies were 85.27% (S:76.87%, D:8.40%, F:5.62%, M:9.09%) and 87.9% (S:80.9%, D:7%, F:5%, M:7.1%) complete, respectively (Supplementary Table 1). The gene completeness score as measured by BUSCO relatively increased in the improved assembly, while the numbers of fragmented

TABLE 2 | Repetitive sequences of draft assemblies A1, A2, and the improved A3 assembly.

Transposable elements	A1 assembly	A2 assembly	A3 assembly
Retro transposons	77,096,057	116,194,477	89,089,240
Gypsy	52,354,920	71,402,096	59,247,991
Copia	19,937,308	37,676,825	24,339,237
Line	5,261,337	6,717,918	5,914,324
Unclassified elements	216,262,607	169,378,278	158,228,382
DNA transposons	9,772,250	27,455,193	19,826,943
Total transposable elements	303,130,914	313,027,948	267,144,565

and missing BUSCO genes were reduced. This genome comparison can be used to help such draft assemblies toward becoming finished.

Functional Annotation of Predicted Gene Content

The FGENESH module of the Molquest v.4.5 software package¹ and Augustus were employed and 51,737 genes were predicted for the improved meta-assembly. The number of predicted genes was less compared to A1 but higher than A2.

In the total gene component prediction, we found 1,303 disease resistance related genes in pigeonpea. The improved assembly yielded a total of 51,737 genes which was less than A1 but more than those reported in the A2 assembly. The variable number of predicted genes observed in the draft assemblies can be attributed to split genes and overestimation during gene finding (Denton et al., 2014). The overestimation of gene numbers often result when fragmented single genes are present on multiple contigs or scaffolds (Pozzi and Salamini, 2007). Improvements in gap filling and read mapping depth resulted in the reduction of the number of genes in meta-assembly A3. The predicted total gene number was less in A3 than in A1 but was slightly higher than the A2 draft assembly (Table 2). In the predicted gene set 54-resistance single copy putative genes containing known conserved domain NBS LRR were selected and *in silico* mapped onto the corresponding chromosomes (Supplementary Table 2).

Identification of Repetitive Sequences and Transposable Elements in the Improved Assembly

Repeat elements are extra copies of DNA sequences generated and planted at various locations in the genome to meet certain challenges and improve the fitness of the organism during the course of evolution. Repetitive elements in pigeonpea occupy nearly half of the genome of *Cajanus cajan* (Macas et al., 2015). Repeats pose many computational challenges in read alignment and assembly (Xu et al., 2019), such as the creation of gaps and overlaps and leads to many mapping inaccuracies in misassemblies. One can always filter and exclude the reads but it is essential to map them onto chromosomal locations where gaps exist. Mate-pair libraries were used for resolving repeat

¹<http://www.softberry.com>

TABLE 3 | Results of gene search.

Parameter	A1 assembly	A2 assembly	A3 assembly
No. of genes predicted	56,888	48,680	51,737
Putative resistance gene analogs against <i>Fusarium</i> wilt	NA	NA	54

problems and obtaining contiguous scaffolds in both prokaryotic (Wetzel et al., 2011) and eukaryotic organisms (Grau et al., 2019). Meta-assembler searches for contigs that can be placed in the gap using mate pairs, and then again checks to see if there exists a recorded shortest path between any of these contigs. In an assembly, overlapping reads are used as edges to connect reads belonging to the same region of the genome. However in complex genomes like pigeonpea, the abundance of repeats cause coverage gaps and read errors thus leaving numerous gaps to fill between contigs while scaffolding. The filling of gaps requires the adoption of robust computational approaches to affectively address repeat problems. Sequenced pair-end and mate-pair reads can potentially bridge gaps efficiently in order to orient contigs by estimating the gap lengths to the edges while filling the scaffolding graph (Ghurye and Pop, 2019).

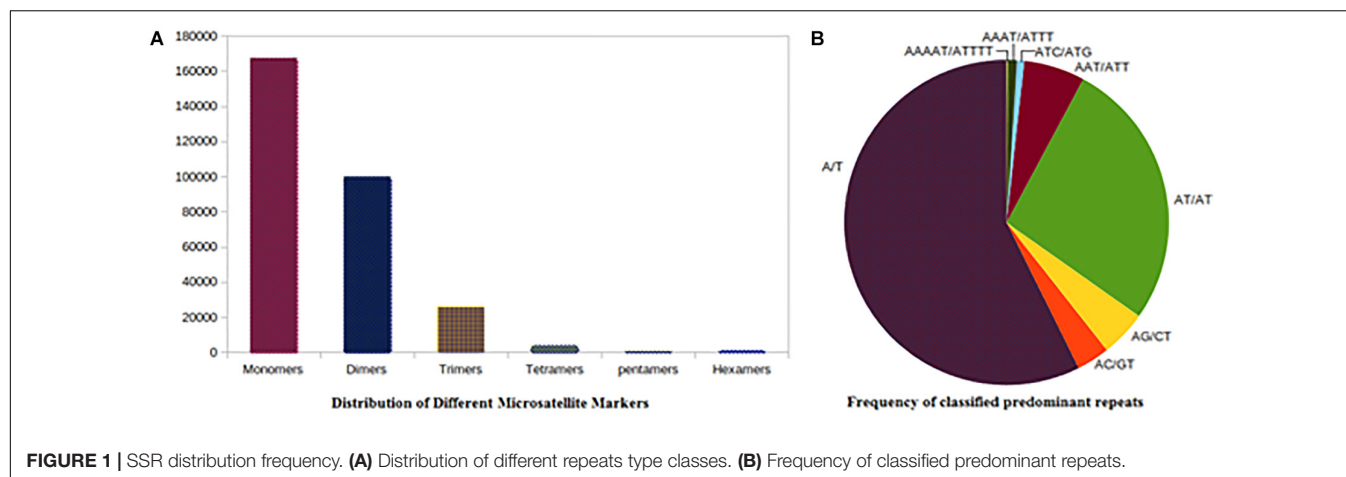
A high level of assembly was achieved using mate-pair reads in wheat, a genome ridden with a large number of repeats (Clavijo et al., 2017). We analyzed the repeat content in comparison to the A1 and A2 assemblies and divided them into various classes (Table 3). Among the different classes identified, transposable elements were found to be rich in AT elements in A3. In the course of the iterative use of reads during assembly, we observed transposon-derived repeats collapse against identical reads resulting in the closure of significant portions of gaps. Similar observations were reported on gap filling using retro transposon-related repeats in human genome assembly (Marin et al., 2018).

Identification of Microsatellites

The improved pigeonpea assembly was mined for simple sequence repeats (SSR). A total of 297,294 were simple repeats out of the total 298,732 repeats and the remaining 1,438 belonged to complex types (Table 4). Mononucleotide repeats were abundant with 56.05% of the total SSRs mined. Dinucleotides occupied 33.45% (99,949), 8.72% (26,069) were trinucleotides, and 1.27% were tetranucleotides (3811) repeats. The remaining SSRs were of the complex type, 0.25% were hexa nucleotides and 0.22% were penta.

TABLE 4 | Results of microsatellite search in the improved pigeonpea assembly A3.

Total number of sequences examined	13,101
Total size of examined sequences (bp)	584,435,790
Total number of identified SSRs	298,732
Number of SSR containing sequences	6,494
Number of sequences containing more than 1 SSR	4,603
Number of SSRs present in compound formation	41,002



Among the 167,465 mononucleotide repeats, the mononucleotide motifs were in majority with A/T repeats of 98.25 and 1.74% were occupied by C/G types. Among the 99,949 dinucleotides microsatellites, the AT/AT type (77.34%) of microsatellites were most common in the genome followed by the AG/CT type (13.21%), and the AC/GT type (9.40%). The CG/CG type dinucleotides microsatellites were present in a very low proportion (0.03%). In trinucleotide SSRs repeats (26,069), around 66.71, 12.31, 8.07, and 5.98% of SSRs were of AAT/AAT, AAG/CTT, ATC/ATG, and AAC/GTT types, and were most abundant, respectively. Among the other types of repeats, the ACG/CGT type was lowest (0.36%) in the genome of pigeonpea. The highest distribution (68.06%) of tetra nucleotides microsatellites was present in the genome of pigeonpea. The maximum numbers of predominant SSRs repeats were of the A/T type followed by AT/AT, AG/CT, AAG/CTT, AAT/ATT, and AAAAT/ATTTT (Supplementary Table 3). The overall analysis showed that the relative abundance of tetra, penta, and hexa SSRs types were low as compared to mono, di, and tri SSRs types in pigeonpea genome sequences (Figure 1).

Characterization and Synteny Analysis of Pigeonpea NBS-LRR Like Resistance Gene Analogs

We verified the presence of already known conserved disease resistance gene families in the refined meta-assembly. The reported resistance (R) genes containing nucleotide-binding site (NBS)-leucine rich repeat (LRR) protein sequences from other important legume genomes were downloaded from Phytozome (Goodstein et al., 2012). The comparison of the predicted coding sequences against bean (*Phaseolus vulgaris*) clusters resulted in the identification of more than 100 resistance gene analogs (RGA). An annotation of the mined predicted genes revealed the presence of known disease resistance domains, such as ARC-NBS-LRR, transmembranes, and kinases. Nucleotide-binding sites (NBS) containing disease resistance genes play an important role in defending plants from a variety of pathogens and insect pests. Many R genes have been identified in various plant species including the pigeonpea genome (Singh

et al., 2012; Varshney et al., 2012). However, functional R genes targeting specific diseases in pigeonpea have not been reported. In this study, an improved A3 meta-assembly using computational analysis of the refined genome identified NBS-LRR resistance (R) proteins. The 1,301 mined putative resistance gene analogs were shown to share up to 78% of their homology with soybean, chickpea, barrel clover, field bean, and other species (Supplementary Table 4). Of them, 251 NBS-LRR domains containing pigeonpea resistance gene analogs were selected. The RGAs had a high amino acid identity in the identified putative pigeonpea disease resistance genes, which showed a high level of proteins in *Glycine max* with several sequences with high homology up to (77–98%) (Supplementary Table 5). We identified 54 NBS-encoding single copy genes and characterized them on the basis of structural diversity and conserved protein motifs.

Synteny analysis revealed significant relationships among the selected legume genomes. *Glycine max* and *Medicago truncatula* genomes revealed the presence of a high level of extensively conserved regions among pigeonpea and other legumes. We observed that nearly 89–91% of the pigeonpea assembly showed significant signs of RGA conservation with other legumes, viz., 41 NBS-LRR orthologs in *Glycine max* and 73 NBS-LRR orthologs in *Medicago truncatula*. A total of 57% of NBS-LRR pigeonpea genes were identified for the closely related organisms. *Glycine max* was found to share the largest number of extended conserved syntenic blocks with *Cajanus cajan* indicating its recent ancestry, followed by *Medicago truncatula*. The reported A3 meta-assembly of pigeonpea comprises 251 R gene homologs of the disease resistance gene, of which 229 are anchored to different pseudomolecules of pigeonpea. Of these, 23 genes are distributed to 57 collinear blocks between pigeonpea and the *Glycine max* genomes displaying a high level of collinearity (Figure 2). Overall, all pigeonpea RGAs displayed extensive collinearity with the different chromosomes of *Glycine max* and *Medicago truncatula*. Synteny analysis revealed homologous blocks connecting chr4 in *C. cajan* with chr4 of *G. max*; chr11 of *C. cajan* with chr20 and chr17 of *G. max*; and chr3 of *C. cajan* with chr19 in *G. max*. Similarly, comparative analysis reported in draft assembly A2 (Varshney et al., 2012) confirms the presence

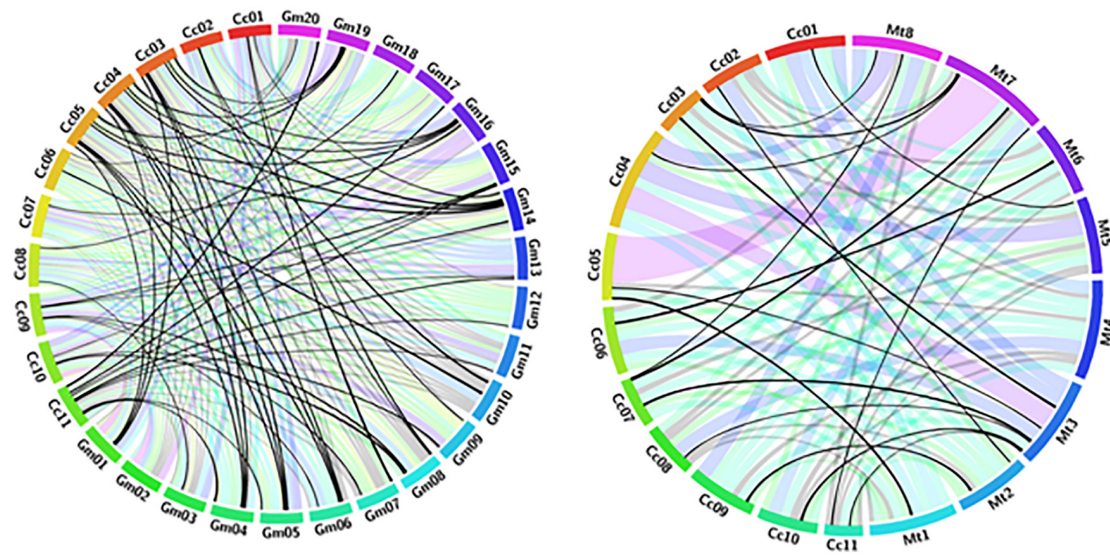


FIGURE 2 | Circos diagram presenting syntenic relationships between NBS-LRR containing R gene proteins in pigeonpea (Cc), *Glycin max* (Gm), and *Medicago truncatula* (Mt) pseudomolecules. Pseudomolecules of the two target species were labeled as Gm01–20 and Mt1–8. Pigeonpea pseudomolecules are labeled in different colors and labeled as Cc01–11. Collinear blocks are colored according to the color of the corresponding pigeonpea pseudomolecules. Each ribbon radiating block from a pigeonpea pseudomolecule represents a NBS-LRR similarity block between pigeonpea and other legumes.

of homologous blocks connecting chr3 in *C. cajan* with chr19 of *G. max* (Varshney et al., 2012).

Cloning, Isolation, and PCR Amplification of Identified Putative R Gene Analogs (RGAs)

The genomic DNA samples from 25 known pigeonpea cultivars were scanned for presence of identified putative R genes. EPrimer (Spapé et al., 2014) was employed for designing the PCR primer sets. A list of the primer sequences used in PCR amplification are given in **Supplementary Table 7**. Eluted PCR amplicons were sequenced by the Sanger sequencing method. Isolated pigeonpea resistance gene analogs were deposited to NCBI (**Supplementary Table 6**).

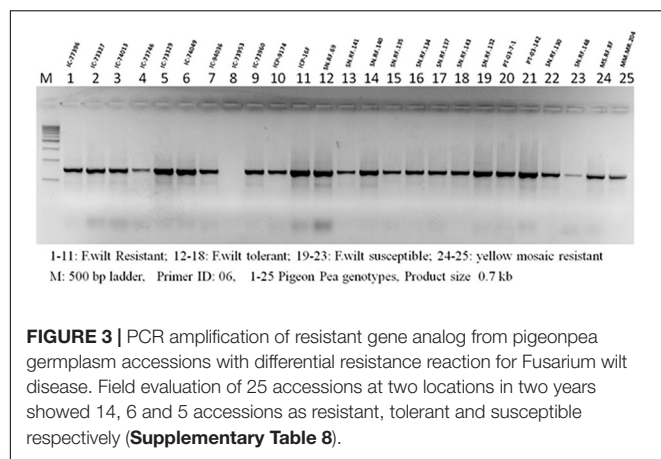
DISCUSSION

In the present work, we chose two available incomplete draft assemblies and employed a reconciliation algorithm to correct any errors. The two compared draft assemblies A1 and A2 had low genome coverage with several repeats and gaps causing disjoints between contigs. A meta-assembler was employed in the present work based on the genome reconciliation algorithm. The computational framework included a merger between the two draft assemblies, A1 and A2, aligning them by selecting common homologous sequence matches and mismatches present in both, resolving gaps, and other sequence errors, to obtain a consensus and complete assembly.

To begin with we wanted to select the order in which the input draft assemblies were to be merged to gain a subsequent superior alignment with higher read depth and read mapping.

After several permutations, we observed that treating assembly A1 as the master and aligning it with assembly A2 yielded better read mapping and lengthier scaffolds of 592,970,700 mb. Merging the two draft assemblies, in course of alignment, the meta-assembler yielded matched and mismatched portions in the merged assembly by identifying homologous genomic regions with a shared set of reads. Mismatches included gaps that had to be filled with right sized read sequences.

The meta-assembler initially utilized all available raw reads from both draft assemblies using conventional read overlapping techniques to fill the existing gaps and join the contigs. However, no notable success was observed in gap filling and repeat resolution. Alternatively, we employed local pigeonpea pair-end and mate-pair libraries to fill the gaps. The meta-assembler generated statistics comparing the distances between the mapped mates and the required sizes of insert reads to fill a gap. For example, gaps measuring <500 mb were filled by pair-end reads while mate-pair reads were utilized for filling gaps measuring 3–5 KB. Mate-pair sizes selected by the meta-assembler were further compared and validated using indexed and hashed mate libraries employing the alignment tool BIMA (Drucker et al., 2014). There are reports on the use of large sized mate-pairs for filling bigger gaps in assembly (Potato Genome Sequencing Consortium, 2011). In the present study, we employed pair-end and mate-pair reads which contributed significantly to filling gaps and thereby in joining the contigs to the full length scaffolds. Further, iterative use of pair-end and mate-pair libraries during successive alignments resulted in the identification of maximal portions shared by the same library of reads. This in turn contributed to the dramatic improvement of genome coverage in the resultant A3 assembly. The quality of the A3 assembly was judged using metrics—contig number, scaffold lengths, N50 and



L50, and genome coverage of 160× with a similarity of 72.7%. The genome similarity score can also be used in estimating the extent of redundancy present in both genomes.

Draft assembly A1 had 360,028 contigs with an N50 and L50 of 5,341 and 30,054, respectively. We obtained genome coverage of 199× with a similarity of 75.6%. Draft assembly A2 had 72,923 contigs with an N50 and L50 of 22,480 and 7,254, respectively. A2 had 592,970,700 scaffolds with a reported genome coverage of 160× with a similarity of 72.7%.

FGENESH predicted 51,737 genes using the improved meta-assembly. The predicted number of genes was less in our improved assembly (**Supplementary Table 2**) compared to A1 but was higher than A2 (**Table 2**). An annotation of the improved assembly yielded 51,737 predicted genes. Wet lab PCR amplification is the gold standard for verifying predicted gene presence and their functionality. For PCR-based gene amplification, 23 primer sets were designed to screen 34 pigeonpea cultivars. Out of the 34 genotypes screened, 14 were found to be *Fusarium* wilt resistant (**Supplementary Table 8**), 6 were *F. wilt* tolerant, 5 were *F. wilt* susceptible, and 5 had yellow mosaic susceptible genotypes (**Figure 3**). Data on yellow mosaic disease reaction are not presented here. PCR amplified genes were isolated, cloned, and submitted to NCBI (**Supplementary Table 6**). Genotype environment interaction in the field determines the phenotypic performance of isolated plant genes. The phenotypic evaluation of predicted resistance genes in field trials is also required for the transfer of obtained results to pigeonpea downstream breeding programs for the development of disease resistant cultivars. Field experiments were conducted to assess the disease reaction of the predicted R genes to *Fusarium* wilt taking cv. Asha (object of the present study) as control with 25 pigeonpea cultivars. The replicated field experiments were conducted at Ranchi (Jharkhand state) and Rahuri (Maharashtra state), India during the rainy season of 2011 and 2012. Of the 25 cultivars screened along with check cv. Asha, 14 resistant and six tolerant disease reactions at the Ranchi farm and eight resistant, one tolerant, and six susceptible disease reactions at the Rahuri farm were observed for the *F. wilt* disease of pigeonpea. The observed variation in disease incidence reflects the natural agro-climatic conditions prevailing at the individual trial sites.

CONCLUSION

In the present work, a genome reconciliation algorithm was adopted to merge and reconstruct draft assemblies to produce an accurate and near complete genome assembly of pigeonpea. We demonstrated the successful implementation of our reassembly framework by merging two chosen draft assemblies employing pair-end and mate-pair libraries to correct gaps and other sequencing errors. The resulting reconstructed meta-assembly was superior compared to the two draft assemblies in terms of measured assembly quality statistics, viz., N50 and scaffold lengths. The quality of the improved assembly was assessed for the presence of known conserved resistance gene loci (imparting resistance to *Fusarium* wilt disease in pigeonpea). An annotation of the improved assembly yielded a prediction of 1,303 resistance genes (including six extra genes gained from the meta-assembly). PCR screens and field experiments validated the resistance reaction of isolated genes against *Fusarium* wilt thus making the results available to pigeonpea breeders.

MATERIALS AND METHODS

We developed a workflow model (**Figure 4**) based on a reconciliation algorithm, that includes: 1. A merger of the two misassemblies, 2. finding matches and mismatches and other sequencing errors, 3. gap closing using pair-ends and mate-pair libraries, and 4. the assessment of improved assembly quality, and the prediction, isolation, and characterization of disease resistance gene families.

Retrieval of Pigeonpea Genome Datasets

Complete data sets belonging to two whole genome sequences of pigeonpea and the associated 23 SRA reads were downloaded from the National Center for Biotechnology Information (NCBI)² to local storage—GCA_000340665.1 (SRA accessions SRR5922904-SRR5922907) and GCA_000230855.2 (SRA accessions SRR6189003-SRR6189021) for the cv Asha.

PCR Amplification

Genomic DNA from 15-day-old seedlings of 34 pigeonpea cultivars was extracted employing the CTAB method. The purity and concentration of DNA was estimated with Nanodrops ND-1000. Nine primers were selected for the polymorphism study (**Supplementary Table 7**). Polymerase chain reaction (PCR) was performed on a total volume of 20 μl containing 60 ng of template DNA, 200 μM of dNTPs, 2.5 mM of MgCl₂, 1× PCR buffer, 0.4 μM of each primer, 0.75 U of Taq DNA polymerase, and water to make the final volume up to 20 μl. For designing the primer sets for the PCR amplification of predicted resistance gene (R) orthologs, BLASTN was employed against the soybean genome. EPrimer (Goodstein et al., 2012) was employed for designing PCR primer sets. A list of primer sequences used in PCR amplification are given in

²<https://www.ncbi.nlm.nih.gov>

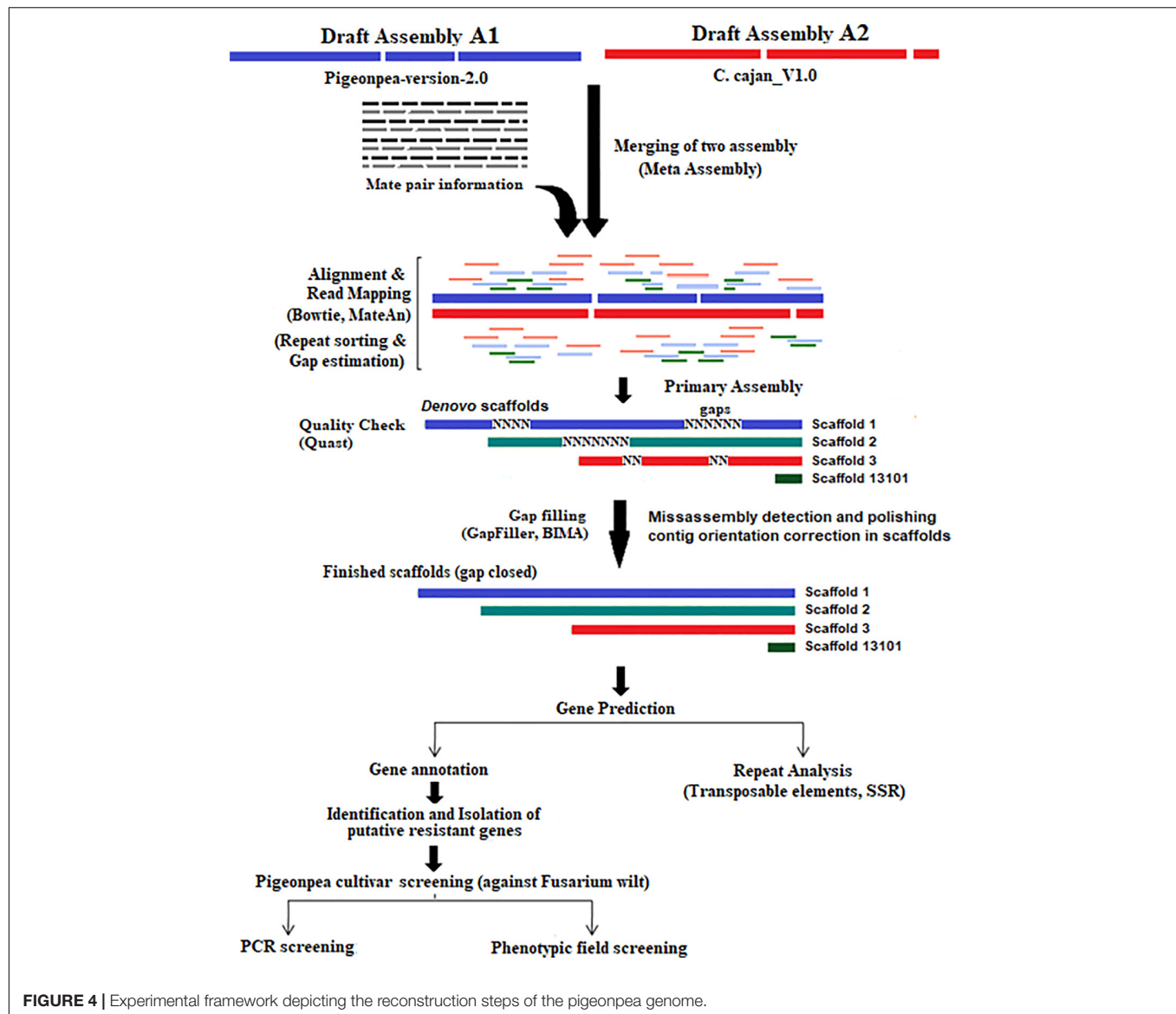


FIGURE 4 | Experimental framework depicting the reconstruction steps of the pigeonpea genome.

Supplementary Table 7. Eluted PCR amplicons were sequenced by the Sanger sequencing method.

Amplifications were carried out using the Bioer Gene Pro thermocycler and PCR conditions were set as an initial denaturation at 94°C for 5 min, 30 cycles of denaturation at 94°C for 30 s, primer annealing at 50°C for 30 s, primer extension at 72°C for 2 min, and a final extension step at 72°C for 7 min. The amplified products were visualized by ethidium bromide-stained 1.5% agarose gels in a SYNGENE G-Box gel documentation unit (Figure 3).

Genome Reconstruction and Quality Assessment

Illumina pair-end and mate-pair library sequence reads of pigeonpea and cv Asha were quality checked using

FASTQC v0.11.8³. Contaminated reads were removed to obtain error-corrected reads. Reads with sequence quality Phred scores of less than Q30 (base calling accuracy with less than 99.99%) were removed using PRINSEQ v0.20.4⁴ and reads were repaired using BBmap v37.66⁵.

Reported pigeonpea draft assemblies A1 (Singh et al., 2012) and A2 (Varshney et al., 2012) were both sequenced using Illumina technology and assembled with the SoapDenovo v2.3.1 assembler. In the present work, data sets A1 (GCA_000340665.1 consisting of 4 SRA read sets) and A2 (GCA_000230855.2 of 19 SRA read sets) were analyzed employing a reconciliation algorithm (Tarailo-Graovac and Chen, 2009). The work flow included the steps: (1) the merger of the two misassemblies,

³<http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

⁴<https://sourceforge.net/projects/prinseq>

⁵<https://sourceforge.net/projects/bbmap>

(2) finding matches, mismatches, and other structural errors, (3) closing gaps using pair-end and mate pair libraries, (4) the validation of mate-pair sizes used in the meta-assembly using indexed and hashed mate-pair library sets, (5) an assessment of the quality of the improved assembly, and (6) a prediction of disease resistance gene families, their isolation, and characterization.

A1 consisted of 360,028 initial contigs (N50 5341, 648 Mb) with 30% of gaps within contigs. A2 contained 72,923 scaffolds (N50 22480, 605 Mb) with 20% intra scaffold gaps. We used all the read datasets available belonging to A1 and A2 with NCBI. All the computations including read pre-processing, quality control, comparison of the two draft assemblies, their alignment, gap filling, assembly merger, map accuracy, quality assessment, and putative gene prediction were performed on a HPC server employing a meta-assembler (Wences and Schatz, 2015).

LG_Gapcloser and GapFiller (Boetzer and Pirovano, 2012; Xu et al., 2019) were employed to find the existing gaps (A1 30%; A2 20%). Mate-pair libraries were hashed and gap sizes were validated using the alignment tool BIMA (Drucker et al., 2014). Initially short reads were used for filling gaps, resulting in a genome size of 648 Mb in A1 and 605 Mb in the A2 draft assembly.

Draft assembly A1 was sequenced in 2011 and had a genome coverage of 199× (Singh et al., 2012). However, using the same raw read data, assembly A3 reported a gain of coverage, i.e., an increase of ~15% (from 60.0 to 75.6%), and was then resubmitted to NCBI. In our present work, we used this recent assembly set and A1 and A2 assembly data (Varshney et al., 2012) for reassembly and improvement (Figure 4).

We observed that in our reassembly, pair-end insert read sizes below 500 bp in our library were utilized for filling smaller gaps. Although mate-pair sizes up to 5.0 kb are available in our library, a 1,350 kb size was the largest used insert. In our meta-assembly, these mate pairs were employed affectively for closing medium and long-distanced gaps (even up to 20–25 kb). Similar results on the use of large sized mate-pairs for filling bigger gaps was reported in the assembly of large genomes (Ghurye and Pop, 2019).

Merging Misassemblies and Gap Closure

Draft assembly sequences A1 and A2 were merged into a single sequence. The alignment and merger of the A1 and A2 assemblies resulted in a total scaffold length of 548 Mb. The resulting merged assembly was compared to the A1 and A2 draft assemblies (75.6 and 72.7%, respectively) and had an improved genome coverage of 82.4%. Yet the merged sequence contained 10% of gaps.

To improve further contiguity and accuracy of the merged sequence, existing intra scaffold gaps were filled. Repeat content and existing gaps were estimated by Gapcloser (Xu et al., 2019) and Gapfiller (Boetzer and Pirovano, 2012). In the second round of gap filling, various computational approaches, such as paired end, mate-pair libraries, and remaining unused short reads were used. The gap content and the estimation of repeats is shown in Table 1. Iterative use of the leftover short reads (300 bp) contributed to filling nearly 20% of the gaps. After polishing and

another round of reassembly, a scaffold length of 13,348 (scaffolds of N50 574,622) with a coverage of 174× was yielded.

Improved Genome Assembly and Quality Assessment

Increased N50, maximum scaffold length and minimum number of contigs, increased N50 values together with longer scaffolds contributed to improving the genome coverage. In the misassemblies, the number of gaps and Ns caused by repeats were measured. In the course of meta-assembly, we strived to minimize gaps and other sequencing errors. We employed Quast v4.5 (Gurevich et al., 2013) to gather extensive assembly statistics. BUSCO v3.2 (Simao et al., 2015) was employed for assessing the genome completeness, and the annotation and sets of predicted genes. Mapping accuracy and the identification of resistant gene analog loci were assessed (Supplementary Table 1). In addition, 75% of unigenes were aligned to the reassembled genome.

Gene Prediction and Function Annotation

The meta-assembly was first repeat-masked using the Repeat Modler and Repeat Masker tools (Smit et al., 2013), followed by *ab initio* gene prediction using the FGENESH module of the Molquest v4.5 software package⁶. The predicted genes were annotated using the BLASTX ($E < 10^6$) search against the NCBI non-redundant (nr) protein database using the Blast2GO software (Conesa et al., 2005). Synteny blocks between the genomes of pigeonpea and other legumes were computed by blastp combined with the Circos (Krzywinski et al., 2009) to understand homology to the NBS-LRR gene from *Glycin max* (Gm) and *Medicago truncatula* (Mt) pseudomolecules.

Identification of Genome Wide SSR

The refined genome sequence of pigeonpea was analyzed to identify various simple sequence repeats (SSR) types using the Microsatellite Identification tool (MISA)⁷. The minimum length for SSR motifs per unit size was set to 10 for mono, 6 for di, and 5 for a tri, tetra, penta, and hexa motifs. We calculated the total lengths of all mono-, di-, tri-, tetra-, penta-, and hexa-nucleotide repeats in terms of base pairs of SSR per mega base pair (Mb) of DNA.

Gene Validation

The genome similarity score recorded a set of sequenced reads originating from one draft genome correctly mapped onto a second genome. To check the accuracy in the improved pigeonpea genome, we wanted to verify the location of certain genomic regions or loci present in the inputted two assemblies. A set of genes imparting resistance against various pests and diseases were located in the B4 cluster on chromosomes in the two examined draft assemblies of pigeonpea (*Cajanus cajan*) Asha. As a test case, the location of B4 gene cluster syntenic regions were verified in the present study to estimate the accuracy of read mapping achieved in the improved assembly.

⁶<http://www.softberry.com>

⁷<http://pgrc.ipkgatersleben.de/misa/>

Computational Resources

We ran all reassembly and merging operations using HPC Cluster with CentOS-Linux version 7, 2.93 GHz $2 \times$ Intel Xeon 8 core processors and 2 TB of RAM. The majority of the running time was spent on the assembly process and about 1/4 of the time was spent on graph construction and analysis. However, Reconciliator used more than 1.5 TB of RAM to merge the pigeonpea draft assemblies.

DATA AVAILABILITY STATEMENT

The improved draft genome assembly of Pigeonpea is available at NCBI GenBank, under the Accession Number GCA_015227855.1. All datasets generated for this study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

SM conceptualized and supervised all the experiments, interpreted the results, and wrote the manuscript. PM and RM performed all the high-throughput bioinformatics analysis, interpreted the results, and formulated the manuscript. MS conducted field phenotyping experiments. DW, AK, NS, and RK updated the final manuscript for publication. MY and SS performed all the wet lab experiments. All authors contributed to the article and approved the submitted version.

REFERENCES

- Alhakami, H., Mirebrahim, H., and Lonardi, S. A. (2017). Comparative evaluation of genome assembly reconciliation tools. *Genome Biol.* 18:93.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Antipov, D., Korobeynikov, A., McLean, J. S., and Pevzner, P. A. (2016). hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32, 1009–1015. doi: 10.1093/bioinformatics/btv688
- Ardui, S., Ameer, A., Vermeesch, J. R., and Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Res.* 46, 2159–2168. doi: 10.1093/nar/gky066
- Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. *Genome Biol.* 13:R56.
- Brown, C. G., and Clarke, J. (2016). Nanopore development at oxford nanopore. *Nat. Biotechnol.* 34, 810–811. doi: 10.1038/nbt.3622
- Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakkottil, G., Wright, J., et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res.* 27, 885–896. doi: 10.1101/gr.217117.116
- Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply *de Bruijn* graphs to genome assembly. *Nat. Biotechnol.* 29, 987–991. doi: 10.1038/nbt.2023
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

ACKNOWLEDGMENTS

We thank the ICAR-National Bureau of Plant Genetic Resources, New Delhi for providing research facilities, the Centre for Agricultural Bioinformatics (CABIN, ICAR-IASRI), New Delhi, India for providing a high performance computing (HPC) facility, and the National Gene Bank, ICAR-NBPGR for access to pigeonpea cultivars. We thank biorxiv for making the preprint available at biorxiv: <https://biorxiv.org/cgi/content/short/2020.08.10.243949v1>.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.607432/full#supplementary-material>

Supplementary Table 1 | BUSCO (Benchmarking Universal Single-Copy Orthologs) genes distribution.

Supplementary Table 2 | Putative disease resistance genes predicted from the improved reassembly of pigeonpea.

Supplementary Table 3 | Numbers of predominant SSRs.

Supplementary Table 4 | Resistance gene analogs sharing homology with species.

Supplementary Table 5 | NBS-LRR domain containing resistance gene analogs to pigeonpea [*Cajanus cajan* (L.) Millsp.].

Supplementary Table 6 | List of pigeonpea disease resistance genes submitted to NCBI.

Supplementary Table 7 | List of primer sequences used in PCR amplification.

Supplementary Table 8 | Pigeonpea phenotypic field scores for *Fusarium* wilt disease reaction.

- David, P., Chen, N. W., Pedrosa-Harand, A., Thareau, V., Sévignac, M., Cannon, S. B., et al. (2009). A nomadic subtelomeric disease resistance gene cluster in common bean. *Plant Physiol.* 151, 1048–1065. doi: 10.1104/pp.109.142109
- Denton, J. F., Lugo-Martinez, J., Tucker, A. E., Schrider, D. R., Warren, W. C., Hahn, M. W., et al. (2014). Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput. Biol.* 10:e1003998. doi: 10.1371/journal.pcbi.1003998
- Drucker, T. M., Johnson, S. H., Murphy, S. J., Cradic, K. W., Therneau, T. M., and Vasmataz, G. (2014). BIMA V3: an aligner customized for mate pair library sequencing. *Bioinformatics* 30, 1627–1629. doi: 10.1093/bioinformatics/btu078
- Ghurye, J., and Pop, M. (2019). Modern technologies and algorithms for scaffolding assembled genomes. *PLoS Comput. Biol.* 15:e1006994. doi: 10.1371/journal.pcbi.1006994
- Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186.
- Grau, J., Hackl, T., Koepfli, K., and Hofreiter, M. (2019). Improving draft genome contiguity with reference-derived *in silico* mate-pair library. *Gigascience* 7:giy029.
- Guizelini, D., Raittz, R. T., Cruz, L. M., Souza, E. M., Steffens, M. B., and Pedrosa, F. O. (2016). GFinisher: a new strategy to refine and finish bacterial genome assemblies. *Sci. Rep.* 6:34963.
- Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUASt: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072–1075. doi: 10.1093/bioinformatics/btt086

- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645.
- Kumar, A., Mishra, P., Maurya, R., Mishra, A. K., Gupta, V. K., Ramteke, P. W., et al. (2018). Improved draft genome sequence of a monoteliosporic culture of the Karnal bunt (*Tilletia indica*) pathogen of wheat. *Genome Announc.* 6:e00015-18.
- Lindner, M. S., Kollock, M., Zickmann, F., and Renard, B. Y. (2013). Analyzing genome coverage profiles with applications to quality control in metagenomics. *Bioinformatics* 29, 1260–1267.
- Macas, J., Novák, P., Čížková, J., Koblížková, A., Neumann, P., Fuková, I., et al. (2015). In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe *Fabeae*. *PLoS One* 10:e0143424. doi: 10.1371/journal.pone.0143424
- Marin, D., Zimmerman, R., Tao, X., Zhan, Y., Scott, R. T. Jr., Treff, N. R., et al. (2018). Validation of a targeted next generation sequencing-based comprehensive chromosome screening platform for detection of triploidy in human blastocysts. *Reprod. Biomed.* 36, 388–395.
- Mishra, P., Maurya, R., Gupta, V. K., Ramteke, P. W., Marla, S. S., and Kumar, A. (2019). Comparative genomic analysis of monosporidial and monoteliosporic cultures for unraveling the complexity of molecular pathogenesis of *Tilletia indica* pathogen of wheat. *Sci. Rep.* 9:8185.
- Pande, S., Sharma, M., and Guvvala, G. (2013). An updated review of biology, pathogenicity, epidemiology and management of wilt disease of pigeonpea (*Cajanus cajan* (L.) Millsp.). *J. Food Legum.* 26, 1–14.
- Potato Genome Sequencing Consortium (2011). Genome sequence and analysis of the tuber crop potato. *Nature* 475:189.
- Pozzi, C., and Salamini, F. (2007). “Genomics of wheat domestication,” in *Genomics-Assisted Crop Improvement*, eds R. K. Varshney and R. Tuberosa (Dordrecht: Springer), 453–481.
- Rao, Y. S., Chai, X. W., Wang, Z. F., Nie, Q. H., and Zhang, X. Q. (2013). Impact of GC content on gene expression pattern in chicken. *Genet. Select. Evol.* 45:9.
- Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212.
- Singh, N. K., Gupta, D. K., Jayaswal, P. K., Mahato, A. K., Dutta, S., Singh, S., et al. (2012). The first draft of the Pigeonpea genome sequence. *J. Plant Biochem. Biotechnol.* 21, 98–112.
- Smit, A. F. A., Hubley, R., and Green, P. (2013). *RepeatMasker Open-4.0*. Available online at: <http://www.repeatmasker.org/RepeatMasker> (accessed October 30, 2019).
- Spapé, M., Verdonchot, R., Dantzig, S. V., and Steenbergen, H. V. (2014). *The E-Primer: An Introduction to Creating Psychological Experiments in E-Prime®*. Leiden: Leiden University Press.
- Tarailo-Graovac, M., and Chen, N. (2009). Using repeat masker to identify repetitive elements in genomic sequences. *Curr. Proto. Bioinform.* 25, 1–14.
- Tsai, I. J., Otto, T. D., and Berriman, M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. *Genome Biol.* 11:R41.
- Varshney, R. K., Chen, W., Li, Y., Bharti, A. K., Saxena, R. K., Schlueter, J. A., et al. (2012). Draft genome sequence of *Pigeonpea* (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nat. Biotechnol.* 30:83.
- Wang, Y., Yu, Y., Hao, P., Li, Y., Shao, Z., Xu, X., et al. (2012). Optimizing hybrid assembly of next-generation sequence data from *Enterococcus faecium*: a microbe with highly divergent genome. *BMC Syst. Biol.* 6:S21. doi: 10.1186/1752-0509-6-S3-S21
- Wences, A. H., and Schatz, M. C. (2015). Metassembler: merging and optimizing de novo genome assemblies. *Genome Biol.* 16:207.
- Wetzel, J., Kingsford, A., and Pop, M. (2011). Assessing benefits of using mate-pairs to resolve repeats in *de novo* short read prokaryotic assemblies. *BMC Bioinform.* 12:95. doi: 10.1186/1471-2105-12-95
- Xu, G. C., Xu, T. J., Zhu, R., Zhang, Y., Li, S. Q., Wang, H. W., et al. (2019). LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* 8:giy157.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Marla, Mishra, Maurya, Singh, Wankhede, Kumar, Yadav, Subbarao, Singh and Kumar. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Identification of Molecular Markers Associated With the Pathophysiology and Treatment of Lupus Nephritis Based on Integrated Transcriptome Analysis

Menghui Yao, Congcong Gao, Chunyi Zhang, Xueqi Di, Wenfang Liang, Wenbo Sun, Qianqian Wang and Zhaohui Zheng*

Department of Rheumatology, The First Affiliated Hospital of Zhengzhou University, Zhengzhou, China

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Ming Zhao,
Central South University, China
Ramalingam Bethunaickan,
National Institute of Research
in Tuberculosis (ICMR), India

*Correspondence:

Zhaohui Zheng
fcczhengzh@zzu.edu.cn

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 15 July 2020

Accepted: 19 November 2020

Published: 15 December 2020

Citation:

Yao M, Gao C, Zhang C, Di X,
Liang W, Sun W, Wang Q and
Zheng Z (2020) Identification
of Molecular Markers Associated With
the Pathophysiology and Treatment
of Lupus Nephritis Based on
Integrated Transcriptome Analysis.
Front. Genet. 11:583629.
doi: 10.3389/fgene.2020.583629

Lupus nephritis (LN) is a well-known complication of systemic lupus erythematosus and is its leading cause of morbidity and mortality. Our study aimed to identify the molecular markers associated with the pathophysiology and treatment of LN. The renal tissue gene expression profiles of LN patients in the GSE32591 dataset were downloaded as a discovery cohort from the Gene Expression Omnibus. Differentially expressed genes (DEGs) were identified; weighted gene co-expression network analysis (WGCNA) was used to identify the co-expression modules of DEGs; and gene function enrichment analysis, molecular crosstalk analysis, and immune cell infiltration analysis were performed to explore the pathophysiological changes in glomeruli and tubulointerstitia of LN patients. The crosstalk genes were validated in another RNA-sequencing cohort. DEGs common in RNA-sequencing dataset and GSE32591 were uploaded to the Connectivity Map (CMap) database to find prospective LN-related drugs. Molecular docking was used to verify the targeting association between candidate small molecular compounds and the potential target. In all, 420 DEGs were identified; five modules and two modules associated with LN were extracted in glomeruli and tubulointerstitia, respectively. Functional enrichment analysis showed that type I interferon (IFN) response was highly active, and some biological processes such as metabolism, detoxification, and ion transport were impaired in LN. Gene transcription in glomeruli and tubulointerstitia might affect each other, and some crosstalk genes, such as *IRF7*, *HLA-DRA*, *ISG15*, *PSMB8*, and *IFITM3*, play important roles in this process. Immune cell infiltration analysis revealed that monocytes and macrophages were increased in glomeruli and tubulointerstitia, respectively. CMap analysis identified proscillaridin as a possible drug to treat LN. Molecular docking showed proscillaridin forms four hydrogen bonds with the SH2 domain of signal transducer and activator of transcription 1 (STAT1). The findings of our study may shed light on the pathophysiology of LN and provide potential therapeutic targets for LN.

Keywords: systemic lupus erythematosus, lupus nephritis, bioinformatics, differentially expressed genes, WGCNA

INTRODUCTION

Systemic lupus erythematosus (SLE) is an autoimmune disease involving multiple organs and systems, and its pathophysiology remains unclear (Mu et al., 2015). Lupus nephritis (LN) is a well-known complication of SLE; about 80% of children and 40% of adults are affected by LN (Brunner et al., 2008), which is the leading cause of morbidity and mortality in SLE patients. We performed a retrospective study for 491 LN patients in China and found that the cumulative probability of survival at 10 and 20 years are 77 and 45%, respectively (Zheng et al., 2012). At present, for the treatment of SLE and LN, most clinicians use high-dose glucocorticoids and immunosuppressants to induce remission, followed by long-term maintenance with small doses. However, only 30–50% of the patients achieve remission, and 10–20% of LN patients progress to end-stage renal disease (ESRD) (Maria and Davidson, 2020). Therefore, the treatment and prognosis of LN are generally not optimistic. It is necessary to strengthen the study of its pathophysiology further and find new treatment methods to improve the survival rates of patients with LN.

In recent years, the combination of molecular biology and information technology has led to the emergence of bioinformatics (Li et al., 2018), which has been used to reinterpret disease at the gene level and has revealed many clinical markers that may be used to diagnose disease or evaluate prognosis, especially in cancer (Zhang et al., 2018). However, there are few studies on bioinformatics in LN. Although the etiology of LN remains uncertain, it is strongly believed that the incidence of LN is associated with genomic and epigenomic mechanisms (Kwon et al., 2019). The various gene expression profiles and their regulatory mechanisms in LN remain to be illuminated.

Here, we obtained the differentially expressed genes (DEGs) of 32 LN renal tissues and 15 healthy renal tissues from the GSE32591 dataset. Functional enrichment analysis, weighted gene co-expression network analysis (WGCNA), molecular crosstalk analysis, and immune cell infiltration analysis were performed to explore the pathophysiological changes in glomeruli and tubulointerstitia of LN patients. The crosstalk genes were then validated in another cohort. Moreover, the DEGs common in an RNA-sequencing dataset and GSE32591 were uploaded to the Connectivity Map (CMap) database to find LN-related drugs. Molecular docking was used to verify the association between candidate small molecular compounds and their potential targets. The analysis of DEGs may shed light on the pathophysiology of LN and provide potential biomarkers for its treatment.

MATERIALS AND METHODS

Subjects and Samples

Six renal tissues were obtained from biopsies of three untreated patients with LN and three patients with renal cancer from the First Affiliated Hospital of Zhengzhou University. The diagnosis of patients with LN met the 1997 American Rheumatology Association SLE Classification Criteria and international renal

pathology criteria. Healthy renal tissues at least 5 cm from the tumor were taken for controls, and their unaffected status was confirmed by microscopic examination. This study was approved by the Ethical Committee of the First Affiliated Hospital of Zhengzhou University (2018-KY-22), and informed consent was obtained from the patients.

Next-Generation Sequencing

Total RNA was extracted from the renal tissues using the TRIzol LS Reagent (Invitrogen, CA, United States). After total RNA quality check, the rRNA was removed using the Ribo-Zero™ rRNA removal kit (Illumina, CA, United States), and purification and fragmentation of RNA were performed at the same time (the fragment length was between 100 and 300 bp to facilitate sequencing). First-strand cDNA was synthesized via reverse transcription, followed by second-strand cDNA synthesis. After terminal repair and purification, the cDNA library was amplified through PCR. Finally, samples were sequenced using a 2 × 150 base paired-end configuration with the Illumina HiSeq 2500 (Illumina, CA, United States).

Gene Expression Omnibus Data Preprocessing

The renal tissue gene expression profiles of GSE32591 from LN patients and healthy controls were downloaded from the Gene Expression Omnibus (GEO) database. GSE32591 is a microarray dataset generated by the Affymetrix GeneChip Human Genome HG-U133A Custom CDF (Berthier et al., 2012). It included 32 patients with SLE and LN and 15 healthy controls. Then, the annotation document of corresponding platforms was used to annotate the gene expression profiling in each dataset. Finally, the matrix with row names as sample names and column names as gene symbols was obtained for subsequent analysis.

Differentially Expressed Gene Analysis

For GSE32591, the DEGs in glomeruli and tubulointerstitia were defined by $p < 0.05$ and $\log_2|\text{fold change}| > 1.0$ using the “limma” package in R software 4.0.0. All the DEGs in glomeruli and tubulointerstitia were defined as total DEGs in GSE32591. For RNA-sequencing data, DESeq2 software was used to analyze the DEGs by comparing the case and control groups. The DEGs were defined by $p < 0.05$ and $\log_2|\text{fold change}| > 1.0$.

Weighted Gene Co-expression Network Analysis

To explore the function of the DEGs more accurately, we identified the co-expression modules in glomeruli and tubulointerstitia using WGCNA, which is an algorithm that can specially screen genes related to the clinical traits and obtain co-expression modules with high biological significance (Langfelder and Horvath, 2008). For glomeruli, to obtain a sufficient number of genes for WGCNA analysis, the genes were ranked by their $\log_2|\text{fold change}|$ value. Finally, the genes with $\log_2|\text{fold change}| > 0.589$ ($|\text{fold change}| > 1.5$) and $p < 0.05$ were selected from the final ranked gene list. For the tubulointerstitia, the genes with $\log_2|\text{fold change}| > 0.380$ ($|\text{fold change}| > 1.3$) and $p < 0.05$

were selected. The WGCNA was performed using the R package “WGCNA” (Langfelder and Horvath, 2008). First, the appropriate soft powers β was selected according to the standard of scale-free network using the algorithm “pickSoftThreshold.” Second, the adjacency coefficient a_{ij} was calculated by the formula: $a_{ij} = |S_{ij}|^\beta$. The S_{ij} was the Pearson correlation coefficient of gene i and gene j , β represents soft powers value. Third, a topological overlap matrix (TOM) and the corresponding dissimilarity (1-TOM) were calculated according to the adjacency coefficient. Then, a hierarchical clustering dendrogram built based on 1-TOM matrix was used to divide co-expressed genes into different modules. Fourth, the module eigengene (ME) that represented the expression patterns of each module was calculated and performed a Pearson correlation analysis with the clinical trait to obtain the modules that were significantly associated with LN.

In this study, the soft threshold was defined as 12 in WGCNA analysis of glomeruli and 18 in WGCNA analysis of tubulointerstitia. The other parameters were the following: minModuleSize = 20, networkType = “unsigned,” deepSplit = 2, and mergeCutHeight = 0.25.

Functional Enrichment Analysis

Gene Ontology (GO) analysis was used to describe the attributes of genes and gene products, including biological process (BP), molecular function (MF), and cellular component (CC). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis was used to obtain pathways at the gene level.

For the co-expression modules obtained from the WGCNA, we focused on the DEGs with $\log_2|\text{fold change}| > 1$ and $p < 0.05$ due to their significant changes and performed the GO and KEGG analyses on DEGs using DAVID¹. The results of the GO analysis related to BP and KEGG pathways were focused, and the p -value represented the significance of the GO terms and pathways; the smaller the p -value, the higher the significance.

Molecular Crosstalk Analysis Between Glomeruli and Tubulointerstitia

As the glomeruli and tubules are closely related anatomically, we wanted to know whether the DEGs in the glomeruli and tubulointerstitia can influence each other. First, we extracted the gene expression data of DEGs from the modules identified from WGCNA and reconstructed the matrices with row names as sample names and column names as DEG symbols. Second, to obtain the correlation among these matrices, we used the principal component analysis (PCA) in SPSS 25.0 to obtain the first principal component of each matrix. Pearson correlation analysis was used to calculate the correlation between these first principal components. The whole analysis process is similar to the WGCNA “relating modules to clinical trait” analysis. Third, to further explore the mechanism of interaction between the glomerulus and tubulointerstitia, we selected the hub genes in each first principal component of matrices based on the following standards: (a) the eigengene connectivity (kME) of genes in modules > 0.9 ; and (b) the correlation coefficient with the first

principal component in factor loading matrix > 0.8 . Then, we used the Search Tool for the Retrieval of Interacting Genes (STRING) database to construct a protein–protein interaction (PPI) network of the hub genes at the protein level. We focused on the interaction between the hub genes located in different modules. The hub genes with the highest degree in the network were defined as crosstalk genes.

Immune Cell Infiltration Analysis

The CIBERSORT algorithm is an analytical tool used to estimate the proportion of various types of immune cells in complex tissues (such as large solid tumors) (Ali et al., 2016). Panousis et al. (2019) have successfully used this algorithm to estimate the proportion of blood immune cell subsets for SLE patients. Therefore, we uploaded the gene expression data of glomeruli and tubulointerstitia to the CIBERSORT website² and obtained the landscapes of immune cells in these tissues, which encompassed T cells, B cells, monocytes, eosinophils, natural killer (NK) cells, macrophages, plasma cells, neutrophils, dendritic cells, and mast cells. Wilcoxon rank sum test was used to compare the proportion of immune cells between LN renal tissues and healthy renal tissues; $p < 0.05$ was considered significant. Pearson correlation was used to evaluate the correlation between the interferon (IFN)-induced genes and immune cells with significantly different proportions.

Validation of Crosstalk Genes

Next-generation sequencing (NGS) technology has developed rapidly in the past decade. It has great advantages for discovering unknown transcripts and comparing alternative splicing microarrays (Levy and Myers, 2016). Our team has performed deep sequencing of three cases of LN renal tissues and normal renal tissues and obtained a large number of DEGs. Therefore, we used the RNA-sequencing dataset to further validate the expression levels of crosstalk genes according to their fold change value.

Connectivity Map Analysis and Molecular Docking

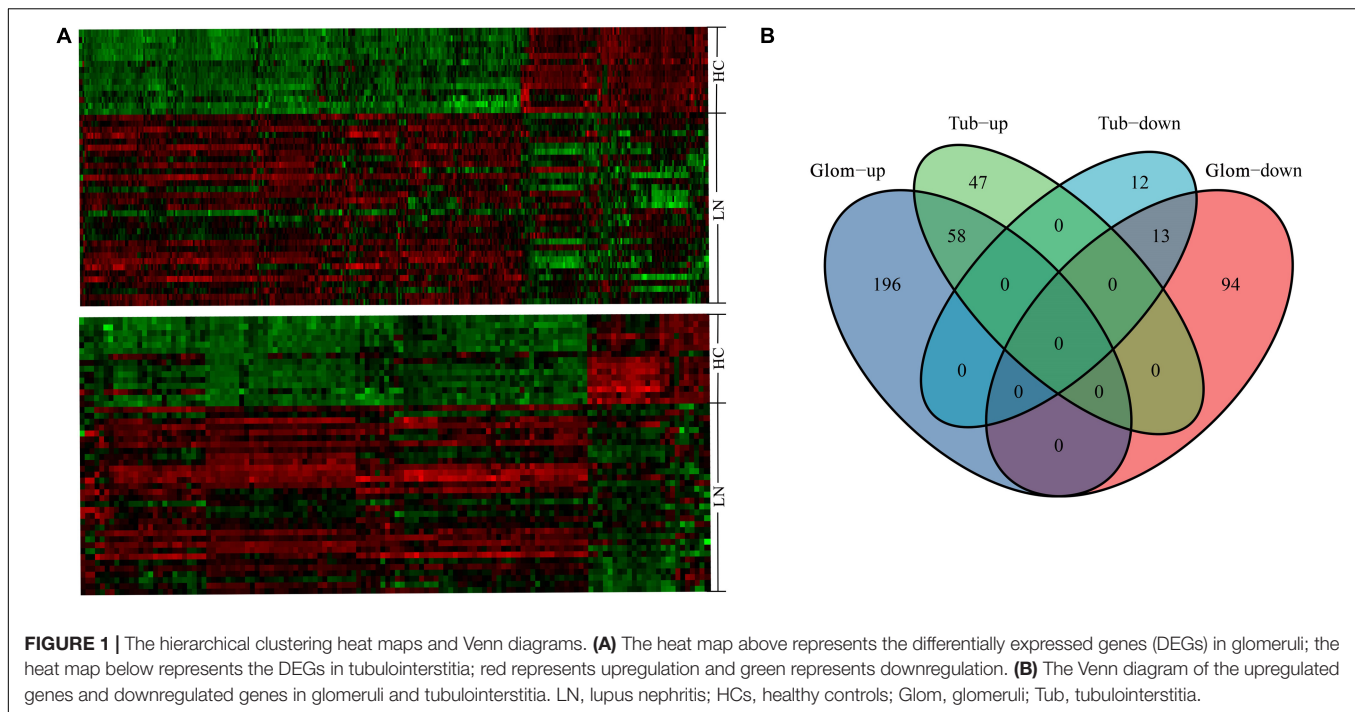
The CMap database is a database of drug-related gene expression profiles, and it consists of a large amount of genome-wide transcriptional expression data of cell lines treated with small molecular compounds to reveal the correlation among genes, diseases, and drugs (Lamb, 2007). Based on the gene expression profiles, researchers could quickly find the drugs with high relevance to diseases.

To improve the accuracy of drug screening further, we selected the common DEGs that had the same expression trend in both GSE32591 and RNA-sequencing dataset. Then, the common DEGs were converted to probe number HG133A through Affymetrix³. The probe numbers of upregulated genes and downregulated genes were transferred into the CMap website for analysis. The $p < 0.05$ and Enrichment < 0 indicated that the changes in the gene expression profiles caused by drugs were

¹<https://david.ncifcrf.gov/>

²<http://cibersort.stanford.edu/>

³<https://www.affymetrix.com>



opposite to those caused by diseases, and these drugs might have a therapeutic effect.

Molecular docking was performed using the Swissdock website to explore whether there was a targeting association between candidate small molecular compounds and DEGs (Grosdidier et al., 2011). The UCSF Chimera software 1.14 was used to visualize the binding interactions between small molecular compounds with three-dimensional (3D) models of the target.

Statistical Analysis

The data in this article were collated from two independent experiments. SPSS 25.0 and R software 4.0.0 were used for statistical analysis; $p < 0.05$ was considered statistically significant.

RESULTS

The Expression Profile of Differentially Expressed Genes in GSE32591

From the GSE32591 dataset, 361 DEGs were identified in glomeruli, including 254 upregulated genes and 107 downregulated genes. In addition, 130 DEGs were identified in tubulointerstitia, including 105 upregulated genes and 25 downregulated genes. Hierarchical clustering heat map was used to reveal the differences in the expressions of the DEGs between LN and control groups (Figure 1A). Among these DEGs, 58 genes were upregulated and 13 genes were downregulated in both glomeruli and tubulointerstitia (Figure 1B). In all, there were 420 DEGs in GSE32591, including 301 upregulated genes and 119 downregulated genes. Furthermore, the DEGs in the

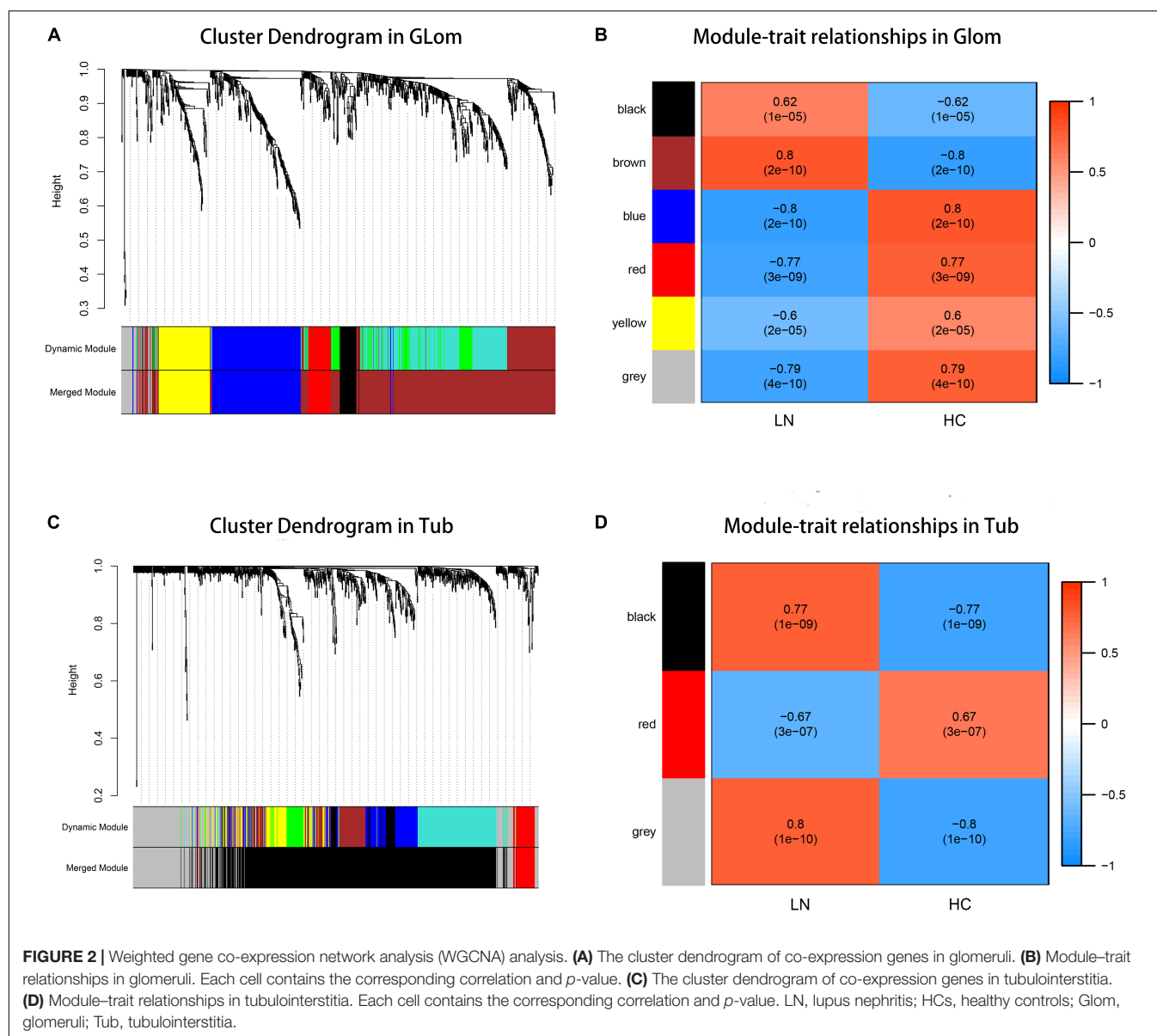
RNA-sequencing dataset were also identified. There were 1,089 DEGs in the RNA-sequencing dataset, including 565 upregulated genes and 524 downregulated genes (Supplementary Figure 1).

The Co-expression Modules in Glomeruli and Tubulointerstitia

According to the previously set criteria, there were 998 genes and 955 genes in the glomeruli and tubulointerstitia, respectively, into the WGCNA analysis. With each module assigned a color, a total of five modules were identified in glomeruli (excluding a gray module that was not assigned into any cluster). Then, a heat map was generated regarding module–trait relationships to evaluate the association between each module and two clinical features (LN and control). As shown in Figure 2, The two modules “brown” and “black” were positively associated with LN, and three modules “red,” “yellow,” and “blue” were negatively associated with LN (Figure 2B). Similarly, two modules in tubulointerstitia were identified; the module “brown” was positively associated with LN, and the module “red” was negatively associated with LN (Figure 2D).

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Pathway Enrichment Analyses

In glomeruli, the DEGs in the brown module and the black module positively correlated with LN were significantly enriched in immune response, especially against virus infection mediated by type I IFN, such as “response to virus,” “defense response to virus,” and “type I interferon signaling pathway.” The KEGG pathway analysis revealed that the abnormal signaling pathways induced during some infectious diseases,



such as those caused by influenza A, herpes simplex, and *Staphylococcus aureus*, were similar to the pathways deployed during the development of LN. The red module was negatively related to LN, and the enrichment analysis showed some biochemical reactions and metabolic pathways are impaired in LN, such as cellular oxidant detoxification, sodium-independent organic anion transport, biosynthesis of amino acids, and protein digestion and absorption. Furthermore, the enrichment analysis for the blue module negatively related to LN also showed the regulation of muscle contraction, response to toxic substances, and Rap1 signaling pathway were also abnormal (Table 1).

In tubulointerstitia, the black module positively related to LN was enriched in the type I IFN pathway, as in the glomerulus. In the red module negatively related to LN, the enrichment analysis showed the DEGs were mainly enriched in cellular response to

hormone stimulus, response to cAMP, transcriptional action, and osteoclast differentiation (Table 2).

Gene Transcription in Glomeruli and Tubulointerstitia Was Affected by Each Other

As shown in Figure 3A, there was a high correlation between the various modules. The black module in the tubulointerstitia had different effects on almost every module in the glomeruli. Positive correlation in the glomeruli was found with the brown module and the black module, but negative correlation with the blue, yellow, and red modules. Similarly, the brown module in glomeruli is positively correlated with the black module but negatively correlated with the red module in the tubulointerstitia. The strong correlation between these modules suggested that

TABLE 1 | GO and KEGG enrichment analysis of DEGs in co-expression modules of glomeruli.

Modules	The number of DEGs	GO and KEGG terms*	p-value
Brown module	256	GO:0009615: Response to virus	6.99E-25
		GO:0051607: Defense response to virus	1.01E-23
		GO:0060337: Type I interferon signaling pathway	4.68E-19
		GO:0045071: Negative regulation of viral genome replication	1.31E-16
		GO:0045087: Innate immune response	2.14E-15
		hsa05150: <i>Staphylococcus aureus</i> infection	2.92E-09
		hsa05164: Influenza A	1.12E-07
		hsa05152: Tuberculosis	4.49E-06
		hsa05133: Pertussis	8.99E-06
Black module	8	hsa05168: Herpes simplex infection	3.19E-05
		GO:0060337: Type I interferon signaling pathway	0.023
		hsa04622: RIG-I-like receptor signaling pathway	0.040
Red module [#]	3	–	–
Yellow module	45	GO:0098869: Cellular oxidant detoxification	6.60E-04
		GO:0043252: Sodium-independent organic anion transport	0.001
		GO:0055114: Oxidation-reduction process	0.003
		GO:0042157: Lipoprotein metabolic process	0.004
		GO:0006094: Gluconeogenesis	0.005
		hsa01100: Metabolic pathways	0.001
		hsa01130: Biosynthesis of antibiotics	0.002
		hsa01230: Biosynthesis of amino acids	0.004
		hsa04974: Protein digestion and absorption	0.007
		hsa00260: Glycine, serine, and threonine metabolism	0.013
Blue module	33	GO:0006937: Regulation of muscle contraction	2.24E-04
		GO:0032972: Regulation of muscle filament sliding speed	0.003
		GO:0009636: Response to toxic substance	0.008
		GO:0055010: Ventricular cardiac muscle tissue morphogenesis	0.040
		GO:0055010: Negative regulation of insulin receptor signaling pathway	0.046
		hsa04015: Rap1 signaling pathway	0.025

*If GO or KEGG terms were more than five, only the top five terms were displayed.

[#]The red module was not enriched in significant terms. DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes.

the genes transcribed in glomeruli and tubulointerstitia may interact with each other. The PPI network between these modules suggested some hub genes acted as bridges between these modules (**Figure 3B**). We calculated the degree of each hub gene using the “Network analysis” tool in Cytoscape 3.7.2. The top 10 genes with the highest degrees were obtained, including *IRF7*, *HLA-DRA*, *ISG15*, *PSMB8*, *IFITM3*, *GBP2*, *OAS2*,

TABLE 2 | GO and KEGG enrichment analysis of DEGs in co-expression modules of tubulointerstitia.

Modules	The number of DEGs	GO and KEGG terms*	p-value
Black module	106	GO:0060337: Type I interferon signaling pathway	2.92E-34
		GO:0009615: Response to virus	1.24E-21
		GO:0051607: Defense response to virus	2.99E-21
		GO:0006955: Immune response	8.47E-19
		GO:0045071: Negative regulation of viral genome replication	3.31E-18
Red module	16	hsa05168: Herpes simplex infection	5.31E-15
		hsa05332: Graft vs. host disease	1.61E-14
		hsa05330: Allograft rejection	5.92E-14
		hsa05150: <i>Staphylococcus aureus</i> infection	8.92E-14
		hsa04940: Type I diabetes mellitus	2.43E-13
		GO:0032870: Cellular response to hormone stimulus	3.23E-10
		GO:0051591: Response to cAMP	6.58E-08
		GO:0006366: Transcription from RNA polymerase II promoter	3.12E-06
		GO:0045944: Positive regulation of transcription from RNA polymerase II promoter	9.67E-06
		GO:0035914: Skeletal muscle cell differentiation	1.04E-05
		hsa05166: HTLV-I infection	3.19E-04
		hsa04380: Osteoclast differentiation	7.34E-04
		hsa05031: Amphetamine addiction	0.004
		hsa04010: MAPK signaling pathway	0.005

*If GO or KEGG terms were more than five, only the top 5 terms were displayed. DEGs, differentially expressed genes; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MAPK, mitogen-activated protein kinase.

SLC27A2, *SLC15A3*, and *IFI44*; hence, these genes were defined as crosstalk genes.

Performance of Immune Cell Infiltration Analysis

As mentioned above, the type I IFN response was very significant in LN. Considering that some immune cells play salient roles in the type I IFN response, we used the CIBERSORT algorithm to estimate the proportion of various types of immune cells in the kidney and explore their relationship with IFN-induced genes. The results showed that the number of monocytes increased significantly in the glomeruli of the LN group compared with that in the control. Moreover, the number of activated NK cells was also increased. On the contrary, the number of memory B cells, T follicular helper cells (Tfh cells), T regulatory cells (Tregs), resting NK cells, resting dendritic cells, and resting memory CD4 T cells was decreased (**Figure 4A**). In the tubulointerstitia, the number of M1 and M2 macrophages, gamma delta T cells, and resting mast cells was increased, whereas that of CD8 T cells, Tfh cells, and resting dendritic cells was decreased (**Figure 4B**).

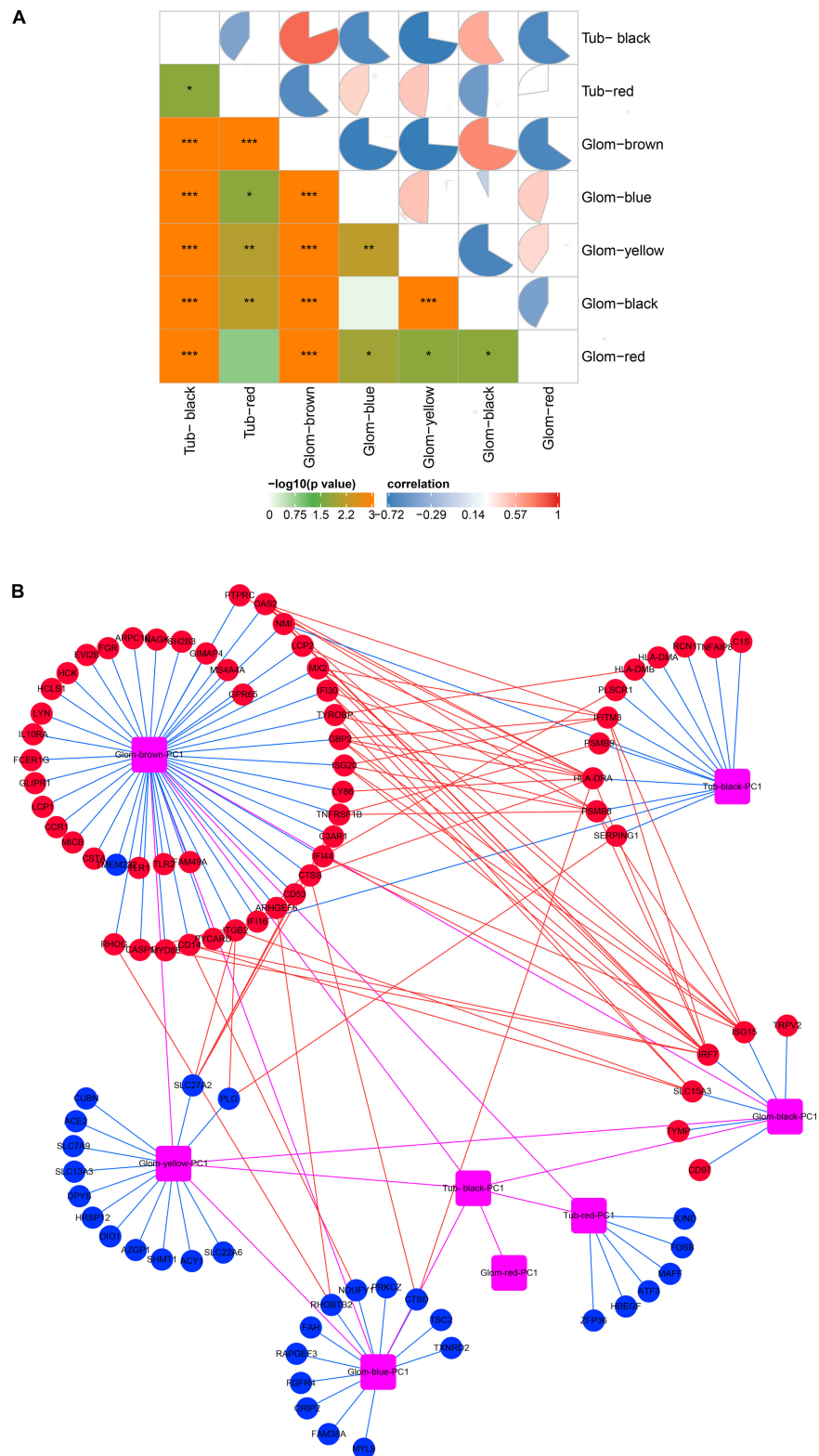
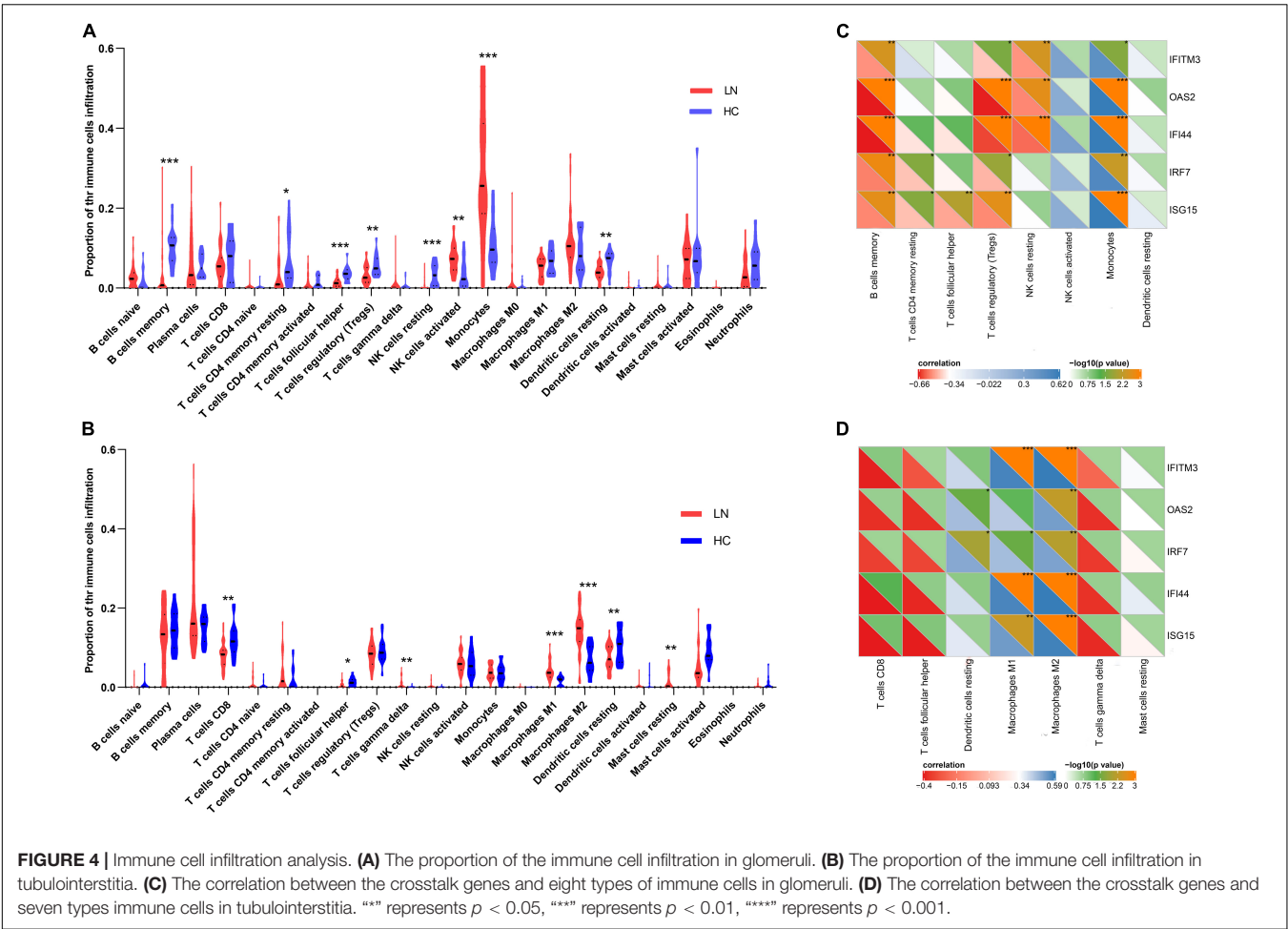


FIGURE 3 | Molecular crosstalk analysis. **(A)** The correlation between modules in glomeruli and tubulointerstitia. **(B)** The interaction of hub genes located in various modules. Glom, glomeruli; Tub, tubulointerstitia; PC1, first principal component. Blue lines represent inclusion relationship of modules to hub genes; red lines represent the interaction between the hub genes located in different modules; magenta lines represent the interaction between various modules; red represents upregulation; blue represents downregulation. *** represents $p < 0.05$, **** represents $p < 0.01$, ***** represents $p < 0.001$.



The Pearson correlation analysis showed that the IFN-induced genes, *IRF7*, *ISG15*, *IFITM3*, *OAS2*, and *IFI44*, in the crosstalk gene set were associated with immune infiltration. In glomeruli, these hub genes were positively correlated with monocytes but negatively correlated with memory B cells and Tregs (Figure 4C). In the tubulointerstitia, the IFN-induced genes were positively correlated with M1 and M2 macrophages (Figure 4D).

Validation of Crosstalk Genes by Next-Generation Sequencing

To verify our analysis, we extracted the expression level of these crosstalk genes using NGS and found that most crosstalk genes had the same changes in the RNA-sequencing dataset (Table 3), illustrating a satisfactory reliability of the result. The expression levels of *HLA-DRA*, *GBP2*, and *SLC27A2* did not differ in our sequencing ($p > 0.05$), but they showed the same trends as microarray sequencing. In the future, we will expand the sample size to validate these crosstalk genes.

Candidate Lupus Nephritis-Related Small Molecular Compounds

To identify LN-related small molecular compounds accurately, we integrated the DEGs between GSE32591 and RNA-sequencing

dataset and obtained 50 common DEGs, including 38 upregulated genes and 12 downregulated genes (Table 4). Most of the common DEGs were IFN-induced genes, and their biological processes are mainly related to type I IFN signaling

TABLE 3 | The FC value of crosstalk genes in GSE32591 and RNA-sequencing dataset.

	Glom		Tub		Kidney	
	FC	p-value	FC	p-value	FC	p-value
<i>IRF7</i>	3.138	< 0.001	1.579	< 0.001	1.989	0.004
<i>HLA-DRA</i>	1.876	< 0.001	2.152	< 0.001	1.015	0.915
<i>ISG15</i>	6.561	< 0.001	9.980	< 0.001	5.732	< 0.001
<i>PSMB8</i>	1.543	0.004	2.836	< 0.001	1.780	< 0.001
<i>IFITM3</i>	2.530	< 0.001	3.278	< 0.001	1.707	0.004
<i>GBP2</i>	3.706	< 0.001	1.509	0.007	1.042	0.878
<i>OAS2</i>	5.979	< 0.001	1.911	< 0.001	3.621	< 0.001
<i>SLC27A2</i>	0.401	< 0.001	0.988	0.895	0.655	0.607
<i>SLC15A3</i>	2.359	< 0.001	1.173	< 0.001	2.228	< 0.001
<i>IFI44</i>	9.088	< 0.001	7.989	< 0.001	3.997	< 0.001

FC, fold change; Glom, glomeruli; Tub, tubulointerstitia.

TABLE 4 | The common DEGs in GSE32591 dataset and RNA-sequencing dataset.

Expression	Genes
Up	<i>STAT1, IFI44L, MX1, IFI44, RSAD2, IFI6, MX2, HERC6, ISG15, OAS2, OAS3, OAS1, HERC5, XAF1, IFI27, IFIT1, IFITM1, IFIT3, PARP12, SAMS1, RTP4, HLA-DQA1, NNMT, PTGER2, LTF, SRGN, PSMB9, TFPI2, SLC15A3, UCP2, ARPC1B, DDX60, LY6E, BST2, MMP7, CFB, UBE2L6, CLU</i>
Down	<i>ATF3, EGR1, ZFPM2, FOS, EGR3, CHI3L1, MYL9, TNNC1, FOSB, JUNB, JUN, ZFP36</i>

DEGs, differentially expressed genes.

TABLE 5 | Ten small molecular compounds for lupus nephritis obtained from the Connectivity Map (CMap) database.

Rank	CMap name	n	Enrichment	p
1	Geldanamycin	15	0.661	0
2	Tanespimycin	62	0.583	0
3	Proscillaridin	3	-0.983	0.00002
4	H-7	4	0.922	0.00004
5	Lisuride	5	-0.807	0.00062
6	5155877	4	-0.860	0.00068
7	Meclocycline	4	-0.859	0.00068
8	Doxorubicin	3	0.921	0.00106
9	Lycorine	5	-0.775	0.00106
10	Lomustine	4	-0.819	0.00203

pathway (Supplementary Figure 2). Then, we queried the CMap database using the upregulated and downregulated genes and identified some compounds that might influence LN; the 10 compounds are shown in Table 5. Doxorubicin and H-7 were the first two small-molecule drugs with the highest enrichment score, and proscillaridin was the small molecular drug with the lowest enrichment scores; their 3D chemical structures were also downloaded from Pubchem database (Figures 5A–C).

Targeting Association Between Signal Transducer and Activator of Transcription 1 and Proscillaridin via Molecular Docking

Proscillaridin was reported to inhibit signal transducer and activator of transcription (STAT)3, and the protein STAT1 encoded by the upregulated DEG *STAT1* has been shown to have a structure similar to that of STAT3. We speculated that proscillaridin could also inhibit STAT1. Molecular docking was performed to preliminarily verify whether there is direct targeting between compounds and the protein. The results showed that the ARG586, HSD675, and ALA676 residues form hydrogen bonds with proscillaridin, which indicated that proscillaridin mainly interacts with the SH2 domain of STAT1 (Figure 6).

DISCUSSION

In recent years, with the wide use of immunosuppressants and biological agents, the prognosis and survival rate of patients with LN have improved; however, 10–20% of the patients with LN progress to ESRD, which is linked to a heavy burden and morbidity (Aljaberi et al., 2019). So, there is a need to study the pathophysiology and discover new therapeutic methods to prevent LN progression and prolong patient survival. Therefore, we performed sequencing in LN renal tissues and healthy renal tissues to identify DEGs and explore their roles in LN.

Through GO and KEGG pathway enrichment analyses of DEGs, we found that innate and adaptive immune response, especially against virus infection mediated by type I IFN, was highly active in both glomeruli and tubulointerstitia, such as the brown module and the black module in glomeruli and the black module in tubulointerstitia. Besides, the results also showed that the metabolism process of carbohydrate, protein, and lipid in LN patients was disordered, and some biochemical reactions involving detoxification were impaired. Interestingly, we found the blue module in glomeruli was enriched in the regulation of muscle contraction, which indicated that the contraction of mesangial cells (Jankowski et al., 2003), podocytes (Saleem et al., 2008), and capillaries might be dysregulated. This may lead to a decrease of the glomerular filtration rate (GFR) and might be one of the causes of urine protein in LN patients (Stockand and Sansom, 1998). In the tubulointerstitia, the red module was enriched in response to hormone stimulus and cAMP. Many types of ion transport are mediated via cAMP, such as Na⁺, K⁺, Ca²⁺, and Cl⁻ (Li et al., 2008). The dysregulation might affect the tubules, then the filtration and reabsorption of tubules would be impaired in LN patients.

Glomerular lesions and tubulointerstitial lesions often occurred together in LN (Cimbaluk and Naumann, 2017), so we wanted to explore whether the two lesions were related at the genetic level. Therefore, we further used PCA and correlation analysis to explore the interaction between glomeruli and tubulointerstitial modules. There was a high correlation between the various modules that suggested that the gene transcription in glomeruli and tubulointerstitia may interact with each other. Combined with gene enrichment results, clearly, the high IFN response in glomeruli and tubulointerstitia revealed a mutual promotion. For a long time, we focused on the fact that IFN could result in autoimmune inflammation in LN (Eloranta et al., 2013); however, our molecular crosstalk analysis showed that the IFN response might also affect some biological processes, such as metabolic pathways, muscle contraction, and detoxification process in glomeruli. In the tubulointerstitia, the cellular response to hormone stimulus and cAMP and transcriptional activation were highly negatively correlated with IFN response, which indicated that the IFN response might have adverse effects on these biological processes in the tubulointerstitia. Most crosstalk genes interpreted from the PPI analysis were IFN-induced genes, which also indicated that IFN-induced genes played an important role in the transcription of each module. Except the IFN-induced genes, we also found some new genes, such as *SLC27A2*, *SLC15A3*,

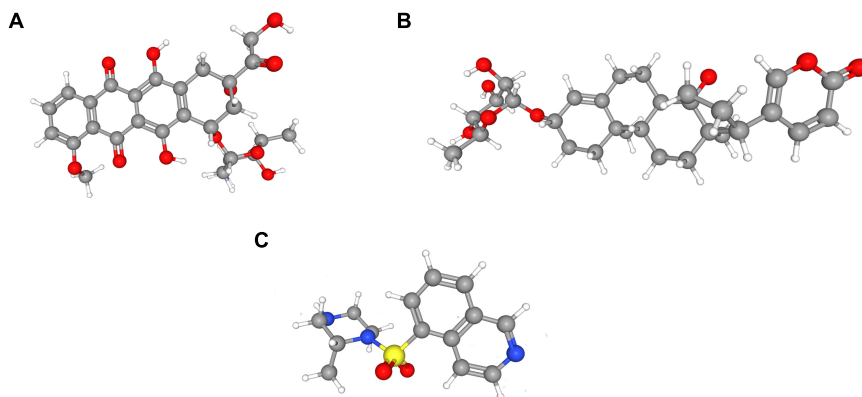


FIGURE 5 | Three-dimensional (3D) chemical structures of the three molecules. **(A)** Doxorubicin. **(B)** Proscillaridin. **(C)** H-7.

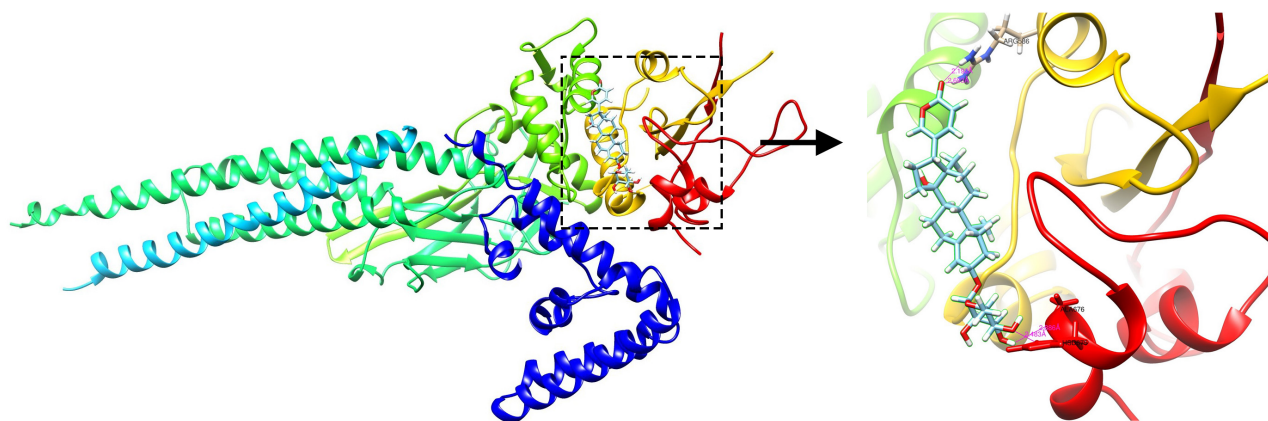


FIGURE 6 | The docking simulation result showing hydrogen bonding between proscillaridin and the ARG586, HSD675, and ALA676 residues in the SH2 domain of signal transducer and activator of transcription 1 (STAT1).

HLA-DRA, and *PSMB8*, which also might be important in kidney gene transcription.

To further explore the relationship between type I IFN response and immune cells in kidney, immune cell infiltration analysis was performed, and the results showed monocytes were the prominent differentially expressed cells in glomeruli and were positively correlated with IFN-induced genes. Monocytes are important subsets of immune cells, participate in various types of immune responses, thereby playing an important role in autoimmune diseases (Auffray et al., 2009). Uccellini and García-Sastre (2018) observed high IFN response in inflammatory monocytes during infection. Monocytes also have been reported to produce IFN and mediate tissue damage in H1N1 IAV-infected mouse models (Lin et al., 2014). Therefore, we speculated that there might be a mutual promotion between the monocytes and the high IFN response in glomeruli. However, we found that these IFN-induced genes seemed to be negatively correlated with Tregs and memory B cells. The function of Tregs is that they suppress autoreactive lymphocytes, especially CD8⁺ T cell and B cell activation, and maintain self-tolerance (Ohl and Tenbrock, 2015). It has been reported that the defects in Tregs or a lack of Tregs

is associated with SLE pathogenesis (Ohl and Tenbrock, 2015). So we speculated that the decrease of Tregs in LN leads to the weakening of the inhibitory effect on B cells, thereby enhancing the B cell intrinsic effect for the augmentation of IFN. Besides, the reduction of memory B cells caused by the disturbance of B cell homeostasis has been observed in active SLE (Odendahl et al., 2000). We speculated that the decreased memory B cells might be related to the abnormal activation of B cells. The activated B cells circulate in the peripheral blood and participate in the formation of autoantibodies and IFN response (Eloranta et al., 2013). Therefore, there is a negative correlation between memory B cells and IFN-induced genes. Macrophages were found to be mainly elevated in the tubulointerstitia and positively correlated with IFN-induced genes. There are two major polarization states for macrophages; “M1” macrophages produce a lot of pro-inflammatory cytokines including IFN- α to cause tissue damage. On the contrary, “M2”-type macrophages can repair tissue damage by secreting anti-inflammatory cytokines such as IL-10 and CCL18 (Wen et al., 2019). The increased numbers of M1 and M2 macrophages will cause repeated injury and repair of the tubulointerstitia, leading to the fibrosis of tubulointerstitia.

Through CMap analysis, 10 drugs (geldanamycin, tanespimycin, proscillaridin, H-7, lisuride, 5155877, meclocycline, doxorubicin, lycorine, lomustine) were identified that might induce the development of LN (enrichment score > 0) or which may be potential drugs for the treatment of LN (enrichment factor < 0). Doxorubicin and H-7 were the first two small-molecule drugs with the highest enrichment scores, which indicated that the use of these small molecules or their analogs might induce or aggravate LN. Huang et al. (2004) reported a patient with SLE developing lupus-like symptoms, such as fever, erythema, and exfoliative dermatitis, with a positive lupus band test after using doxorubicin. Yang et al. (2009) found that doxorubicin treatment in mice significantly increased albuminuria and decreased podocytes. These results showed that patients with LN should be cautious when using doxorubicin. H-7 is a protein kinase inhibitor (Steele and Brahmi, 1988) and has not been reported to be associated with LN. Proscillaridin was the first small molecular drugs with the lowest enrichment score, indicating that it might be a potential therapeutic strategy for LN. In short, the abovementioned drugs might affect LN through a variety of small molecular pathways.

The DEG *STAT1* was upregulated in the common DEGs (Table 2). *STAT1* is known to occupy a central position in the type I IFN signaling pathway. If drugs that can inhibit *STAT1* and change the high IFN-response signature are identified, they may be considered as potential candidate drugs for LN treatment. Proscillaridin belongs to cardiac glycosides (Maryam et al., 2018), and Ye et al. (2011) have reported that cardiac glycosides could potentially inhibit the induction of the IFN genes induced by virus, double-stranded RNA, and double-stranded DNA, which was consistent with our analysis. Proscillaridin was also reported to have an inhibitory effect on *STAT3* (Maryam et al., 2018). As *STAT1* and *STAT3* belong to the *STAT* protein family and have similar structures, and proscillaridin reverses the high IFN-response signature, we speculated that it could also inhibit *STAT1*. Through molecular docking, we found that proscillaridin formed four hydrogen bonds with the SH2 domain of *STAT1*. The SH2 domain is the most critical and conserved domain in *STAT1*, located between amino acid residues 577 and 683; it is vital for the activation and function of *STAT1* (Levy and Darnell, 2002). Proscillaridin might inhibit the activation of *STAT1* and the type I IFN signaling pathway by binding to the SH2 domain. However, more details of the specific interactions between proscillaridin and *STAT1* need to be confirmed by future experiments.

However, there remain several limitations that need to be resolved in the future. For example, our research was a bioinformatic analysis based on sequencing data; therefore, further verifications by cell and animal experiments are needed. Besides, whether the small molecular compounds screened in our study could influence LN and the specific interactions and mechanisms between proscillaridin and *STAT1* need further confirmation. Next, better-designed experiments need to be carried out based on our findings.

In conclusion, we found that type I IFN response was highly active, and some biological processes such as metabolism, detoxification, ion transport were impaired in LN through the WGCNA analysis of DEGs. The gene transcription in glomeruli

and tubulointerstitia might affect each other, and some crosstalk genes, such as *IRF7*, *HLA-DRA*, *ISG15*, *SLC15A3*, and *IFITM3*, play important roles in this process. Monocytes and macrophages may be associated with high IFN response in kidney tissues. Proscillaridin may play a therapeutic role by targeting *STAT1*. Therefore, the analysis for DEGs provided a new perspective for the pathophysiology and treatment of LN.

DATA AVAILABILITY STATEMENT

The original sequencing data has been deposited in SRA. To review GEO accession GSE157293: Go to <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157293>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of the First Affiliated Hospital of Zhengzhou University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MY and CG participated in kidney sample collection, data analysis, and manuscript writing. ZZ designed and conducted the whole experiment and finalized the manuscript. CZ and XD participated in data analysis. WS, WL, and QW collected kidney tissues from patients. All authors discussed the outline and commented on the manuscript.

FUNDING

This work was support by the Research Project of Medical Science and Technology of Henan Province of China (Grant No. LHGJ20190260).

ACKNOWLEDGMENTS

We thank the technical assistance and expertise of Genesky Biotechnologies Inc. (Shanghai, China) for its perfect technical assistance. We would like to thank Editage (www.editage.cn) for English language editing. And we thank the Sangerbox tools for providing the platform for data analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.583629/full#supplementary-material>

Supplementary Figure 1 | The volcano maps of DEGs in the RNA-sequencing dataset.

Supplementary Figure 2 | The GO enrichment analysis of the common DEGs. The number represents the gene counts in the GO terms.

REFERENCES

- Ali, H. R., Chlon, L., Pharoah, P. D., Markowitz, F., and Caldas, C. (2016). Patterns of immune infiltration in breast cancer and their clinical implications: a gene-expression-based retrospective study. *PLoS Med.* 13:e1002194. doi: 10.1371/journal.pmed.1002194
- Aljaberi, N., Bennett, M., Brunner, H. I., and Devarajan, P. (2019). Proteomic profiling of urine: implications for lupus nephritis. *Expert Rev. Proteomics* 16, 303–313. doi: 10.1080/14789450.2019.1592681
- Auffray, C., Sieweke, M. H., and Geissmann, F. (2009). Blood monocytes: development, heterogeneity, and relationship with dendritic cells. *Annu. Rev. Immunol.* 27, 669–692. doi: 10.1146/annurev.immunol.021908.132557
- Berthier, C. C., Bethunaikkan, R., Gonzalez-Rivera, T., Nair, V., Ramanujam, M., Zhang, W., et al. (2012). Cross-species transcriptional network analysis defines shared inflammatory responses in murine and human lupus nephritis. *J. Immunol.* 189, 988–1001. doi: 10.4049/jimmunol.1103031
- Brunner, H. I., Gladman, D. D., Ibanez, D., Urowitz, M. D., and Silverman, E. D. (2008). Difference in disease features between childhood-onset and adult-onset systemic lupus erythematosus. *Arthritis Rheum.* 58, 556–562. doi: 10.1002/art.23204
- Cimbaluk, D., and Naumann, A. (2017). Renal involvement in systemic lupus erythematosus: glomerular pathology, classification, and future directions. *Diagnostic Histopathol.* 23, 109–116. doi: 10.1016/j.mpdhp.2017.03.007
- Eloranta, M. L., Alm, G. V., and Rönnblom, L. (2013). Disease mechanisms in rheumatology—tools and pathways: plasmacytoid dendritic cells and their role in autoimmune rheumatic diseases. *Arthritis Rheum* 65, 853–863. doi: 10.1002/art.37821
- Grosdidier, A., Zoete, V., and Michielin, O. (2011). SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res.* 39, W270–W277. doi: 10.1093/nar/gkr366
- Huang, J. Y., Wu, C. H., Shih, I. H., and Lai, P. C. (2004). Complications mimicking lupus flare-up in a uremic patient undergoing pegylated liposomal doxorubicin therapy for cervical cancer. *Anticancer Drugs* 15, 239–241. doi: 10.1097/00001813-200403000-00008
- Jankowski, M., Szczepańska-Konkel, K., Kalinowski, L., and Angielski, S. (2003). Involvement of Rho-kinase in P2Y-receptor-mediated contraction of renal glomeruli. *Biochem. Biophys. Res. Commun.* 302, 855–859. doi: 10.1016/s0006-291x(03)00272-9
- Kwon, Y.-C., Chun, S., Kim, K., and Mak, A. (2019). Update on the genetics of systemic lupus erythematosus: genome-wide association studies and beyond. *Cells* 8:1180. doi: 10.3390/cells8101180
- Lamb, J. (2007). The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer* 7, 54–60. doi: 10.1038/nrc2044
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559
- Levy, D. E., and Darnell, J. E. Jr. (2002). Stats: transcriptional control and biological impact. *Nat. Rev. Mol. Cell Biol.* 3, 651–662. doi: 10.1038/nrm909
- Levy, S. E., and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annu. Rev. Genomics Hum. Genet.* 17, 95–115. doi: 10.1146/annurev-genom-083115-022413
- Li, S., Yang, R., Sun, X., Miao, S., Lu, T., Wang, Y., et al. (2018). Identification of SPP1 as a promising biomarker to predict clinical outcome of lung adenocarcinoma individuals. *Gene* 679, 398–404. doi: 10.1016/j.gene.2018.09.030
- Li, Y., Konings, I. B., Zhao, J., Price, L. S., de Heer, E., and Deen, P. M. (2008). Renal expression of exchange protein directly activated by cAMP (Epac) 1 and 2. *Am. J. Physiol. Renal. Physiol.* 295, F525–F533. doi: 10.1152/ajprenal.00448.2007
- Lin, S.-J., Lo, M., Kuo, R.-L., Shih, S.-R., Ojcius, D. M., Lu, J., et al. (2014). The pathological effects of CCR2+ inflammatory monocytes are amplified by an IFNAR1-triggered chemokine feedback loop in highly pathogenic influenza infection. *J. Biomed. Sci.* 21, 99–99. doi: 10.1186/s12929-014-0099-6
- Maria, N. I., and Davidson, A. (2020). Protecting the kidney in systemic lupus erythematosus: from diagnosis to therapy. *Nat. Rev. Rheumatol.* 16, 255–267. doi: 10.1038/s41584-020-0401-9
- Maryam, A., Mehmood, T., Yan, Q., Li, Y., Khan, M., and Ma, T. (2018). Proscillaridin A promotes oxidative stress and ER stress, inhibits STAT3 activation, and induces apoptosis in A549 lung adenocarcinoma cells. *Oxid. Med. Cell Longev.* 2018:3853409. doi: 10.1155/2018/3853409
- Mu, Q., Zhang, H., and Luo, X. M. (2015). SLE: Another autoimmune disorder influenced by microbes and diet? *Front. Immunol.* 6:608. doi: 10.3389/fimmu.2015.00608
- Odendahl, M., Jacobi, A., Hansen, A., Feist, E., Hiepe, F., Burmester, G. R., et al. (2000). Disturbed peripheral B lymphocyte homeostasis in systemic lupus erythematosus. *J. Immunol.* 165, 5970–5979. doi: 10.4049/jimmunol.165.10.5970
- Ohl, K., and Tenbrock, K. (2015). Regulatory T cells in systemic lupus erythematosus. *Eur. J. Immunol.* 45, 344–355. doi: 10.1002/eji.201344280
- Panousis, N. I., Bertsias, G. K., Ongen, H., Gergianaki, I., Tektonidou, M. G., Trachana, M., et al. (2019). Combined genetic and transcriptome analysis of patients with SLE: distinct, targetable signatures for susceptibility and severity. *Ann. Rheum. Dis.* 78, 1079–1089. doi: 10.1136/annrheumdis-2018-214379
- Saleem, M. A., Zavadil, J., Bailly, M., McGee, K., Witherden, I. R., Pavenstadt, H., et al. (2008). The molecular and functional phenotype of glomerular podocytes reveals key features of contractile smooth muscle cells. *Am. J. Physiol. Renal. Physiol.* 295, F959–F970. doi: 10.1152/ajprenal.00559.2007
- Steele, T. A., and Brahmi, Z. (1988). Inhibition of human natural killer cell activity by the protein kinase C inhibitor 1-(5-isoquinolinesulfonyl)-2-methylpiperazine is an early but post-binding event. *J. Immunol.* 141, 3164–3169.
- Stockand, J. D., and Sansom, S. C. (1998). Glomerular mesangial cells: electrophysiology and regulation of contraction. *Physiol. Rev.* 78, 723–744. doi: 10.1152/physrev.1998.78.3.723
- Uccellini, M. B., and García-Sastre, A. (2018). ISRE-Reporter mouse reveals high basal and induced type I IFN responses in inflammatory monocytes. *Cell Rep.* 25, 2784–2796.e2783. doi: 10.1016/j.celrep.2018.11.030
- Wen, Y., Lu, X., Ren, J., Privratsky, J. R., Yang, B., Rudemiller, N. P., et al. (2019). KLF4 in macrophages attenuates TNF α -mediated kidney injury and fibrosis. *J. Am. Soc. Nephrol.* 30, 1925–1938. doi: 10.1681/asn.2019020111
- Yang, L., Zheng, S., and Epstein, P. N. (2009). Metallothionein over-expression in podocytes reduces adriamycin nephrotoxicity. *Free Radic. Res.* 43, 174–182. doi: 10.1080/10715760802657308
- Ye, J., Chen, S., and Maniatis, T. (2011). Cardiac glycosides are potent inhibitors of interferon- β gene expression. *Nat. Chem. Biol.* 7, 25–33. doi: 10.1038/nchembio.476
- Zhang, T., Xu, J., Deng, S., Zhou, F., Li, J., Zhang, L., et al. (2018). Core signaling pathways in ovarian cancer stem cell revealed by integrative analysis of multi-marker genomics data. *PLoS One* 13:e0196351. doi: 10.1371/journal.pone.0196351
- Zheng, Z. H., Zhang, L. J., Liu, W. X., Lei, Y. S., Xing, G. L., Zhang, J. J., et al. (2012). Predictors of survival in Chinese patients with lupus nephritis. *Lupus* 21, 1049–1056. doi: 10.1177/0961203312445230

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Yao, Gao, Zhang, Di, Liang, Sun, Wang and Zheng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



An Immune Gene-Related Five-lncRNA Signature for to Predict Glioma Prognosis

Xinzhuan Wang^{1,2,3†}, Ming Gao^{1,2,3†}, Junyi Ye^{1,2,3}, Qiuyi Jiang^{1,2,3}, Quan Yang^{1,2,3}, Cheng Zhang⁴, Shengtao Wang^{1,2,3}, Jian Zhang⁵, Ligang Wang^{1,2,3}, Jianing Wu^{1,2,3}, Hua Zhan^{1,2,3}, Xu Hou^{1,2,3*}, Dayong Han^{1,2,3*} and Shiguang Zhao^{1,2,3,6*}

¹ Department of Neurosurgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China, ² Key Colleges and Universities Laboratory of Neurosurgery in Heilongjiang Province, Harbin, China, ³ Institute of Neuroscience, Sino-Russian Medical Research Center, Harbin Medical University, Harbin, China, ⁴ North Broward Preparatory School, Coconut Creek, FL, United States, ⁵ Department of General Surgery, The First Affiliated Hospital of Harbin Medical University, Harbin, China, ⁶ Department of Neurosurgery, The Pinghu Hospital of Shenzhen University, Shenzhen, China

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Qing Liu,
Central South University, China
Ranjit Das,
Yenepoya University, India

*Correspondence:

Shiguang Zhao
guangsz@hotmail.com

Xu Hou
houxu1983@hotmail.com

Dayong Han
hdybsn@163.com

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 30 September 2020

Accepted: 09 November 2020

Published: 16 December 2020

Citation:

Wang XZ, Gao M, Ye JY, Jiang QY,
Yang Q, Zhang C, Wang ST, Zhang J,
Wang LG, Wu JN, Zhan H, Hou X,
Han DY and Zhao SG (2020) An
Immune Gene-Related Five-lncRNA
Signature for to Predict Glioma
Prognosis. *Front. Genet.* 11:612037.
doi: 10.3389/fgene.2020.612037

Background: The tumor immune microenvironment is closely related to the malignant progression and treatment resistance of glioma. Long non-coding RNA (lncRNA) plays a regulatory role in this process. We investigated the pathological mechanisms within the glioma microenvironment and potential immunotherapy resistance related to lncRNAs.

Method: We downloaded datasets derived from glioma patients and analyzed them by hierarchical clustering. Next, we analyzed the immune microenvironment of glioma, related gene expression, and patient survival. Coexpressed lncRNAs were analyzed to generate a model of lncRNAs and immune-related genes. We analyzed the model using survival and Cox regression. Then, univariate, multivariate, receiver operating characteristic (ROC), and principle component analysis (PCA) methods were used to verify the accuracy of the model. Finally, GSEA was used to evaluate which functions and pathways were associated with the differential genes.

Results: Normal brain tissue maintains a low-medium immune state, and gliomas are clearly divided into three groups (low to high immunity). The stromal, immune, and estimate scores increased along with immunity, while tumor purity decreased. Further, human leukocyte antigen (HLA), programmed cell death-1 (PDL1), T cell immunoglobulin and mucin domain 3 (TIM-3), B7-H3, and cytotoxic T lymphocyte-associated antigen-4 (CTLA4) expression increases concomitantly with immune state, and the patient prognosis worsens. Five immune gene-related lncRNAs (AP001007.1, LBX-AS1, MIR155HG, MAPT-AS1, and LINC00515) were screened to construct risk models. We found that risk scores are related to patient prognosis and clinical characteristics, and are positively correlated with PDL1, TIM-3, and B7-H3 expression. These lncRNAs may regulate the tumor immune microenvironment through cytokine-cytokine receptor interactions, complement, and coagulation cascades, and may promote CD8 + T cell, regulatory T cell, M1 macrophage, and infiltrating neutrophils activity in the high-immunity group. *In vitro*, the abnormal expression of immune-related

lncRNAs and the relationship between risk scores and immune-related indicators (PDL1, CTLA4, CD3, CD8, iNOS) were verified by q-PCR and immunohistochemistry (IHC).

Conclusion: For the first time, we constructed immune gene-related lncRNA risk models. The risk score may be a new biomarker for tumor immune subtypes and provide molecular targets for glioma immunotherapy.

Keywords: tumor immune microenvironment, immune gene sets, lncRNA, glioma, risk score

INTRODUCTION

Glioma is a primary malignant tumor derived from glial cells in the central nervous system. Its annual incidence rate is 7.08 per 100,000 people, and accounts for about 75% of whole brain and other central nervous system malignancies (Lapointe et al., 2018; Ostrom et al., 2019). Clinically, gliomas are often divided into low-grade gliomas (LGGs) and glioblastomas, which have different treatment methods and prognoses. For example, LGGs are slow growing and are mainly treated by total surgical resection. The patient prognosis is relatively good (Sturm et al., 2017). However, the median survival period is less than 2 years with malignant glioblastoma progression, even with standard treatment (surgical resection, adjuvant radiotherapy, and chemotherapy) (Tan et al., 2020). In 2016, the WHO classified gliomas into five categories based on their morphology and molecular characteristics (Louis et al., 2016). Recently, immunotherapy has been used in clinical applications. However, the overall prognosis of glioblastoma patients varies greatly. This may be due to the formation of unique tumor microenvironments during long-term tumor formation and limited molecular markers that distinguish tumor subtypes (Hanahan and Weinberg, 2011). Therefore, it is important to understand the glioma immune microenvironment and screen new molecular markers, which will guide future glioma treatment.

The extracellular matrix, soluble molecules, and tumor stromal cells are the basic components of the tumor microenvironment (Cui et al., 2017). Immune cells and stromal cells are the most common non-tumor cells. Macrophages are the most abundant immune cells in brain tumors (Quail and Joyce, 2017). Glioma often recruits T cells, bone marrow-derived suppressor cells, and macrophages through several pathways to promote immune cell accumulation and transformation into different cell types (Giering et al., 2017). Microglia and macrophages are often activated to control anti-tumor immune responses, promote tumor cell proliferation and invasion, and achieve immune escape (Hambardzumyan et al., 2016). Human leukocyte antigen (HLA) (Machulla et al., 2001), programmed cell death-1 (PDL1) (Jackson et al., 2019), cytotoxic T lymphocyte-associated antigen-4 (CTLA4) (Nduom et al., 2015), T cell immunoglobulin and mucin domain 3 (TIM-3) (Das et al., 2017), and other immune-related genes participate in the immune escape process. Therefore, treatments targeting immune checkpoints, microglia, and macrophages are used in the clinic (Poon et al., 2017). However, some patients are in a state of immune tolerance. To improve the quality of

medical care and increase the understanding of the immune microenvironment, tumor immune gene analysis is common. Considering tumor-associated immune genes, investigating immune gene sets with guided evolutionary simulated annealing (GESA) can more comprehensively reflect the glioma immune microenvironment *in vivo* to better establish a prognostic model, find effective molecular markers, and perform effective targeted treatment (Molinaro et al., 2019).

With the development of high-throughput technology and the establishment of public databases, the molecular understanding of tumors has rapidly developed (Serrati et al., 2016), leading to improved understanding of tumor pathogenesis and improved biomarker screening. Importantly, some long non-coding RNA (lncRNA) has been identified as potential glioma biomarkers (Peng et al., 2018). Previously, lncRNAs were hypothesized to have no coding function and were regarded as transcriptional noise. However, lncRNAs play an important regulatory role in gene transcription and post-transcriptional modification. Indeed, lncRNA can regulate inflammation and participate in immune gene expression, thus affecting the tumor immune microenvironment (Chen, 2016; Mathy and Chen, 2017). For example, lincRNA-Cox2 regulates chromatin complex remodeling and participates in inflammatory gene expression (Hu et al., 2016). lncRNA nuclear-enriched abundant transcript 1 (NEAT1) participates in the regulation of interleukin (IL)-8 transcription, thus affecting cytokine response, and induces immune gene expression (Hirose et al., 2014). High HOTAIR lncRNA expression promotes the secretion of monocyte chemoattractant protein-1 (MCP-1/CCL2) by tumor cells and promotes the proliferation of tumor-associated macrophages (TAM) and myeloid-derived suppressor cells (MDSC) in the immune microenvironment (Botti et al., 2019). The complex relationship between lncRNAs and the tumor immune microenvironment has been gradually revealed, and the mechanism of immune-related lncRNA in a variety of tumors has been reported (Hu et al., 2019). However, the relationship between lncRNAs and the glioma immune environment remains unclear.

We analyzed glioma samples downloaded from The Cancer Genome Atlas (TCGA) and Chinese Glioma Genome Atlas (CGGA), to examine the glioma immune microenvironment using the single-sample GSEA method. Then, we screened lncRNAs related to the analyzed immune gene set. Using survival curve and Cox regression analysis, a five-lncRNA prognosis model related to the immune gene set was constructed, and the relationship between the risk score and the glioma patient

prognosis was explored. Our results provide new ideas for the clinical immunotherapy of glioma.

MATERIALS AND METHODS

Patient and Glioma Samples

This study was approved by the patients and the Ethics Committee of the First Affiliated Hospital of Harbin Medical University. All glioma tissue samples were obtained from the surgical resection tissue of glioma patients ($n = 18$); non-tumor brain tissue was used as the negative control group ($n = 5$). Tissue samples are stored separately in liquid nitrogen and paraffin embedded.

Data Extraction

Sequencing data collected from glioma patients were downloaded from public databases. We excluded samples with incomplete clinical information. In total, we downloaded 697 (168 GBM, 529 LGG) glioma RNA-seq and 669 (510 LGG, 159 GBM) clinical sample information datasets from the TCGA database¹, 1018 (375 GBM, 643 LGG) glioma RNA-seq and 971 (596 LGG, 375 GBM) clinical sample information datasets from the CGGA database² (Jiang et al., 2016; Hu et al., 2018), and 1152 normal brain RNA-seq datasets from the Genotype-Tissue Expression (GTEx) database³ (GTEx Consortium, 2015).

Immune Grouping and Correlation Analysis

In the single-sample GSEA method, each sample was scored according to 29 immune gene sets and divided into three groups by hierarchical clustering (Molinari et al., 2019). We used Estimate package to calculate the tumor microenvironment indicators for each sample and analyze the tumor purity (Yoshihara et al., 2013). Then, we used the R-x64-4.0.2 language package to analyze the three immune-related gene and patient prognosis groups. Finally, we analyzed immune cell infiltration in each tumor sample using the CIBERSORT method (Newman et al., 2015) ($p < 0.05$).

Risk Model

Nine lncRNAs were screened based on the correlation between identified lncRNAs and the immune gene sets ($R^2 \geq 0.62$) in CGGA. An additional five prognosis-related lncRNAs were identified using univariate and multivariate survival analyses by Cox regression model (Malinchoc et al., 2000). We divided the samples from the CGGA database into high- and low-risk groups according to the median risk score (Risk score = correlation_lncRNA1 $\times$ expression_lncRNA1 + correlation_lncRNA2 $\times$ expression_lncRNA2 + correlation_lncRNA n $\times$ expression_lncRNA n) (Chen et al., 2007; Zhang et al., 2020b). Survival curve and Cox regression analysis were used to construct the immune gene set-related lncRNA risk model.

¹<https://www.cancer.gov/tcga>

²<http://www.cgga.org.cn/>

³<https://commonfund.nih.gov/GTEx>

Risk Model Assessment

We used cor.test function to detect the relationships between lncRNAs (Zhang et al., 2020a). Then, we evaluated the accuracy of the risk model using univariate, multivariate, and receiver operating characteristic (ROC) curves. ggpubr package was used to show the relationship between lncRNAs, clinical symptoms, and immune status. Then, we use principal component analysis for model clustering through scatterplot3d package (Ma and Dai, 2011).

GSEA for Enrichment Analysis

We used clusterProfiler, colorspace, and enrichplot package to perform GO and KEGG analysis based on the sequence of genes which was sorted each gene in descending order of log2FoldChange [log2 (Mean of high immune group genes/Mean of low immune group genes)], and drew a bubble chart ($p < 0.05$) through ggplot2 package (Cheng et al., 2019).

Quantitative RT-PCR (qRT-PCR)

Total RNA was prepared using TRIzol Reagent (Invitrogen, Carlsbad, CA, United States) according to the manufacturer's instructions. The concentration of the total RNA was detected by NanoDrop 2000 (Thermo Scientific™). Total RNA (1000 ng) was reverse transcribed into cDNA using qPCR RT Kit (TOYOBO, Japan). Relative expression of target gene to the housekeeping gene GAPDH was determined by qRT-PCR using FastStart Universal 96 SYBR Green Master (ROX) (Roche, Germany). All primer sequence used in this study is listed in **Supplementary Table 1**. Analysis between the two groups was performed by an unpaired t -test; $P < 0.05$ was considered statistically significant.

Immunohistochemistry (IHC)

The tissue sample immersed in formalin is wrapped in paraffin and sliced into 5 μ m thick sections. Then sample sections were incubated for PDL1, CTLA4, CD3, CD8, and INOS primary antibodies at 4°C overnight and secondary antibodies at 37°C for 30 min. Next, samples were visualized by using the diaminobenzidine (DAB) substrate kit for 10 min. After intensive washing, samples were counterstained with hematoxylin, then dehydrated and coverslipped according to manufacturer's protocol. The results of immunohistochemistry (IHC) were taken with Leica microscope.

Statistical Analysis

All analyses were performed with GraphPad Prism 7, R version 3.6.1 and corresponding packages. For all data, the statistical significance is: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

RESULTS

The Tumor Immune Microenvironment Reflects Tumor Purity

Normal brain tissue maintains a low-medium immune state, while gliomas are clearly divided into low-immunity groups

(immunity_L), medium-immunity groups (immunity_M), and high-immunity groups (immunity_H) (**Supplementary Figure 1A** and **Figures 1A,B**). From immunity_L to immunity_H, the stromal score, immune score, and estimate score (stromal score combined with immune score) increase, and the tumor purity decreases. We further quantified different immunity groups scores and drew violin plots. The changes of immune stromal cells in the tumor microenvironment and the decrease in tumor purity are consistent with **Figures 1C,E**, **Supplementary Figure 1B**, **Figure 1G** (TCGA, $p < 0.001$), **Figures 1D,F**, **Supplementary Figure 1C**, and **Figure 1H** (CGGA, $p < 0.001$). In order to better understand the tumor microenvironment and find potential therapeutic targets, whether there are differences in immune-related genes is worthy of our further study.

Immune Gene Expression in the Three Groups

We generated boxplots to evaluate the expression of immune-related genes during the immune response. As shown in **Figures 2A,B**, HLA-related gene expression gradually increased from the immunity_L to immunity_H groups ($p < 0.001$). We also found that PDL1 (**Figure 3A**, TCGA; **Figure 3B**, CGGA), CTLA4 (**Figure 3C**, TCGA; **Figure 3D**, CGGA), CD96 (**Figure 3E**, TCGA; **Figure 3F**, CGGA), TIM-3 (**Figure 3G**, TCGA; **Figure 3H**, CGGA), and CD276 (**Supplementary Figure 1D**, TCGA; **Supplementary Figure 1E**, CGGA) expression levels also increased from the immunity_L to immunity_H groups. However, HLA-related gene expression promotes immune responses to clear tumors, while immune checkpoint genes (PDL1, CTLA4, TIM-3, and CD276) suppress immune responses and facilitate tumor proliferation and metastasis. Therefore, we further analyzed patient outcomes.

Patient Prognosis in the Different Immune Groups

To analyze the effect of gene expression on patient prognosis in the different immune groups, we drew survival curves for the TCGA (669 samples: 510 LGG and 159 GBM samples) and the CGGA (971 samples: 596 LGG and 375 GBM samples) (**Figure 4**). Among the glioma patients, patients in the immune_L group had the best prognosis, followed by the immune_M group, and the immune_H group had the worst prognosis (**Figure 4B**, $p < 0.001$). Among LGGs, prognosis in the different immune groups was similar (**Figures 4C,D**, $p < 0.001$). In GBM, the prognosis in the immune_H group tended to be worse than in the immune_L group, but the difference was not statistically significant ($p > 0.05$).

Risk Models of Five lncRNAs Related to the Immune Gene Sets

Nine lncRNAs were screened based on their coexpression with immune-related genes. The nine lncRNAs we screened were AC084018.1, AP001007.1, DICER1-AS1, HCP5, LBX2-AS1, LINC00515, MAPT-AS1, USP30-AS1, and MIR155HG. After univariate (**Figure 5A**) and multivariate (**Figure 5B**) analyses,

AP001007.1, MIR155HG, and LBX2-AS1 were identified as independent risk factors [hazard ratio (HR) > 1 , $P < 0.05$], and LINC00515 and MAPT-AS1 were identified as independent protective factors (HR < 1 , $P < 0.001$). All the lncRNAs were related to prognosis in CGGA-mRNAseq_325 and CGGA-mRNAseq_625 samples (**Supplementary Figure 2**, $p < 0.001$). Then, five lncRNAs (AP001007.1, MIR155HG, LBX2-AS1, LINC00515, and MAPT-AS1) were used to construct a risk model and draw survival (**Figure 5C**, $p < 0.001$) and risk curves (**Figure 5D**). The results show that as the patient risk increases, the survival time decreases, and the overall death rate increases. Finally, correlation analysis showed that the primary, recurrent, and secondary (PRS) type, World Health Organization (WHO) grade, isocitrate dehydrogenase (IDH)-mutant, 1p/19q co-deleted, age, and risk score were independent prognostic factors (**Figures 5E,F**, $p < 0.05$). Importantly, the risk score [area under the curve (AUC) = 0.732] and WHO (AUC = 0.747) had potential diagnostic value (**Figure 6A**). Thus, our risk model has clear diagnostic value.

Clinical Characteristics of the Five lncRNAs

We next clarified the correlation between lncRNAs and clinical characteristics based on CGGA database. The results indicated that as the WHO level increased, AP001007.1, LBX-AS1, and MIR155HG expression also increased, while MAPT-AS1 and LINC00515 expression decreased (**Figure 6B**, $p < 0.001$). In addition, 1p19q no-codeletion (**Figure 6C**), IDH1 wildtype (**Figure 6D**), MGMT un-methylated (**Figure 6E**), and recurrent glioma (**Figure 6F**) compared with 1p19q deletion (**Figure 6C**), IDH1 mutant (**Figure 6D**), MGMT methylated (**Figure 6E**), and primary glioma (**Figure 6F**), AP001007.1, LBX-AS1, and MIR155HG also was high expression, while MAPT-AS1 and LINC00515 were also low in CGGA (except for LINC00515 in **Figure 6E** and MAPT-AS1 in **Figure 6F**, $p > 0.05$). Then, principal component analysis also showed that the risk model could divide the high- and low-risk groups into different subgroups (**Figure 6G**).

The Correlation Between lncRNA and Immunity

Using correlation analysis, we found that the lncRNAs in the risk model are associated (**Supplementary Figure 3A**). The risk score is closely related to the lncRNAs, PDL1, TIM-3, and B7-H3 (**Supplementary Figures 3B–I**). In addition, we found that AP001007.1, LBX2-AS1, and MIR155HG had the highest expression, while MAPT-AS1 and LINC00515 expression was the lowest in the immune_H group. In contrast, the expression of AP001007.1, LBX-AS1, and MIR155HG were relatively low, while the expressions of MAPT-AS1 and LINC00515 were relatively high in the immune_L group (**Figure 6H**). We next analyzed the immune-infiltrating cells in each group. In the immune-H group, we found that naive B cells, plasma cells, CD8 + T cells, regulatory T cells (Tregs), M1 macrophages, M2 macrophages, resting mast cells resting, and infiltrating neutrophils increased. CD4 + naive T cell, inactivated

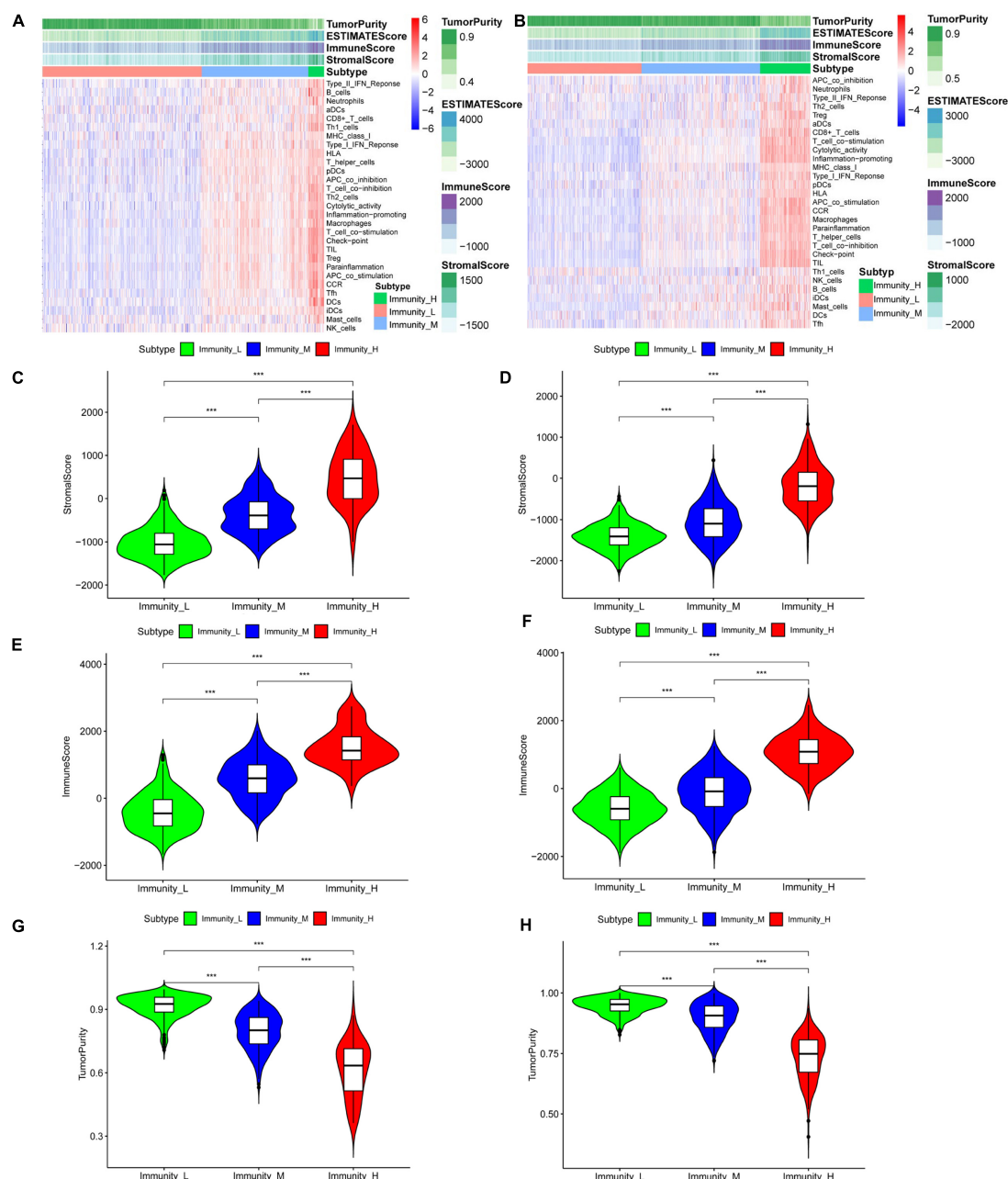


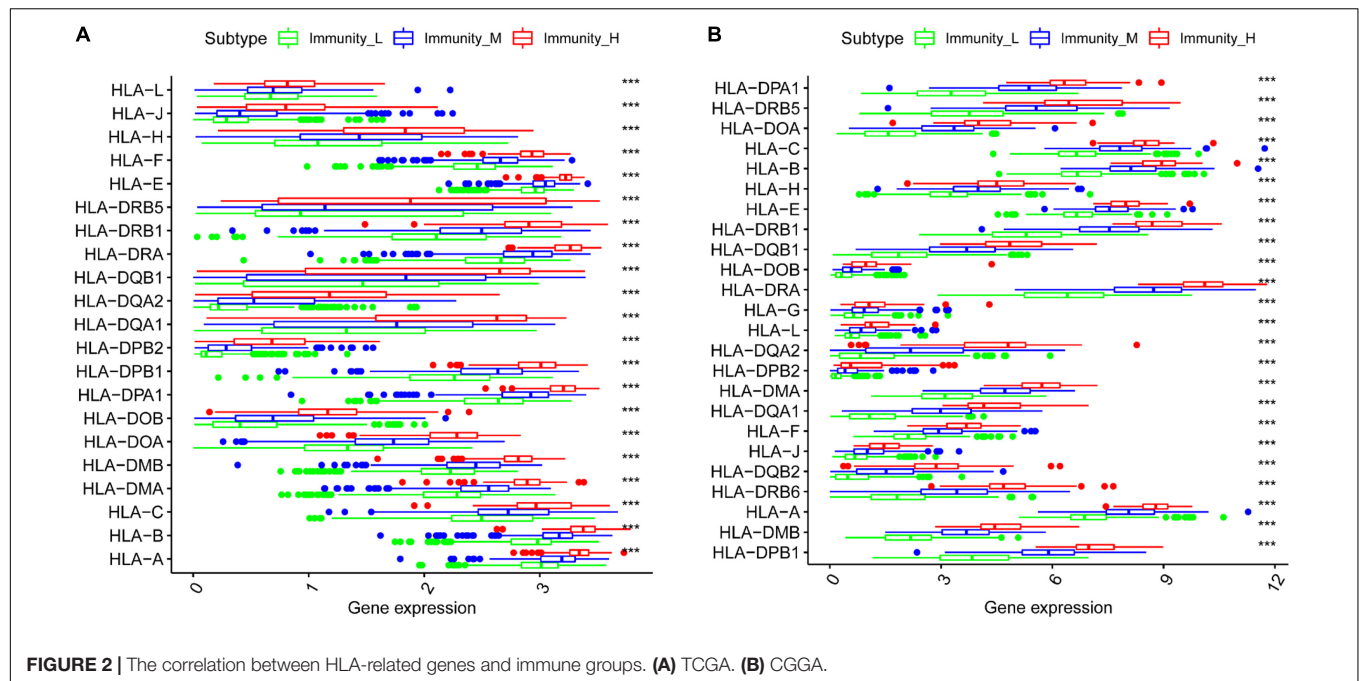
FIGURE 1 | The tumor immune microenvironment is related to the expressed immune genes. Heatmaps of the tumor immune microenvironment in the TCGA (A) and CGGA (B) datasets. Violin plots of the stromal cell scores among immune groups in the TCGA (C) and CGGA (D) datasets. Violin plots of the immune scores among the immune groups in the TCGA (E) and CGGA (F) datasets. Violin plots of tumor purity in the TCGA (G) and CGGA (H) datasets, *** $P < 0.001$.

CD4 + memory T cell, monocyte, inactivated natural-killer (NK) cell, and activated NK cell infiltration decreased (Figure 7A, TCGA; Figure 7B, CGGA $p < 0.05$).

GO Enrichment and KEGG Pathway Analysis

We used GSEA to analyze enriched differential genes in the immune_H and immune_L groups. We observed that the

differential genes were enriched in immunoglobulin complex, circulating, immunoglobulin receptor binding, and MHC protein complex (Figures 7C–F). Further KEGG function analysis (Figure 8A and Supplementary Figure 3J) showed allograft rejection, asthma, intestinal immune network for IgA production, and cytokine–cytokine receptor interaction may be activated cell signaling pathways. The intersection of the two data sets revealed 81 cell signal pathways involved in glioma (Figure 8B). The five lncRNA we identified may regulate the



immune microenvironment through cytokine–cytokine receptor interaction, antigen processing and presentation, complement and coagulation cascades, and intestinal immune network for IgA production (Figures 8C,D).

***In vitro* Validation of the Risk Model**

Through qRT-PCR, we confirmed that AP001007.1, MIR155HG, and LBX2-AS1 are highly expressed, while LINC00515 and MAPT-AS1 are low expressed in gliomas compared to the control group (Figure 9A). Then, we further found that the risk score was positively correlated with the expression of PDL1, CTLA4, CD3, CD8, and INOS (Figure 9B, $\text{cor} > 0.5$). Finally, the immunohistochemical results also confirmed that the expression of PDL1, CTLA4, CD3, CD8, and INOS in the high risk group was significantly higher than low risk group base on protein level (Figure 9C).

DISCUSSION

Glioma cells form a complex regulatory network via the extracellular matrix, stromal cells, and infiltrating immune cells (Hanahan and Coussens, 2012). Some cells secrete factors and lncRNAs to promote inflammation and angiogenesis in tumors, thereby promoting malignant tumor progression and immune escape (Pitt et al., 2016; Dagogo-Jack and Shaw, 2018). It is critical to understand the tumor immune microenvironment and screen new markers to enable targeted glioma therapy for glioma. Abundant research on immune cells has been performed (Charoentong et al., 2017; Jia et al., 2018). However, the cell types, functions, and pathways associated with glioma remain unclear. Therefore, we analyzed 1715 glioma and 1152 normal brain tissue samples using the single-sample GSEA method. We found that

the immune environment in gliomas was very different from the immune environment in normal brain tissues. In the immune_H group, the tumor immune cell, and stromal cell content increases, the tumor purity decreases, and tumor heterogeneity becomes greater. These conclusions are in line with previous findings (Hanahan and Coussens, 2012; Pitt et al., 2016; Dagogo-Jack and Shaw, 2018) indicating that this method can accurately reflect the basic conditions of the tumor microenvironment.

Human leukocyte antigens and immune checkpoints are an indispensable regulator of the immune microenvironment (Topalian et al., 2016; Pereira et al., 2019). In the immune_H group, PDL1, CTLA4, TIM-3, and CD96 expression were increased. Immune checkpoints act to negatively regulate immune regulation. Normal immune surveillance and cell killing ability are weakened in many tumors. Further, tumors often have immune escape or immunotherapy resistance mechanisms, leading to ineffective clinical treatment (Field et al., 2017; Qian et al., 2018). We also observed that among all the glioma samples, the immune_H group had the worst prognosis, followed by the immune_M group, and the immune_L group had the best prognosis. This conclusion also supports previous results. Moreover, GSEA has been used in many studies and has a certain degree of credibility (Ma et al., 2020; Wang et al., 2020). Therefore, the single-sample GSEA method based on expressed immune genes can distinguish the biological characteristics of the immune microenvironment between different gliomas, which provides the possibility for screening the immune microenvironment-related biomarkers.

Long non-coding RNAs can regulate the tumor immune microenvironment and can be used as biomarkers. For example, NF-kappaB interacting lncRNA (NKILA) can promote the immune escape of tumor cells by regulating T cell activity (Huang et al., 2018). Further, SATB2-AS (the antisense transcript

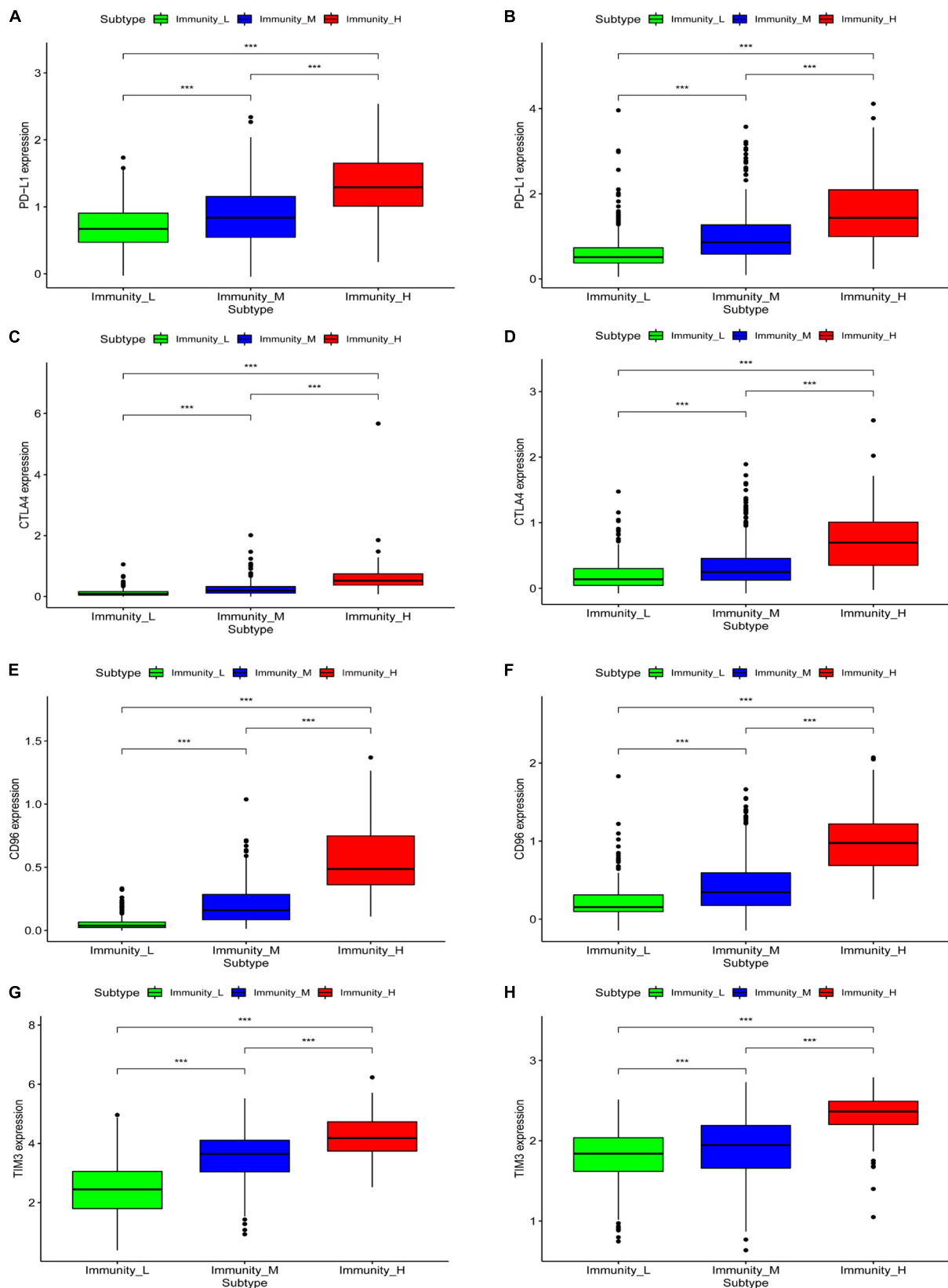


FIGURE 3 | The correlation between immune groups and immune checkpoints. PDL1 expression is based on TCGA (A) and CGGA (B), CTLA-4 expression is based on TCGA (C) and CGGA (D), CD96 expression is based on TCGA (E) and CGGA (F), TIM-3 expression is based on TCGA (G) and CGGA (H), *** $p < 0.001$.

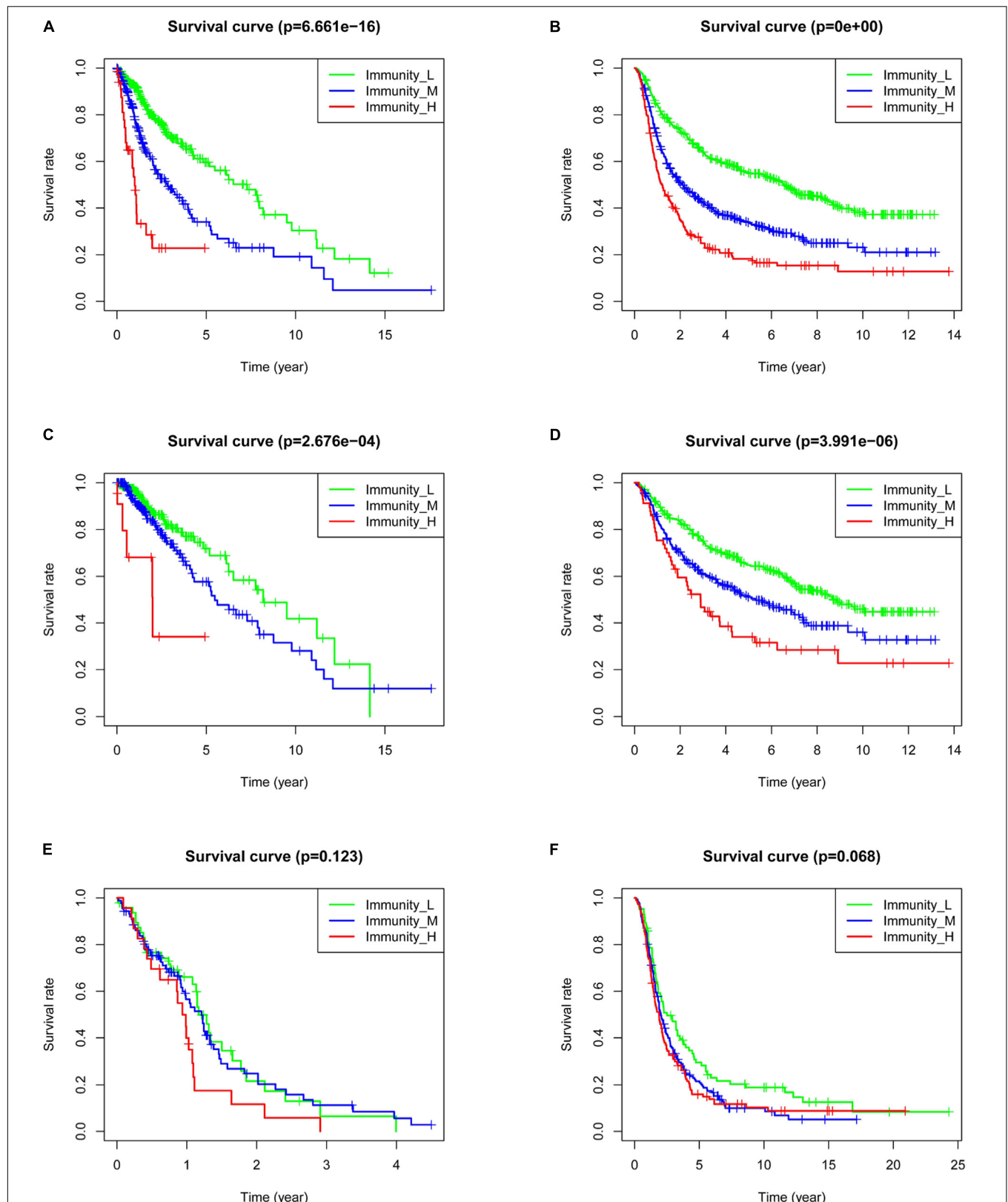
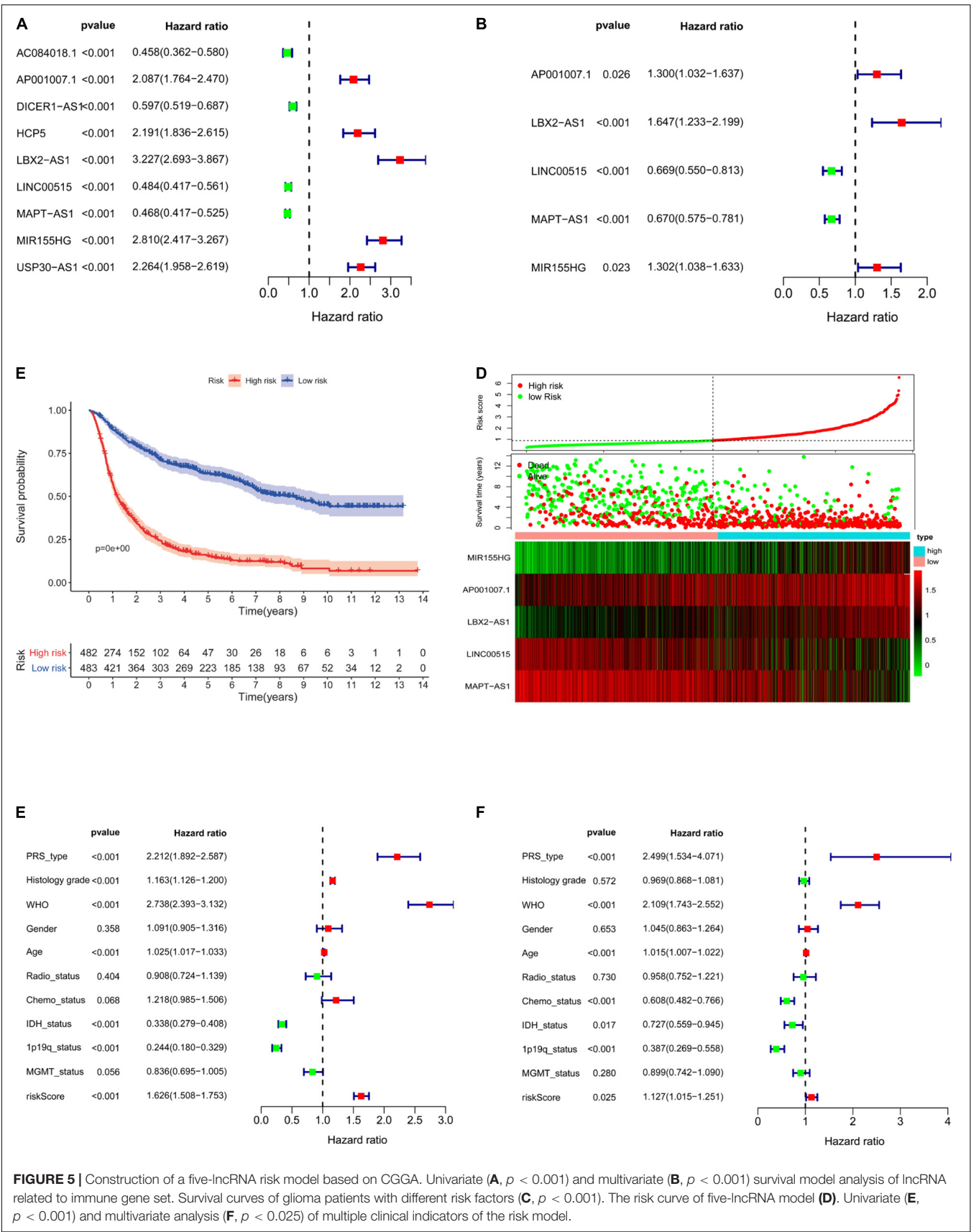


FIGURE 4 | Correlation between immune grouping and survival time of glioma patients. **(A)** TCGA and **(B)** CGGA in all glioma patients. **(C)** TCGA and **(D)** CGGA in low-grade glioma patients. **(E)** TCGA and **(F)** CGGA in GBM patients.



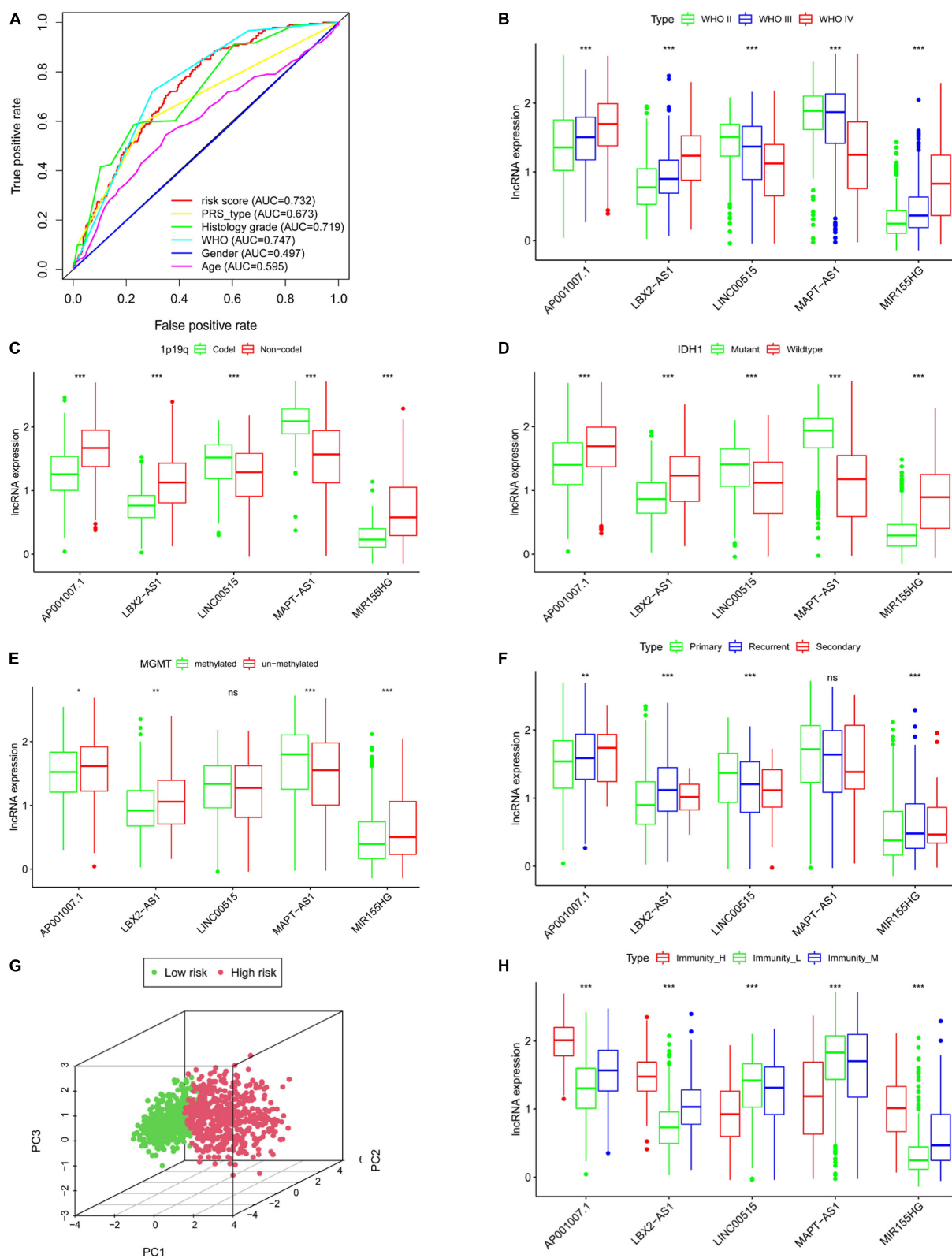
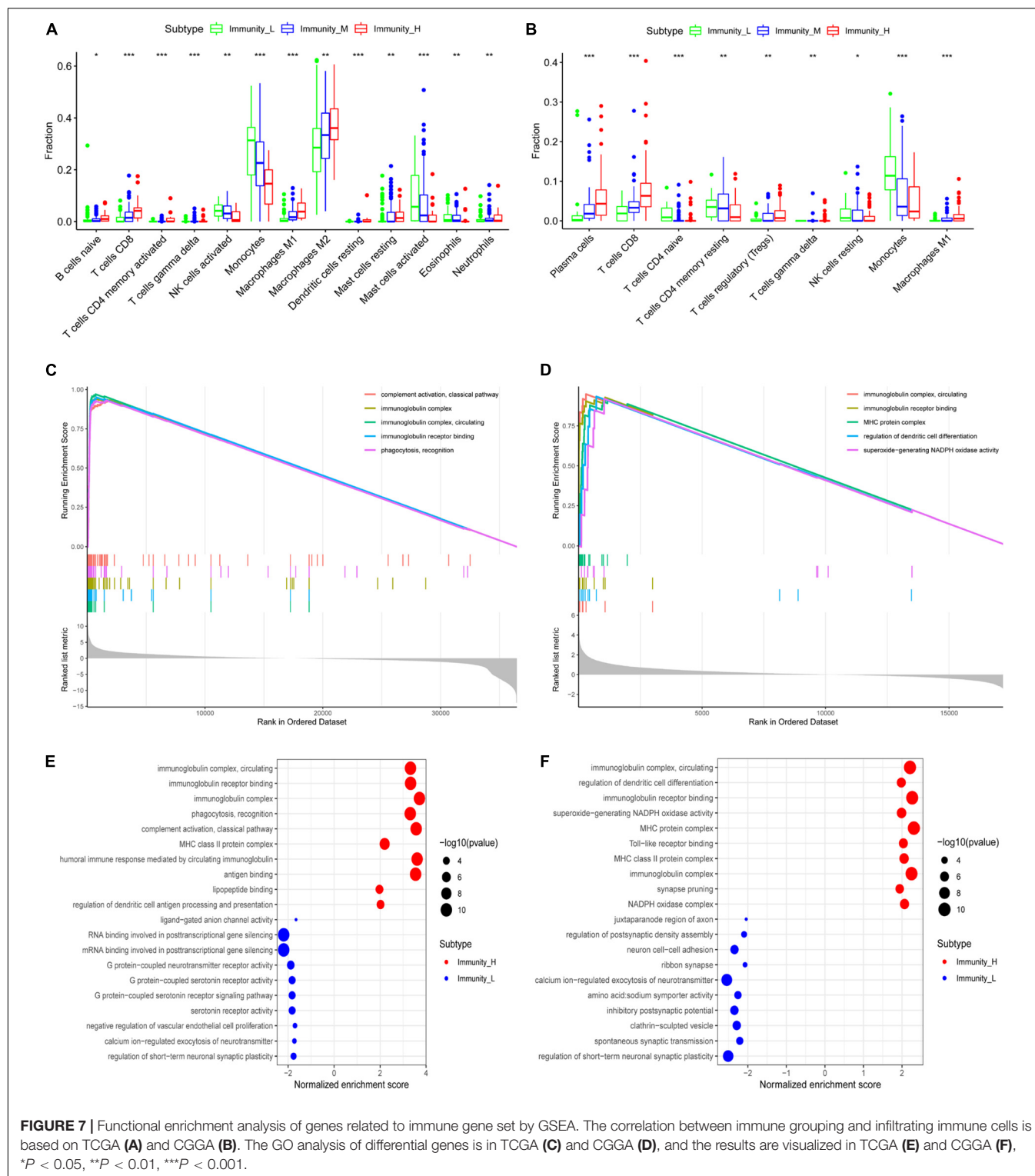
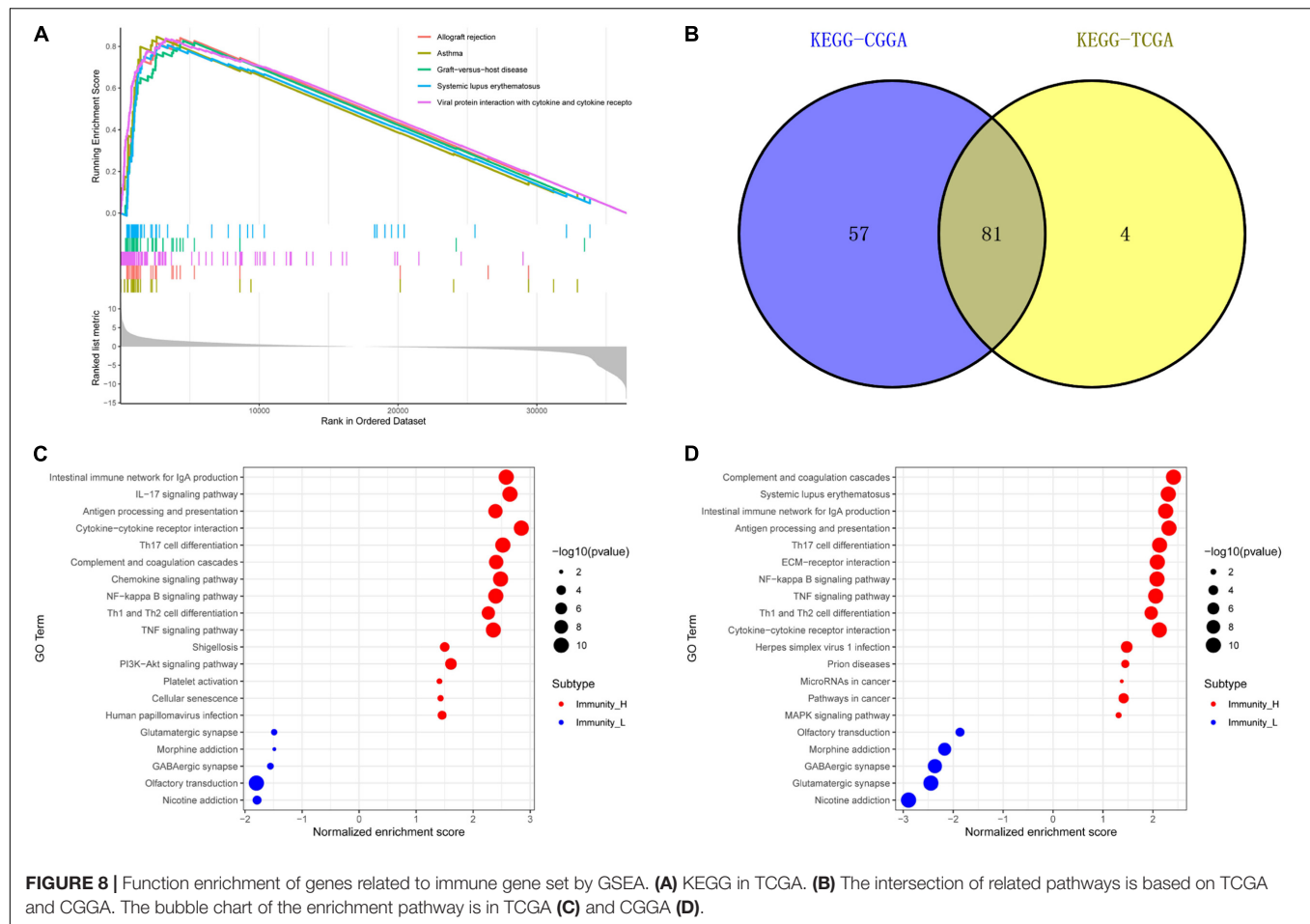


FIGURE 6 | The clinical characteristics of the risk model are based on CGGA. **(A)** Roc curves of multiple clinical indicators. WHO grade **(B)**, 1p19q status **(C)**, IDH status **(D)**, MGMT methylation status **(E)**, PRS type **(F)**. **(G)** Principal component analysis of lncRNA related to immune gene set. **(H)** The expression of lncRNAs in different immune groups, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns: not statistically significant.



of SATB2 protein) can directly combine with WDR5 (WD repeat containing protein 5) and GADD45A (growth arrest and DNA damage protein 45A) to regulate SATB2 expression, thereby inhibiting tumor cell metastasis and regulating the tumor immune microenvironment (Xu et al., 2019). Immune-related

lncRNAs have carcinogenic effects in several tumors, and can be used as biomarkers (Li et al., 2020). Thus, it is very important to determine whether lncRNAs related to the immune gene set have clinical diagnostic and prognostic value. We screened five lncRNAs using the Cox regression method and constructed a

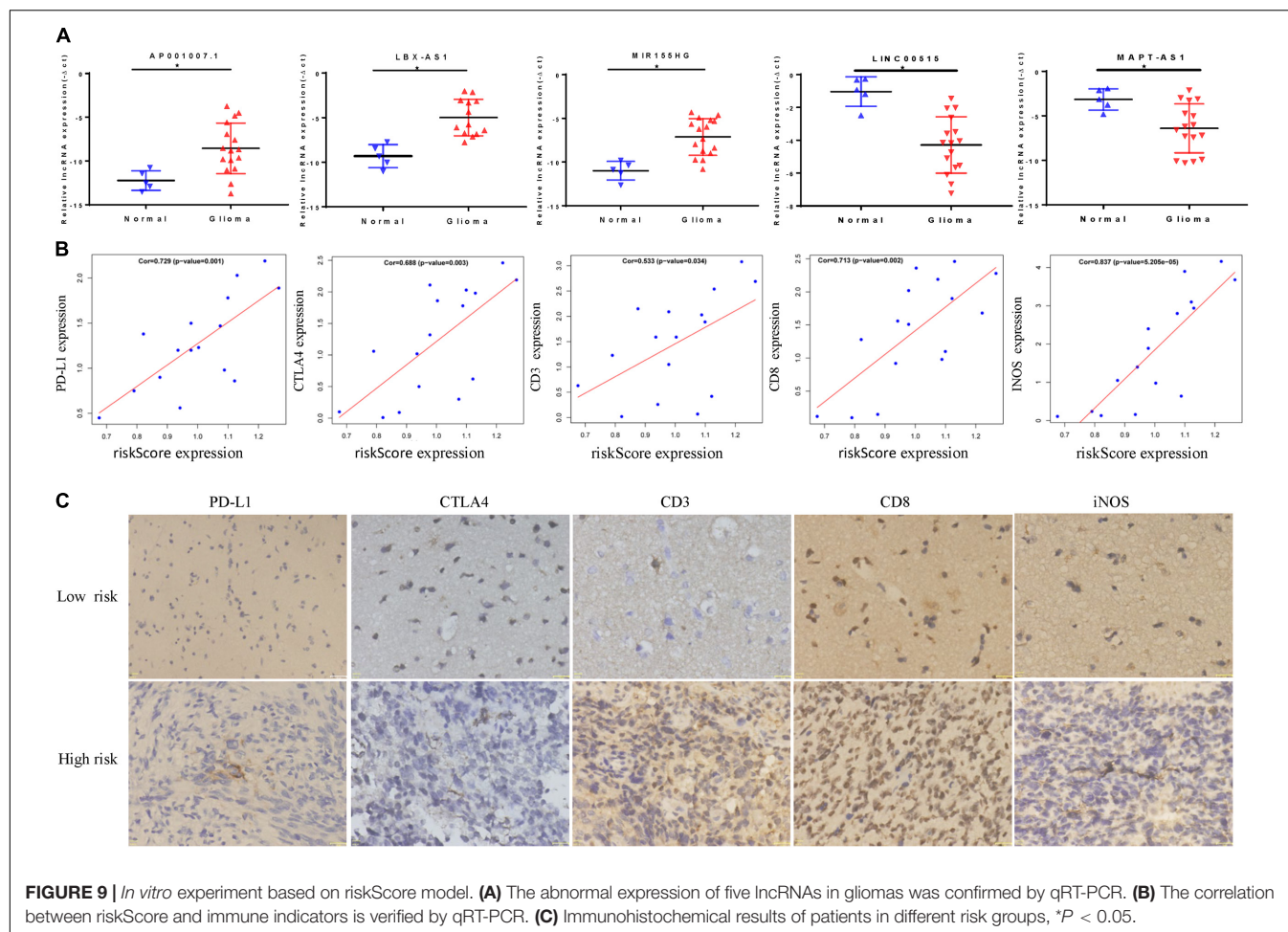


prognostic model. We found that the risk score is related to prognosis and is an independent factor that can be used for clinical diagnosis. We further observed that five lncRNAs interact and are closely related to the clinical symptoms of glioma patients (WHO grade, IDH1 status, 1q19q status, and MGMT). Principle component analysis (PCA) analysis showed that subgroups within the high- and low-risk groups can be well distinguished using our method. These conclusions show that the risk scores of the five lncRNAs related to the immune gene set can predict patient prognosis and clinical characteristics, and can be used as a new biomarker to inform clinical diagnosis and treatment.

We used boxplots to visualize lncRNA expression in each immune group. In the immune_H group, we found that AP001007.1, LBX2-AS1, and MIR155HG were highly expressed, while LINC00515 and MAPT-AS1 expression was low. The immune_L group showed the opposite trend. Survival analysis showed that AP001007.1, LBX2-AS1, and MIR155HG were risk factors, and their high expression predicted poor patient outcomes. LINC00515 and MAPT-AS expression were protective indicators, and low expression predicted poor patient prognosis. LBX2-AS1 produces malignant behavior in gliomas by conferring resistance to cell apoptosis (Chen et al., 2020). MIR155HG promotes immune cell infiltration and immune resistance

(Peng et al., 2019). In contrast, MAPT-AS1 expression indicates a good prognosis for cancer patients (Wang et al., 2019). Therefore, the immune gene set-related model we constructed has considerable credibility.

We found that the five lncRNAs we analyzed may promote immune cell infiltration through cytokine-cytokine receptor interaction, antigen processing and presentation, complement and coagulation cascades, and may contribute to immune resistance and tolerance, ultimately leading to poor patient prognosis. However, this study also has some limitations. First, because there was no information regarding MGMT, 1p19q, and other related molecules in the TCGA dataset, only a single CGGA cohort was used for statistical analysis. The CGGA sample information is mainly clinical sample information from Chinese patients, which may only indicate region-specific effects. Second, basic experiments have also verified the important role of some immune gene-related lncRNAs in regulating glioma development. However, the mechanism underlying our prognostic model remains unclear and requires additional experiments to verify our *in silico* results. Third, we confirmed that the five lncRNA have potential clinical value to identify risk factors, but more factors should be considered, especially considering multimodal glioma development.



CONCLUSION

Immune-related genes can reflect the characteristics of the immune microenvironment. To reveal the mechanism of partial resistance or treatment resistance within a new risk model, five immune-related lncRNAs were analyzed and shown to have good stability and feasibility (AP001007.1, LBX-AS1, MIR155HG, MAPT-AS1, and LINC00515). Thus, our study reveals biomarkers that distinguish specific glioma groups and can be used in the clinical diagnosis and treatment of glioma.

DATA AVAILABILITY STATEMENT

The public databases mined in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

This study was approved by the Ethics Committee of the First Affiliated Hospital of Harbin Medical University.

AUTHOR CONTRIBUTIONS

XZW, MG, JYY, and QY collated and analyzed the data. CZ, QY, STW, DYH, XH, and LGW completed the writing and repair of the manuscript. JZ, HZ, JNW, and SGZ designed and guided the subject. All authors reviewed and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81972363 and 81802501), the Natural Science Foundation of The First Affiliated Hospital of Harbin Medical University (2016B006), and Scientific Research Innovation Project of the First Affiliated Hospital of Harbin Medical University (YJSKYCX2019-42HYD).

ACKNOWLEDGMENTS

We thank all the members in Key Colleges and Universities Laboratory of Neurosurgery for help with our study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.612037/full#supplementary-material>

Supplementary Figure 1 | (A) The heatmap of the immune microenvironment of normal brain tissue base on GTEx. The relationship between immune grouping and ESTIMATEScore in TCGA (B) and CGGA (C). The correlation between CD276 and immune grouping in TCGA (D) and CGGA (E).

Supplementary Figure 2 | Survival curve of glioma patients in CGGA mRNAseq_693 and CGGA mRNAseq_325. (A) ap0001007.1, (C) LBX2-AS1, (E) LINC00515, (G) MAPT-AS1, (I) MIR155HG in CGGA mRNAseq_325. (B) ap0001007.1, (D) LBX2-AS1, (F) LINC00515, (H) MAPT-AS1, (G) MIR155HG in CGGA mRNAseq_325 database.

Supplementary Figure 3 | Correlation analysis of risk score, lncRNA, and immune checkpoint. (A) Correlation analysis of lncRNAs related to immune gene set. (B) KEGG in CGGA. (C) Analysis of the correlation between risk scores and lncRNAs or immune checkpoints in (B) LINC00515, (C) MIR155HG, (D) LBX2-AS1, (E) MAPT-AS1, (F) ap0001007.1 or (G) TIM-3, (H) B7-H3, (I) PDL1, (J) KEGG in CGGA.

REFERENCES

- Botti, G., Scognamiglio, G., Aquino, G., Liguori, G., and Cantile, M. (2019). LncRNA HOTAIR in tumor microenvironment: what role? *Int. J. Mol. Sci.* 20:2279. doi: 10.3390/ijms20092279
- Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019
- Chen, H. Y., Yu, S. L., Chen, C. H., Chang, G. C., Chen, C. Y., Yuan, A., et al. (2007). A five-gene signature and clinical outcome in non-small-cell lung cancer. *N. Engl. J. Med.* 356, 11–20. doi: 10.1056/NEJMoa060096
- Chen, L. L. (2016). Linking long noncoding RNA localization and function. *Trends Biochem. Sci.* 41, 761–772. doi: 10.1016/j.tibs.2016.07.003
- Chen, Q., Gao, J., Zhao, Y., and Hou, R. (2020). Long non-coding RNA LBX2-AS1 enhances glioma proliferation through downregulating microRNA-491-5p. *Cancer Cell Int.* 20:411. doi: 10.1186/s12935-020-01433-2
- Cheng, Q., Huang, C., Cao, H., Lin, J., Gong, X., Li, J., et al. (2019). A novel prognostic signature of transcription factors for the prediction in patients with GBM. *Front. Genet.* 10:906. doi: 10.3389/fgene.2019.00906
- Cui, D., Huang, Z., Liu, Y., and Ouyang, G. (2017). The multifaceted role of periostin in priming the tumor microenvironments for tumor progression. *Cell Mol. Life Sci.* 74, 4287–4291. doi: 10.1007/s00018-017-2646-2
- Dagogo-Jack, I., and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15, 81–94. doi: 10.1038/nrclinonc.2017.166
- Das, M., Zhu, C., and Kuchroo, V. K. (2017). Tim-3 and its role in regulating anti-tumor immunity. *Immunol. Rev.* 276, 97–111. doi: 10.1111/imr.12520
- Field, C. S., Hunn, M. K., Ferguson, P. M., Ruedl, C., Ancelet, L. R., and Hermans, I. F. (2017). Blocking CTLA-4 while priming with a whole cell vaccine reshapes the oligoclonal T cell infiltrate and eradicates tumors in an orthotopic glioma model. *Oncoimmunology* 7:e1376154. doi: 10.1080/2162402X.2017.1376154
- Gierzyng, A., Pszczolkowska, D., Walentynowicz, K. A., Rajan, W. D., and Kaminska, B. (2017). Immune microenvironment of gliomas. *Lab Invest.* 97, 498–518. doi: 10.1038/labinvest.2017.19
- GTEx Consortium (2015). Human genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648–660. doi: 10.1126/science.1262110
- Hambardzumyan, D., Gutmann, D. H., and Kettenmann, H. (2016). The role of microglia and macrophages in glioma maintenance and progression. *Nat. Neurosci.* 19, 20–27. doi: 10.1038/nn.4185
- Hanahan, D., and Coussens, L. M. (2012). Accessories to the crime: functions of cells recruited to the tumor microenvironment. *Cancer Cell* 21, 309–322. doi: 10.1016/j.ccr.2012.02.022
- Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013
- Hirose, T., Virnecchi, G., Tanigawa, A., Naganuma, T., Li, R., Kimura, H., et al. (2014). NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol. Biol. Cell* 25, 169–183. doi: 10.1091/mbc.E13-09-0558
- Hu, G., Gong, A. Y., Wang, Y., Ma, S., Chen, X., Chen, J., et al. (2016). LincRNA-Cox2 promotes late inflammatory gene transcription in macrophages through modulating SWI/SNF-mediated chromatin remodeling. *J. Immunol.* 196, 2799–2808. doi: 10.4049/jimmunol.1502146
- Hu, H., Mu, Q., Bao, Z., Chen, Y., Liu, Y., Chen, J., et al. (2018). Mutational landscape of secondary glioblastoma guides MET-Targeted trial in brain tumor. *Cell* 175, 1665.e18–1678.e18. doi: 10.1016/j.cell.2018.09.038
- Hu, Q., Ye, Y., Chan, L. C., Li, Y., Liang, K., Lin, A., et al. (2019). Oncogenic lncRNA downregulates cancer cell antigen presentation and intrinsic tumor suppression. *Nat. Immunol.* 20, 835–851. doi: 10.1038/s41590-019-0400-7
- Huang, D., Chen, J., Yang, L., Ouyang, Q., Li, J., Lao, L., et al. (2018). NKILA lncRNA promotes tumor immune evasion by sensitizing T cells to activation-induced cell death. *Nat. Immunol.* 19, 1112–1125. doi: 10.1038/s41590-018-0207-y
- Jackson, C. M., Choi, J., and Lim, M. (2019). Mechanisms of immunotherapy resistance: lessons from glioblastoma. *Nat. Immunol.* 20, 1100–1109. doi: 10.1038/s41590-019-0433-y
- Jia, Q., Wu, W., Wang, Y., Alexander, P. B., Sun, C., Gong, Z., et al. (2018). Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. *Nat. Commun.* 9:5361. doi: 10.1038/s41467-018-07767-w
- Jiang, T., Mao, Y., Ma, W., Mao, Q., You, Y., Yang, X., et al. (2016). CGCG clinical practice guidelines for the management of adult diffuse gliomas. *Cancer Lett.* 375, 263–273. doi: 10.1016/j.canlet.2016.01.024
- Lapointe, S., Perry, A., and Butowski, N. A. (2018). Primary brain tumours in adults. *Lancet* 392, 432–446. doi: 10.1016/S0140-6736(18)30990-5
- Li, Y., Jiang, T., Zhou, W., Li, J., Li, X., Wang, Q., et al. (2020). Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat. Commun.* 11:1000. doi: 10.1038/s41467-020-14802-2
- Louis, D. N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W. K., et al. (2016). The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol.* 131, 803–820. doi: 10.1007/s00401-016-1545-1
- Ma, B., Li, Y., and Ren, Y. (2020). Identification of a 6-lncRNA prognostic signature based on microarray re-annotation in gastric cancer. *Cancer Med.* 9, 335–349. doi: 10.1002/cam4.2621
- Ma, S., and Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. *Brief. Bioinform.* 12, 714–722. doi: 10.1093/bib/bbq090
- Machulla, H. K., Steinborn, F., Schaaf, A., Heidecke, V., and Rainov, N. G. (2001). Brain glioma and human leukocyte antigens (HLA)- is there an association. *J. Neurooncol.* 52, 253–261. doi: 10.1023/a:1010612327647
- Malinchoc, M., Kamath, P. S., Gordon, F. D., Peine, C. J., Rank, J., and ter Borg, P. C. (2000). A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts. *Hepatology* 31, 864–871. doi: 10.1053/he.2000.5852
- Mathy, N. W., and Chen, X. M. (2017). Long non-coding RNAs (lncRNAs) and their transcriptional control of inflammatory responses. *J. Biol. Chem.* 292, 12375–12382. doi: 10.1074/jbc.R116.760884
- Molinaro, A. M., Taylor, J. W., Wiencke, J. K., and Wrensch, M. R. (2019). Genetic and molecular epidemiology of adult diffuse glioma. *Nat. Rev. Neurol.* 15, 405–417. doi: 10.1038/s41582-019-0220-2
- Nduom, E. K., Weller, M., and Heimberger, A. B. (2015). Immunosuppressive mechanisms in glioblastoma. *Neuro Oncol.* 17(Suppl. 7), vii9–vii14. doi: 10.1093/neuonc/nov151
- Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* 12, 453–457. doi: 10.1038/nmeth.3337

- Ostrom, Q. T., Cioffi, G., Gittleman, H., Patil, N., Waite, K., Kruchko, C., et al. (2019). CBTRUS statistical report: primary brain and other central nervous system tumors diagnosed in the United States in 2012-2016. *Neuro Oncol.* 21(Suppl. 5), v1-v100. doi: 10.1093/neuonc/noz150
- Peng, L., Chen, Z., Chen, Y., Wang, X., and Tang, N. (2019). MIR155HG is a prognostic biomarker and associated with immune infiltration and immune checkpoint molecules expression in multiple cancers. *Cancer Med.* 8, 7161-7173. doi: 10.1002/cam4.2583
- Peng, Z., Liu, C., and Wu, M. (2018). New insights into long noncoding RNAs and their roles in glioma. *Mol Cancer* 17:61. doi: 10.1186/s12943-018-0812-2
- Pereira, B. I., Devine, O. P., Vukmanovic-Stejić, M., Chambers, E. S., Subramanian, P., Patel, N., et al. (2019). Senescent cells evade immune clearance via HLA-E-mediated NK and CD8+ T cell inhibition. *Nat. Commun.* 10:2387. doi: 10.1038/s41467-019-10335-5
- Pitt, J. M., Marabelle, A., Eggermont, A., Soria, J. C., Kroemer, G., and Zitvogel, L. (2016). Targeting the tumor microenvironment: removing obstruction to anticancer immune responses and immunotherapy. *Ann. Oncol.* 27, 1482-1492. doi: 10.1093/annonc/mdw168
- Poon, C. C., Sarkar, S., Yong, V. W., and Kelly, J. J. P. (2017). Glioblastoma-associated microglia and macrophages: targets for therapies to improve prognosis. *Brain* 140, 1548-1560. doi: 10.1093/brain/aww355
- Qian, J., Wang, C., Wang, B., Yang, J., Wang, Y., Luo, F., et al. (2018). The IFN- γ /PD-L1 axis between T cells and tumor microenvironment: hints for glioma anti-PD-1/PD-L1 therapy. *J. Neuroinflammation* 15:290. doi: 10.1186/s12974-018-1330-2
- Quail, D. F., and Joyce, J. A. (2017). The Microenvironmental Landscape of Brain Tumors. *Cancer Cell* 31, 326-341. doi: 10.1016/j.ccell.2017.02.009
- Serrati, S., De Summa, S., Pilato, B., Petriella, D., Lacalamita, R., Tommasi, S., et al. (2016). Next-generation sequencing: advances and applications in cancer diagnosis. *Onco Targets Ther.* 9, 7355-7365. doi: 10.2147/OTT.S99807
- Sturm, D., Pfister, S. M., and Jones, D. T. W. (2017). Pediatric gliomas: current concepts on diagnosis, biology, and clinical management. *J. Clin. Oncol.* 35, 2370-2377. doi: 10.1200/JCO.2017.73.0242
- Tan, A. C., Ashley, D. M., López, G. Y., Malinzak, M., Friedman, H. S., and Khasraw, M. (2020). Management of glioblastoma: state of the art and future directions. *CA Cancer J. Clin.* 70, 299-312. doi: 10.3322/caac.21613
- Topalian, S. L., Taube, J. M., Anders, R. A., and Pardoll, D. M. (2016). Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat. Rev. Cancer* 16, 275-287. doi: 10.1038/nrc.2016.36
- Wang, D., Li, J., Cai, F., Xu, Z., Li, L., Zhu, H., et al. (2019). Overexpression of MAPT-AS1 is associated with better patient survival in breast cancer. *Biochem. Cell Biol.* 97, 158-164. doi: 10.1139/bcb-2018-0039
- Wang, F., Cao, X., Yin, L., Wang, Q., and He, Z. (2020). Identification of SCARA5 gene as a potential immune-related biomarker for triple-negative breast cancer by integrated analysis. *DNA Cell Biol.* 39, 1813-1824. doi: 10.1089/dna.2020.5449
- Xu, M., Xu, X., Pan, B., Chen, X., Lin, K., Zeng, K., et al. (2019). LncRNA SATB2-AS1 inhibits tumor metastasis and affects the tumor immune cell microenvironment in colorectal cancer by regulating SATB2. *Mol. Cancer* 18:135. doi: 10.1186/s12943-019-1063-6
- Yoshihara, K., Shahmoradgol, M., Martínez, E., Vegesna, R., Kim, H., Torres-García, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612
- Zhang, M., Sun, L., Ru, Y., Zhang, S., Miao, J., Guo, P., et al. (2020a). A risk score system based on DNA methylation levels and a nomogram survival model for lung squamous cell carcinoma. *Int. J. Mol. Med.* 46, 252-264. doi: 10.3892/ijmm.2020.4590
- Zhang, M., Wang, X., Chen, X., Zhang, Q., and Hong, J. (2020b). Novel immune-related gene signature for risk stratification and prognosis of survival in lower-grade glioma. *Front. Genet.* 11:363. doi: 10.3389/fgene.2020.00363

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Wang, Gao, Ye, Jiang, Yang, Zhang, Wang, Zhang, Wang, Wu, Zhan, Hou, Han and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Mutation Frequency in Main Susceptibility Genes Among Patients With Head and Neck Paragangliomas

Anastasiya V. Snezhkina^{1*}, Maria S. Fedorova¹, Vladislav S. Pavlov¹, Dmitry V. Kalinin², Alexander L. Golovyuk², Elena A. Pudova¹, Zulfiya G. Guvatova¹, Nataliya V. Melnikova¹, Alexey A. Dmitriev¹, George S. Razmakhaev³, Andrey A. Poloznikov³, Galina S. Alekseeva³, Andrey D. Kaprin³, George S. Krasnov¹ and Anna V. Kudryavtseva^{1*}

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Lyudmila Gulyaeva,
Novosibirsk State University, Russia
Aleksandr Khudiakov,
Almazov National Medical Research
Centre, Russia

*Correspondence:

Anastasiya V. Snezhkina
leftger@rambler.ru
Anna V. Kudryavtseva
rhizamoeba@mail.ru

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 07 October 2020

Accepted: 26 November 2020

Published: 18 December 2020

Citation:

Snezhkina AV, Fedorova MS,
Pavlov VS, Kalinin DV, Golovyuk AL,
Pudova EA, Guvatova ZG,
Melnikova NV, Dmitriev AA,
Razmakhaev GS, Poloznikov AA,
Alekseeva GS, Kaprin AD,
Krasnov GS and Kudryavtseva AV
(2020) Mutation Frequency in Main
Susceptibility Genes Among Patients
With Head and Neck Paragangliomas.
Front. Genet. 11:614908.
doi: 10.3389/fgene.2020.614908

¹ Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, ² Vishnevsky Institute of Surgery, Ministry of Health of the Russian Federation, Moscow, Russia, ³ National Medical Research Radiological Center, Ministry of Health of the Russian Federation, Moscow, Russia

Head and neck paragangliomas (HNPGLs) are rare neuroendocrine tumors that have a high degree of heritability and are predominantly associated with mutations in ten genes, such as *SDHx*, *SDHAF2*, *VHL*, *RET*, *NF1*, *TMEM127*, *MAX*, *FH*, *MEN2*, and *SLC25A11*. Elucidating the mutation prevalence is crucial for the development of genetic testing. In this study, we identified pathogenic/likely pathogenic variants in the main susceptibility genes in 102 Russian patients with HNPGLs (82 carotid and 23 vagal paragangliomas) using whole exome sequencing. Pathogenic/likely pathogenic variants were detected in 43% (44/102) of patients. We identified the following variant distribution of the tested genes: *SDHA* (1%), *SDHB* (10%), *SDHC* (5%), *SDHD* (24.5%), and *RET* (5%). *SDHD* variants were observed in the majority of the patients with bilateral/multiple paragangliomas. Thus, among Russian patients with HNPGLs the most frequently mutated gene was *SDHD* followed by *SDHB*, *SDHC*, *RET*, and *SDHA*.

Keywords: head and neck paragangliomas, carotid and vagal paragangliomas, susceptibility genes, *SDHx*, mutation frequency, pathogenic mutations

INTRODUCTION

Head and neck (HN) paragangliomas (PGLs) are rare neuroendocrine tumors of four distinct localizations: carotid, vagal, laryngeal, and middle ear PGLs. Carotid paragangliomas (CPGLs) arise from the carotid glomus at the carotid artery bifurcation and are the most common form of HNPGLs (60%) (El-Naggar et al., 2017). Middle ear and vagal paragangliomas (MEPGLs and VPGLs) are less frequent than CPGLs and account for 29% and 13%, respectively; laryngeal PGLs are very rare (El-Naggar et al., 2017). All the HNPGLs are often characterized by slow growth and non-aggressive behavior, but exhibit metastatic potential. The overall metastatic rate for HNPGLs vary depending on the site of tumor localization: 2% for larynx and middle ear, 4–6% for carotid, and up to 16% for vagal PGLs (Williams, 2017). HNPGLs can also develop as bilateral or multiple tumors and pose significant treatment challenges.

HNPGLs can as familial or sporadic forms (Dahia, 2014). Familial HNPGLs together with pheochromocytomas (PCCs) account for about 40% and are associated with four types of hereditary paraganglioma syndromes (PGL1–5) caused by mutations in the following genes: *SDHD* (PGL1), *SDHAF2* (PGL2), *SDHC* (PGL3), *SDHB* (PGL4), and *SDHA* (PGL5) (Burnichon et al., 2010; Boedeker et al., 2014). Patients with hereditary HNPGLs less frequently harbor germline mutations in *TMEM127* (Bausch et al., 2017), *RET* (Kudryavtseva et al., 2019), *MAX* (Burnichon et al., 2012), *FH* (Castro-Vega et al., 2014), and *SLC25A11* (Buffet et al., 2018). Germline mutations in *VHL*, *NF1*, and *MEN2* have been detected in HNPGLs in association with other tumoral and clinical features (Boedeker et al., 2009).

In this study, we aimed to assess the frequency of variants in the main susceptibility genes for HNPGLs, such as *SDHx*, *SDHAF2*, *VHL*, *RET*, *NF1*, *TMEM127*, *MAX*, *FH*, *MEN2*, and *SLC25A11*, among a representative set of Russian patients with CPGLs and VPGLs.

MATERIALS AND METHODS

Patients

In total, 102 Russian patients with HNPGLs, including 82 patients with CPGLs and 23 with VPGLs, were enrolled at the Vishnevsky Institute of Surgery, Ministry of Health of the Russian Federation. Informed consent was obtained from all patients. This study was approved by the Ethics Committee of Vishnevsky Institute of Surgery with ethics committee approval no. 004-2020, 03.07.2020, and was performed in accordance with the Declaration of Helsinki (1964) (World Medical Association, 2001). The clinicopathologic characteristics of the patients are presented in Table 1.

Exome Library Preparation and Sequencing

DNA was extracted from formaldehyde fixed-paraffin embedded (FFPE) tumor tissues using High Pure FFPE DNA Isolation Kit (Roche, Basel, Switzerland) and quantified with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). DNA quality was assessed by quantitative PCR (qPCR) using QuantumDNA Kit (Evrogen, Moscow, Russia). Exome libraries were prepared from DNA using Rapid Capture Exome Kit (Illumina, San Diego, CA, USA) or TruSeq Exome Library Prep Kit (Illumina) according to the manufacturer's recommendations. High-throughput sequencing of the exome libraries was performed on a NextSeq 500 System (Illumina) in a paired-end mode of 76 × 2 bp. The average coverage for each sample was at least 300X. Sequencing data are available at NCBI Sequence Read Archive (SRA) BioProject PRJNA639937.

Mutation Analysis

Raw sequencing read quality was assessed using FastQC (v. 0.11.9). The reads were trimmed for quality (less than Q20), and adapter sequences were removed using Trimmomatic (v. 0.39) (Bolger et al., 2014). Alignment of reads to the reference

human genome GRCh37.75/hg19 was performed with Burrows-Wheeler Aligner (v. 0.7.17) (Li and Durbin, 2010). To report alignment statistics and determine read duplicates, we applied SAMtools (v. 1.10) (Li et al., 2009) and Picard-tools (v. 2.23.4). Base quality score recalibration was carried out with GATK4 (v. 4.1.2) (McKenna et al., 2010) and dbSNP (common variants 2015-06-05) (Sherry et al., 2001). Variant calling was performed with GATK HaplotypeCaller (McKenna et al., 2010). We excluded false positives using StrandBiasBySample, StrandOddsRatio, and BaseQualityRankSumTest annotations, as well as mis-sequenced single-nucleotide variants in polyN motifs, such as GGGTG > GGGGG, CCCCC > CCCCC, and others. For functional annotation of variants, ANNOVAR (v. 20200316) (Wang et al., 2010) was used. Annotations included allele frequency data [gnomAD (Karczewski et al., 2020), Kaviar (Glusman et al., 2011), and ESP-6500 (<http://evs.gs.washington.edu/EVS/>)], information about reported genomic variations and its association with human pathologies [ClinVar (Landrum et al., 2018), dbSNP, and COSMIC (Tate et al., 2019)], score for the conservation of mutated sites [PhastCons (Siepel et al., 2005) and PhyloP (both PHAST v. 1.5) (Pollard et al., 2010)], localization of variants in protein domains [InterPro (v. 81.0) (Mitchell et al., 2019)], as well as pathogenicity prediction score [SIFT (v. 6.2.1) (Vaser et al., 2016), PolyPhen2 (v. 2.2.2) (Adzhubei et al., 2010), MutationTaster (v. 2013-03-20) (Schwarz et al., 2014), LRT (v. 0.2) (Chun and Fay, 2009), PROVEAN (Choi and Chan, 2015), MetaSVM and MetaLR (Dong et al., 2015), CADD (v. 1.6) (Kircher et al., 2014), and DANN (Quang et al., 2015)].

Sanger Sequencing

To validate the whole exome sequencing data, Sanger sequencing was performed in Evrogen. Primer sequences are available on request.

RESULTS

A representative set of HNPGL samples were collected from 102 Russian patients diagnosed with CPGLs ($n = 82$) and VPGLs ($n = 23$), including 76 patients with single CPGL, 20 patients with single VPGL, and 6 patients with bilateral/multiple PGLs (three of them had both carotid and vagal paragangliomas) (Table 1). These tumor samples were analyzed for the presence of pathogenic/likely pathogenic variants in the main susceptibility genes for HNPGLs: *SDHx*, *SDHAF2*, *VHL*, *RET*, *NF1*, *TMEM127*, *MAX*, *FH*, *MEN2*, and *SLC25A11*. Variants were classified as pathogenic or likely pathogenic according to the annotations in the ClinVar database or by using the criteria of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG-AMP) (Richards et al., 2015).

Pathogenic/likely pathogenic variants were revealed in 44 out of 102 (43%) patients with HNPGLs (Supplementary Table 1). The prevalence of variants was as follows: *SDHA* (1%, 1/102), *SDHB* (10%, 10/102), *SDHC* (5%, 5/102), *SDHD* (24.5%, 25/102), and *RET* (5%, 5/102). Almost all patients with bilateral/multiple paragangliomas (except a patient with no variants in any of

TABLE 1 | Clinicopathologic characteristics of patients with HNPGLs.

Characteristic	Number of patients, n (%)
Total patients	102
Tumor localization	
CPGLs	82 (80%)
VPGLs	23 (22.5%)
Sex	
Male	25 (24.5%)
Female	77 (75.5%)
Age at diagnosis	
≥40	71 (70%)
<40	31 (30%)
Mean	48 (16–79)
Tumor characteristics	
Single	96 (94%)
Bilateral/multiple	6 (6%)
Recurrent	8 (8%)
Metastasis	1 (1%)

the main susceptibility genes tested) demonstrated pathogenic variants in *SDHD* (Table 2). A pathogenic variant NM_003002: c.A305G, p.H102R (chr11: 111959726, rs104894302) was the most frequent *SDHD* mutation detected among the Russian patients. This variant has been found in nine patients with CPGLs and three with VPGLs, including two patients with bilateral/multiple paragangliomas (one patient among them had both tumors subjected to genetic testing). In addition, it is noteworthy that only one variant NM_020975: c.A2372T, p.Y791F (chr10: 43613908, rs77724903) was determined among all *RET*-mutated HNPGLs indicating its high frequency in this population.

Also, we analyzed the frequency of variants in the main susceptibility genes separately for CPGLs and VPGLs. In CPGLs, pathogenic/likely pathogenic variants were detected in 38 of 82 (46%) patients and were distributed as follows: *SDHA* (1%, 1/82), *SDHB* (8.5%, 7/82), *SDHC* (6%, 5/82), *SDHD* (28% 23/82), and *RET* (5%, 4/82) (Supplementary Table 1, Table 2). The majority of patients had pathogenic/likely pathogenic variants in one of these genes, however, in two patients, a likely pathogenic variant NM_020975: c.A2372T, p.Y791F (chr10: 43613908, rs77724903) in *RET* were corepresented with *SDHA* (Pat16) and *SDHB* (Pat142) variants. The same variant in *RET* was identified in two other patients (Pat35 and Pat155) with CPGLs. A likely pathogenic variant NM_003001: c.G149A, p.R50H (chr1: 161298257, rs769177037) in *SDHC* were also detected simultaneously in two patients (Pat102 and Pat152). Nine patients with CPGLs carried one pathogenic variant NM_003002: c.A305G, p.H102R (chr11: 111959726, rs104894302) in *SDHD*, including two patients (Pat1 and Pat5) with bilateral/multiple paragangliomas.

Pathogenic/likely pathogenic variants in *SDHB*, *SDHD*, and *RET* genes were found in 9 out of 23 (39%) patients with VPGLs (Supplementary Table 1, Table 2). The frequency of variants

had the following distribution: *SDHB* (13%, 3/23), *SDHD* (22%, 5/23), and *RET* (4%, 1/23). All the patients with VPGLs carried pathogenic/likely pathogenic variants only in one of these genes. Notably, a likely pathogenic variant NM_020975: c.A2372T, p.Y791F (chr10: 43613908, rs77724903) in *RET* identified in four patients with CPGLs was also found in a patient with VPGL (Pat158). Moreover, three patients with VPGLs harbored the pathogenic variant NM_003002: c.A305G, p.H102R (chr11: 111959726, rs104894302) in the *SDHD* gene detected with high frequency in a set of CPGLs.

Next, we analyzed the age at diagnosis and sex ratio for *SDHB*, *SDHC*, *SDHD*, and *RET* variants within the cohort of patients with HNPGLs (Table 3). Variants in the *SDHC* and *SDHD* genes were diagnosed with approximately equal frequency in males and females taking into account 1:3 male to female ratio among the studied cohort. *SDHB* variants were found third-fold more frequent in females when *RET* variants were detected about two-fold more frequent in males. Variants in *SDHB*, *SDHC*, and *SDHD* were observed in all age groups and were more often detected in patients aged between 19–40 and 41–60 years. *RET* variants were identified in two groups of patients aged 19–40 and 41–60 years. Notably, frequency of variants in all the genes was the lowest in patients aged 61–80 years.

DISCUSSION

In HNPGLs, the occurrence and frequency of mutations in the *SDHx* genes are extensively studied and are of importance for the diagnosis and management of the disease. The prevalence of mutations in other susceptibility genes has been poorly investigated. Notably, most studies have been focused on germline variants. In this work, we cannot establish germline and somatic mutation status for identified variants, since we used the archival collection of tumors for which blood or other normal tissues were not available. We obtained data on the overall frequency of pathogenic/likely pathogenic variants in the main susceptibility genes that allows to better understand molecular basis of the tumor development in Russian patients. A more similar study was performed for 24 Spanish patients with HNPGLs, who were subjected to genetic testing for germline and somatic mutations in the *SDHx* genes (Curras-Freixes et al., 2015). In contrast to our data, the majority of mutations were detected in *SDHB* (33%, 8/24) followed by *SDHD* (21%, 5/24) and *SDHC* (4%, 1/24). Two patients with *SDHB* mutations and one patient with *SDHD* mutation were characterized by metastatic tumors but no *SDHA* variants were detected.

In Russian patients, we revealed the likely pathogenic variant in *RET* in 5% of cases (5/102), including four patients with CPGLs and one patient with VPGL. Activation mutations in the proto-oncogene *RET* lead to the development of an autosomal dominant syndrome called multiple endocrine neoplasia type 2 (MEN2). Up to 50% of patients with MEN2 develop PCCs (Pedulla et al., 2014). HNPGLs have been rarely described in patients with MEN2 syndrome (Boedeker et al., 2009).

TABLE 2 | Types of HNPGLs and number of patients with identified pathogenic/likely pathogenic variants.

Tumor type	Total patients, <i>n</i>	Patients with variants, <i>n</i>	Gene variants, <i>n</i>				
			<i>SDHA</i>	<i>SDHB</i>	<i>SDHC</i>	<i>SDHD</i>	<i>RET</i>
CPGLs	82	38	1	7	5	23	4
VPGLs	23	9	0	3	0	5	1
Bilateral/multiple PGLs	6	5	0	0	0	5	0

TABLE 3 | Age and sex of patients with *SDHB*, *SDHC*, *SDHD*, and *RET* variants.

Gene	Age at presentation			Sex (M:F)
	19–40 yr	41–60 yr	61–80 yr	
<i>SDHB</i>	4	5	1	1:9
<i>SDHC</i>	2	2	1	1:4
<i>SDHD</i>	8	12	5	7:18
<i>RET</i>	4	1	0	2:4

M, males; F, females.

However, several studies have reported *RET* mutations in HNPGLs without any association with MEN2 syndrome. A *RET* mutation was previously detected in a patient with multiple paragangliomas (Ding et al., 2019). Moreover, the likely pathogenic germline variant NM_020975: c.A2372T, p.Y791F (chr10: 43613908, rs77724903) in *RET* was identified in two out of four members of a family with multiple and malignant paragangliomas (Choi Jdo et al., 2014). All four members carried pathogenic *SDHD* mutations. In addition, in this family, the *RET* mutation was observed in the male adult with bilateral carotid body and jugulotympanic paragangliomas and his son with unilateral CPGLs. Here, we also detected this *RET* variant in all *RET*-mutated tumors. Moreover, this variant co-occurred with the pathogenic start-loss variant in *SDHA* and the pathogenic splice site variant in *SDHB*. Collectively, data from our study and previous studies suggest that this *RET* variant can occur both alone and together with *SDHD*, *SDHA*, and *SDHB* pathogenic variants. However, according to the ClinVar database, this *RET* mutation was annotated as pathogenic variant associated with the MEN2 syndrome and its pathogenicity has not been proved by any functional studies. Thus, the role of the *RET* variant in the development of HNPGLs is controversial taking into account that we detected this variant in patients, who were not diagnosed with the MEN2 syndrome.

No pathogenic/likely pathogenic variants in *SDHAF2*, *VHL*, *MAX*, *MEN2*, *NF1*, *FH*, *TMEM127*, and *SLC25A11* were identified among the Russian patients. Variants in all the genes were previously reported only as germline at a very low frequency in HNPGLs (Boedeker et al., 2009; Burnichon et al., 2012; Castro-Vega et al., 2014; Bausch et al., 2017; Buffet et al., 2018; Guha et al., 2019).

In the study, we also analyzed male to female ratio and age groups for identified variants in the *SDHx* and *RET* genes. Currently, clinic-genetic correlations in association with mutations in these genes in HNPGLs have been poorly investigated. Moreover, reported data are related to germline mutations and can be compared only conditionally with our results. Thus, Mario Hermesen with colleagues studied 23 males and 51 females and with HNPGLs and found 8:6 and 5:9 ratios for *SDHB* and *SDHD* mutations, respectively (Hermesen et al., 2010). Taking into account the initial male to female ratio approximately 1:2, we can see that *SDHD* mutations were detected about equally in males and females that is concordance with our data. *SDHB* mutations were more frequently revealed in males compared with females, while we obtained opposite results. In addition, it was observed that *SDHB* and *SDHD* mutations were diagnosed predominantly in patients aged <50 years. We showed that patients aged <61 years carried variants in these genes more frequently. Also, age-related penetrance for *SDHB* and *SDHD* mutations was shown to be increased by 50 years in HNPGLs (Neumann et al., 2004). In our study, we also observed higher number of patients aged 41–60 years with *SDHB* and *SDHD* variants.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found here: <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA639937>.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee of Vishnevsky Institute of Surgery. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

AS and AVK: conceptualization and writing—review and editing. GK, VP, AD, AS, and NM: formal analysis. MF, EP, and ZG: investigation. GR, DK, and AG: resources. AP, GA, and ADK: writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Russian Science Foundation, Grant Number 19-15-00419.

ACKNOWLEDGMENTS

The authors thank the Vishnevsky Institute of Surgery for sample collection and National Medical Research Center of Radiology for assistance in data analysis and discussing. The authors thank the

Center for Precision Genome Editing and Genetic Technologies for Biomedicine, EIMB RAS for providing the computing power and techniques for the data analysis. This work was performed using the equipment of EIMB RAS “Genome” center (http://www.eimb.ru/rus/ckp/ccu_genome_c.php).

REFERENCES

- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248
- Bausch, B., Schiavi, F., Ni, Y., Welander, J., Patocs, A., Ngeow, J., et al. (2017). Clinical Characterization of the pheochromocytoma and paraganglioma susceptibility genes SDHA, TMEM127, MAX, and SDHAF2 for gene-informed prevention. *JAMA Oncol.* 3, 1204–1212. doi: 10.1001/jamaoncol.2017.0223
- Boedeker, C. C., Erlic, Z., Richard, S., Kontny, U., Gimenez-Roqueplo, A. P., Cascon, A., et al. (2009). Head and neck paragangliomas in von Hippel-Lindau disease and multiple endocrine neoplasia type 2. *J. Clin. Endocrinol. Metab.* 94, 1938–1944. doi: 10.1210/jc.2009-0354
- Boedeker, C. C., Hensen, E. F., Neumann, H. P., Maier, W., Van Nederveen, F. H., Suarez, C., et al. (2014). Genetics of hereditary head and neck paragangliomas. *Head Neck* 36, 907–916. doi: 10.1002/hed.23436
- Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170
- Buffet, A., Morin, A., Castro-Vega, L. J., Habarou, F., Lussey-Lepoutre, C., Letouze, E., et al. (2018). Germline mutations in the mitochondrial 2-oxoglutarate/malate carrier SLC25A11 gene confer a predisposition to metastatic paragangliomas. *Cancer Res.* 78, 1914–1922. doi: 10.1158/0008-5472.CAN-17-2463
- Burnichon, N., Briere, J. J., Libe, R., Vescovo, L., Riviere, J., Tissier, F., et al. (2010). SDHA is a tumor suppressor gene causing paraganglioma. *Hum. Mol. Genet.* 19, 3011–3020. doi: 10.1093/hmg/ddq206
- Burnichon, N., Cascon, A., Schiavi, F., Morales, N. P., Comino-Mendez, I., Abermil, N., et al. (2012). MAX mutations cause hereditary and sporadic pheochromocytoma and paraganglioma. *Clin. Cancer Res.* 18, 2828–2837. doi: 10.1158/1078-0432.CCR-12-0160
- Castro-Vega, L. J., Buffet, A., De Cubas, A. A., Cascon, A., Menara, M., Khalifa, E., et al. (2014). Germline mutations in FH confer predisposition to malignant pheochromocytomas and paragangliomas. *Hum. Mol. Genet.* 23, 2440–2446. doi: 10.1093/hmg/ddt639
- Choi Jdo, W., Tucker, K. M., Lee, T. T., and Chong, G. C. (2014). Hereditary paraganglioma-pheochromocytoma syndromes associated with SDHD and RET mutations. *Head Neck* 36, E99–E102. doi: 10.1002/hed.23598
- Choi, Y., and Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 31, 2745–2747. doi: 10.1093/bioinformatics/btv195
- Chun, S., and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Res.* 19, 1553–1561. doi: 10.1101/gr.092619.109
- Curras-Freixes, M., Inglada-Perez, L., Mancikova, V., Montero-Conde, C., Leton, R., Comino-Mendez, I., et al. (2015). Recommendations for somatic and germline genetic testing of single pheochromocytoma and paraganglioma based on findings from a series of 329 patients. *J. Med. Genet.* 52, 647–656. doi: 10.1136/jmedgenet-2015-103218
- Dahia, P. L. (2014). Pheochromocytoma and paraganglioma pathogenesis: learning from genetic heterogeneity. *Nat. Rev. Cancer* 14, 108–119. doi: 10.1038/nrc3648
- Ding, Y., Feng, Y., Wells, M., Huang, Z., and Chen, X. (2019). SDHx gene detection and clinical Phenotypic analysis of multiple paraganglioma in the head and neck. *Laryngoscope* 129, E67–E71. doi: 10.1002/lary.27509
- Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., et al. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* 24, 2125–2137. doi: 10.1093/hmg/ddu733
- El-Naggar, A. K., Chan, J. K. C., Grandis, J. R., Takata, T., and Slootweg, P. J. (2017). *WHO Classification of Head and Neck Tumours*. Lyon: International Agency for Research on Cancer.
- Glusman, G., Caballero, J., Mauldin, D. E., Hood, L., and Roach, J. C. (2011). Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* 27, 3216–3217. doi: 10.1093/bioinformatics/btr540
- Guha, A., Musil, Z., Vicha, A., Zelinka, T., Pacak, K., Astl, J., et al. (2019). A systematic review on the genetic analysis of paragangliomas: primarily focused on head and neck paragangliomas. *Neoplasia* 66, 671–680. doi: 10.4149/neo_2018_181208N933
- Hermesen, M. A., Sevilla, M. A., Llorente, J. L., Weiss, M. M., Grimbergen, A., Allonca, E., et al. (2010). Relevance of germline mutation screening in both familial and sporadic head and neck paraganglioma for early diagnosis and clinical management. *Cell Oncol.* 32, 275–283. doi: 10.3233/CLO-2009-0498
- Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7
- Kircher, M., Witten, D. M., Jain, P., O’roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310–315. doi: 10.1038/ng.2892
- Kudryavtseva, A. V., Lukyanova, E. N., Kalinin, D. V., Zaretsky, A. R., Pokrovsky, A. V., Golovyuk, A. L., et al. (2019). Mutational load in carotid body tumor. *BMC Med. Genomics* 12:39. doi: 10.1186/s12920-019-0483-x
- Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* 46, D1062–D1067. doi: 10.1093/nar/gkx1153
- Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110
- Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., et al. (2019). InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* 47, D351–D360. doi: 10.1093/nar/gky1100
- Neumann, H. P., Pawlu, C., Peczkowska, M., Bausch, B., McWhinney, S. R., Muresan, M., et al. (2004). Distinct clinical features of paraganglioma syndromes associated with SDHB and SDHD gene mutations. *JAMA* 292, 943–951. doi: 10.1001/jama.292.8.943
- Pedulla, G., Crocetti, D., Paliotta, A., Tarallo, M. R., De Gori, A., Cavallaro, G., et al. (2014). Surgical treatment of pheochromocytoma in MEN 2. *Ann. Ital. Chir.* 85, 443–447.
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121. doi: 10.1101/gr.097857.109
- Quang, D., Chen, Y., and Xie, X. (2015). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763. doi: 10.1093/bioinformatics/btu703

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.614908/full#supplementary-material>

- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–424. doi: 10.1038/gim.2015.30
- Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11, 361–362. doi: 10.1038/nmeth.2890
- Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308
- Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005
- Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015
- Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., and Ng, P. C. (2016). SIFT missense predictions for genomes. *Nat. Protoc.* 11, 1–9. doi: 10.1038/nprot.2015.123
- Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38:e164. doi: 10.1093/nar/gkq603
- Williams, M. D. (2017). Paragangliomas of the head and neck: an overview from diagnosis to genetics. *Head Neck Pathol.* 11, 278–287. doi: 10.1007/s12105-017-0803-4
- World Medical Association (2001). World medical association declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bulletin of the World Health Organ.* 79, 373–374.
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Snezhkina, Fedorova, Pavlov, Kalinin, Golovyuk, Pudova, Guvatova, Melnikova, Dmitriev, Razmakhaev, Poloznikov, Alekseeva, Kaprin, Krasnov and Kudryavtseva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Automated Extraction of Information From Texts of Scientific Publications: Insights Into HIV Treatment Strategies

Nadezhda Biziukova¹, Olga Tarasova^{1*}, Sergey Ivanov^{1,2} and Vladimir Poroikov¹

¹ Laboratory of Structure-Function Based Drug Design, Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russia, ² Department of Bioinformatics, Faculty of Biomedicine, Pirogov Russian National Research Medical University, Moscow, Russia

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Vladimir Aleksandrovich Ivanisenko,
Russian Academy of Sciences, Russia
Anastasia Khandazhinskaya,
Engelhardt Institute of Molecular
Biology (RAS), Russia

*Correspondence:

Olga Tarasova
olga.a.tarasova@gmail.com

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 18 October 2020

Accepted: 26 November 2020

Published: 22 December 2020

Citation:

Biziukova N, Tarasova O, Ivanov S and
Poroikov V (2020) Automated
Extraction of Information From Texts of
Scientific Publications: Insights Into
HIV Treatment Strategies.
Front. Genet. 11:618862.
doi: 10.3389/fgene.2020.618862

Text analysis can help to identify named entities (NEs) of small molecules, proteins, and genes. Such data are very important for the analysis of molecular mechanisms of disease progression and development of new strategies for the treatment of various diseases and pathological conditions. The texts of publications represent a primary source of information, which is especially important to collect the data of the highest quality due to the immediate obtaining information, in comparison with databases. In our study, we aimed at the development and testing of an approach to the named entity recognition in the abstracts of publications. More specifically, we have developed and tested an algorithm based on the conditional random fields, which provides recognition of NEs of (i) genes and proteins and (ii) chemicals. Careful selection of abstracts strictly related to the subject of interest leads to the possibility of extracting the NEs strongly associated with the subject. To test the applicability of our approach, we have applied it for the extraction of (i) potential HIV inhibitors and (ii) a set of proteins and genes potentially responsible for viremic control in HIV-positive patients. The computational experiments performed provide the estimations of evaluating the accuracy of recognition of chemical NEs and proteins (genes). The precision of the chemical NEs recognition is over 0.91; recall is 0.86, and the F1-score (harmonic mean of precision and recall) is 0.89; the precision of recognition of proteins and genes names is over 0.86; recall is 0.83; while F1-score is above 0.85. Evaluation of the algorithm on two case studies related to HIV treatment confirms our suggestion about the possibility of extracting the NEs strongly relevant to (i) HIV inhibitors and (ii) a group of patients i.e., the group of HIV-positive individuals with an ability to maintain an undetectable HIV-1 viral load overtime in the absence of antiretroviral therapy. Analysis of the results obtained provides insights into the function of proteins that can be responsible for viremic control. Our study demonstrated the applicability of the developed approach for the extraction of useful data on HIV treatment.

Keywords: text mining, data mining, named entity recognition, NER, virus-host interactions, HIV, viremic control

INTRODUCTION

Scientific publications represent the main source of knowledge for researchers in different fields of biology and medicine. Besides, the more pressing the problem for humanity is, the more articles devoted to this problem can be found in the repositories of scientific publications. The extraction of records from scientific publications provides the opportunity to analyze the information derived from primary sources; therefore, such an approach helps to obtain the most contemporary information (Cash, 2004; Tarasova et al., 2015, 2019; Saik et al., 2016). Currently, text-mining technologies aimed at rapid automated extraction of specific information are under rigorous development.

Analysis of interactions between named entities (NEs) representing proteins, genes, and chemical compounds can help investigate the particular molecular mechanisms of disease progression, the effect of drugs, and reveal the drug-drug interactions important for efficacy of therapy (Chen et al., 2014; Tannenbaum and Sheehan, 2014; Lim et al., 2016; Szklarczyk et al., 2019).

Identification of associations between NEs in the texts of scientific publications includes two steps: (i) extraction of named entities from the texts, and (ii) recognition of associations. This is the focus of the Named Entity Recognition (NER) methods. There are two main groups of approaches used for NER: (i) based on rules and dictionaries and (ii) based on machine learning methods. The main disadvantage of rule and dictionary-based algorithms is the inability to extract information about entities not included in dictionaries. Another drawback is the requirements for the allocation of memory for storing dictionaries.

Machine learning methods require sets of texts, in which the names of proteins, genes, chemical compounds, and so on are labeled by an expert or a group of experts. Then, using such texts as a training set, it is possible to adjust the algorithm to recognize NEs in a large number of articles. Finally, it is possible to identify relationships between the NEs extracted. Machine learning methods have advantages over dictionary-based methods because they provide the recognition of new NEs not included in dictionaries and, therefore, are the best option for the careful extraction of information. At present, many novel text corpora are constantly developing for the purposes of the scientific community and provide the possibility to extract information about a variety of NEs: genes, proteins and chemicals names, symptoms and syndromes of the diseases, side effects and toxicity of drugs, revealed during the clinical trials or as a result of medical studies, cases studies, etc.

There are methods aimed at NER that have been developing during the last years (Kaewphan et al., 2018; Korvigo et al., 2018; Hemati and Mehler, 2019; Hong and Lee, 2020; Huang et al., 2020; Kilicoglu et al., 2020). Most of them are based on algorithms for NER related either to chemicals or biological objects. In this study, we aim to develop and test an algorithm for the extraction of named entities of genes/proteins and chemical compounds and identify associations between them. Our method of NER is based on the conditional random fields and uses a

set of originally developed word features that allow for context consideration. Thus, we suggest selecting a set of publications strictly relevant to the subject and extracting a set of chemical NEs, proteins, and genes to derive the NEs associated with one another. After that, their functions in a particular molecular mechanism of the disease can be analyzed.

Human immunodeficiency virus (HIV) still remains one of the challenges for humanity (Rojas-Celis et al., 2019; Tarasova et al., 2020). The number of new HIV cases per year reached 1.7 million (WHO, HIV incidence)¹ while is, Number of People (All Ages) Living With HIV² is around 38 million (WHO). Antiretroviral therapy (ART) helps to reduce viral load and disease progression, but antiretroviral medicine should be taken by a patient for term of life. The risk of HIV drug resistance and side effects of antiretroviral medicines decrease the effectiveness of ART (Iyidogan and Anderson, 2014; Tarasova and Poroikov, 2018). At the same time, an effective HIV vaccine does not exist (Ventura, 2020).

Taking into account the importance of the problem, we consider two main approaches for the HIV/AIDS treatment as case studies: (1) the usage of antiretroviral drugs and (2) studies of the ways of HIV/AIDS development in different groups of patients and attempts to affect the key proteins of the pathways providing a long period of disease progression. The mechanisms identified can be used for the development of novel strategies of HIV treatment and vaccine development. We have validated and tested our algorithm on the tasks of identification of (i) chemicals that can be considered as HIV inhibitors and (ii) groups of proteins that may be important for the different velocity of HIV/AIDS. To reach these purposes, we used the abstracts of publications relevant to HIV/AIDS treatment for case studies of HIV inhibition and HIV viremic control. More specifically, the algorithm developed is aimed at extracting (a) the names of HIV reverse transcriptase (RT) inhibitors and (b) the protein (gene) names described in articles relevant to the studies of HIV elite controllers, i.e., the group of HIV-positive individuals with an ability to maintain an undetectable HIV-1 viral load overtime in the absence of antiretroviral therapy.

MATERIALS AND METHODS

The extraction of the NE names includes several stages. First, we collected text corpora and made their preprocessing. Second, we developed an algorithm of NER and its parameter optimization. Third, we carried out validation and testing of an algorithm and analysis of the information obtained.

To extract NEs, we used annotated text corpora and applied an algorithm based on conditional random fields (CRF). Text annotation in the corpus implies an indication of NE position inside the text.

¹WHO HIV Incidence. Available online at: <https://apps.who.int/gho/data/node.main.HIVINCIDENCE?lang=en> (accessed October 15, 2020).

²Number of People (All Ages) Living With HIV. Available online at: <https://apps.who.int/gho/data/view.main.22100WHO?lang=en> (accessed October 15, 2020).

Text Corpora

We used the CHEMDNER³ (Krallinger et al., 2015) and ChemProt⁴ as annotated corpora. Both corpora consist of freely available abstracts of articles.

CHEMDNER consists of three sets (training, evaluation, and development sets) and includes 10,000 abstracts. It has been developed for the purpose of chemical NER (Krallinger et al., 2015). Annotations include the position of NE in the text and also the NE type: ABBREVIATION (as ATP), FAMILY (as steroid hormones), FORMULA [as S or C (sulfur or carbon)], IDENTIFIER (as GRN-529), MULTIPLE (as nucleoside tri- and di-phosphates), SYSTEMATIC (as sphingosine-1-phosphate), TRIVIAL (as progesterone).

ChemProt also consists of three sets—training, development, and test—and includes 2,482 abstracts in total. It also includes annotations of genes, proteins, and chemical compounds. ChemProt includes the following types of NEs: (i) GENE-Y is for proteins/genes that can be normalized or associated with a biological database identifier, (ii) GENE-N is for proteins/genes that cannot be normalized or associated with a biological database identifier and (iii) CHEMICAL is for chemical compounds⁴. The proteins and genes are not considered in ChemProt separately, but search in the databases containing data on genes and proteins can help identify protein names and names of genes.

Text Preprocessing

Algorithms for entity recognition in texts require tokenization. It is the process of splitting the whole text into elementary text units. As a result of tokenization, the text is presented as a set of tokens. Words, symbols, numbers, can be used as tokens. Thus, we used this method for all symbols and spaces implemented in the “wordpunct_tokenize” function of the NLTK Python library.

Tokens can be divided into groups: those that belong and do not belong to NEs. If an NE originally consists of two words (for example, reverse transcriptase) or includes any symbols (for example, sphingosine-1-phosphate), then after tokenization it will be presented by a list or array of tokens rather than a simple string. We used the labeling system SOBIE (Rocktäschel et al., 2012; Batista-Navarro et al., 2015; Dai et al., 2015; Leaman et al., 2015) to indicate the position of the term associated with a particular token.

SOBIE is an abbreviation for tags: “S” (Single)—if a token belongs to NE and NE consists of one token, “O” (Out)—if a word doesn’t belong to NE, “B” (Begin) is a label of the first word of composite NE if NE consists of two or more words, “E” (End)—is a label of the last word of composite NE, if NE consists of two or more words and I (Inside)—is a label of words belong to NE that are between “B” and “E” if NE consists of three or more words. The example of labeling text by SOBIE is presented in **Supplementary Figure 1**.

³BioCreative IV, CHEMDNER Corpus. Available online at: <https://biocreative.bioinformatics.udel.edu/resources/biocreative-iv/chemdner-corpus/> (accessed August 21, 2020).

⁴BioCreative VI, ChemProt Corpus. Available online at: <https://biocreative.bioinformatics.udel.edu/news/corpora/chemprot-corpus-biocreative-vi> (accessed September 4, 2020).

Tokenized corpora texts with a set of SOBIE labels were placed to the database managed by PostgreSQL DBMS. The schema of the database is provided in **Supplementary Figure 2**.

We compiled a features set for each word to train a model. For NER, it is essential to take into account the context, which means the words should follow one another in order without mixing. In our model, each word *W* is characterized by its features along with the features of one word before and after *W*. Each token was described with a triple set of features (**Table 1**) - the features of a particular token and the features of two tokens: one before and one after the considered one.

We have compiled a list of non-specific terms and used the belonging of a token to one of them as a feature. Non-specific terms are general words that can indicate the presence of the term in the proximity to NEs of a protein (gene) or a chemical compound in the text. For example, such terms may include the words “inhibit,” “chemical” for the names of chemical compounds and “target,” “genes” for the names of proteins and genes. The list of all these non-specific terms for NEs of proteins/genes and chemical compounds is presented in **Supplementary Materials**. We used marks “C” for non-specific terms used for chemical compounds and mark “G” – for proteins and genes. If a non-specific term was used in both lists, it received both labels: “C” and “G.”

All the features were obtained using scripts prepared using Python 3.7. The features were represented as dictionaries to be able to access each feature individually by its keyword. Finally, the data were represented in the format that is shown in **Figure 1**.

Algorithm Realization

As the next step, we built a model. For our approach, we used an algorithm named CRF in realization using Python 3.7 and SciKitLearn library. This library provides a lot of algorithms using machine learning, different metrics, etc. The CRF algorithm allows us to take the context of a phrase into account. We suggest that the set of features developed can help to improve the recognition of context near the NEs. We used the hyperparameter optimization function, which is included in the SciKitLearn to achieve the highest accuracy of the model.

We built several models. The first model (i) was built to provide the recognition labels of chemical entity mentions. The second model (ii) aimed at recognizing names of proteins and genes. The third model (iii) allowed recognition of types (ABBREVIATION, FORMULA, IDENTIFIER etc.) of chemical entity mentions. Based on the models (i–iii), we built the fourth model (iv) to extract chemical compounds and proteins/genes. This model combines algorithm for recognizing the names of chemical compounds and proteins and identifying the types of chemical compounds. Thus, the models (i) and (iii) were built based on CHEMDNER as a training corpus. The model (ii) was built using ChemProt. And the model (iv) was built using both CHEMDNER and ChemProt. Initially, models were tested using 5-fold cross-validation (Stone, 1974).

We used precision, recall, and F1-score to assess the quality of model recognition.

Precision is the proportion of positive identifications that was actually correct. It is calculated as the ratio of the number

TABLE 1 | The set of features used for CRF.

Feature	Type	Meaning
Word	String	Word string
Lower	String	Lowercase word string
Isupper	Boolean	If word is uppercase
Istitle	Boolean	If word's first symbol is uppercase
Isdigit	Boolean	If word is digit
Hasdigits	Boolean	If word has digits
Isnonspecific	Boolean	If word belongs to non-specific terms list
Isstopword	Boolean	If word belongs to stop-words list
Hassymbols	Boolean	If word has symbols
Word[n-3:n]	String	Three last chars of word string
Word[n-2:n]	String	Two last chars of word string
Firstchar	String	Word's first char
Length	Integer	Number of chars in word string
Postag	String	Word's part of speech tag

Features which do not belong to the standard set are highlighted by gray color.

of true positive identification to the sum of true positive and false-positive identifications (1).

$$\text{Precision} = \frac{TP}{TP + FP}; \quad (1)$$

Recall is the proportion of actual positive NE mentions that were identified correctly. It is calculated as the ratio of the number of true positive identifications to the sum of true positive and false negative decisions.

$$\text{Recall} = \frac{TP}{TP + FN}; \quad (2)$$

F1-score is the harmonic mean of precision and recall.

$$F1 = 2 \frac{\text{precision} \bullet \text{recall}}{\text{precision} + \text{recall}}; \quad (3)$$

The test set was formed based on the idea that if an article is strictly relevant to reverse transcriptase inhibition and our algorithm is able to recognize the NE of a protein and a name of a chemical compound or drug then there is a high probability that the chemical named entity identified is the name of HIV reverse transcriptase inhibitor.

We used the set of 148 publications abstracts collected from NCBI PubMed. We used the workflow developed earlier (Tarasova et al., 2019). In this workflow we were focused on the publications that included the description of HIV inhibitors and included the details of biological experiments used for their testing. Using the Python script, we automatically extracted over 15,000 abstracts of articles from PubMed using the query "HIV AND reverse transcriptase AND inhibitors." Then, using the Lingpipe 4.1.2 tool (Carpenter, 2007), we selected a set of over 1,000 abstracts strictly relevant to the development, synthesis, and testing of HIV RT inhibitors. We used Lingpipe 4.1.2 since it provides the possibility of building the models

of text classification into the classes according to the content, and the selection of the texts strictly relevant to the particular subject. Earlier we used Lingpipe 4.1.2 to perform selection of the abstracts of publications, which included the description of HIV-1 reverse transcriptase inhibitors including the details of their biological testing (Tarasova et al., 2019) with mean accuracy over 0.83. We assume that the data on the biological experiment details in the publication text can help to confirm its relevance to the inhibition of HIV-1 RT.

After that, we manually selected 148 of them, which consisted the description of HIV-1 reverse transcriptase inhibitors and the details of their biological testing. This set of abstracts was used to evaluate the accuracy of extracting the names of reverse transcriptase inhibitors.

The texts of abstracts can include the NEs of such chemical compounds as ATP, various ions, DNA, etc. If we want to extract the small drug-like compounds mainly, we would like to filter out ions and biological molecules remaining only drug-like compounds with molecular weight ranged from 300 to 700 (Da). So, to filter them out, we have additionally introduced filters for the classes of chemical compounds obtained based on the results of recognition using CRF. The classes FORMULA and FAMILY were excluded because the class FORMULA is mainly represented by ions, and the class FAMILY contains groups of chemical compounds that do not include the specific names of small molecule inhibitors of HIV RT.

Once we have developed an algorithm providing recognition of proteins (genes) names, we also tested the applicability of our algorithm to the extraction of proteins and genes responsible for the slow disease progressions of HIV-positive patients. To perform this analysis, we have collected a set of abstracts of publications from NCBI PubMed and NCBI PubMed Central (PMC) databases. We collected the abstracts strictly relevant to (1) the HIV elite controllers (ECs), a group of patients who do not progress into HIV/AIDS for years in the absence of antiretroviral therapy, and (2) the whole cohort of HIV-positive patients. The first set related to HIV elite controllers was obtained based on the query "(HIV[Title/Abstract] OR "human immunodeficiency virus"[Title/Abstract] OR "HIV"[Mesh] OR AIDS[Title/Abstract] OR "acquired immunodeficiency syndrome"[Title/Abstract] OR "Acquired Immunodeficiency Syndrome"[Mesh]) AND ("elite control*"[Title/Abstract] OR "Elite suppress*"[Title/Abstract])." We obtained over 840 abstracts strictly relevant to the HIV elite controllers. The second set of abstracts was collected using the query "HIV positive AND HIV/AIDS" and included over 30 thousand of abstracts. We excluded from the group (1) the abstracts belonging to the group (2) because the group (2) may include HIV-positives ECs and we are interested in the differences between the protein profiles of groups (1, ECs) and (2, HIV-positives excluding ECs) mainly. The abstracts of groups (1) and (2) were processed using the NER algorithm developed, and names of proteins (genes) were extracted from them. Then, we excluded the proteins found in the abstracts of the group (2) from a list of proteins obtained for the group (1). We suggest that it provides the opportunity to compile a list of proteins responsible for the slow HIV/AIDS disease progression.

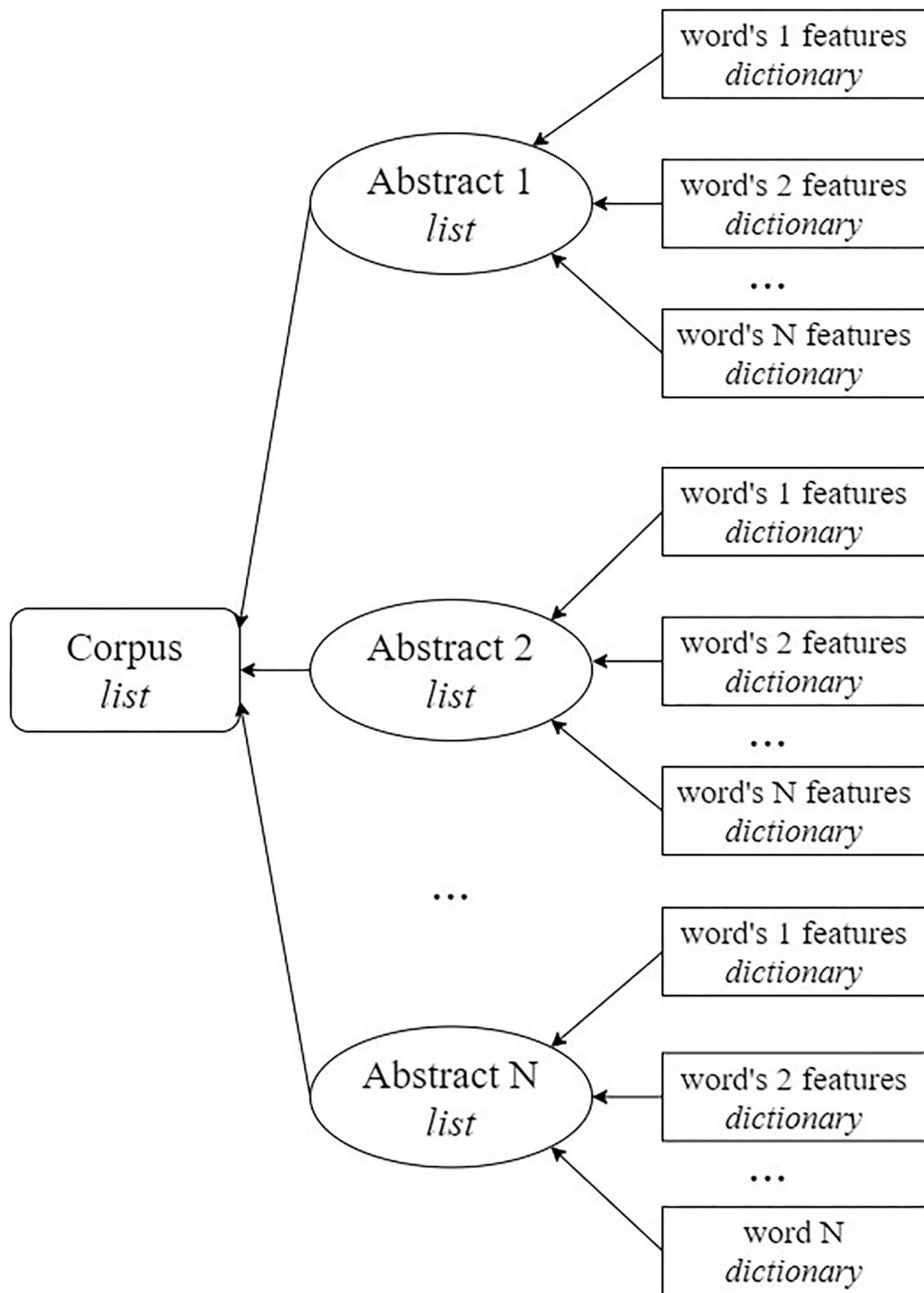


FIGURE 1 | Scheme for the organization of data that are input for the models.

The results of our computational experiments are described and discussed below.

RESULTS

Extraction of Chemical Named Entities

In our study, we performed several computational experiments using CHEMDNER and ChemProt corpora and the texts related to (i) the inhibition of HIV-1 reverse transcriptase and (ii) HIV ECs. Earlier (Tarasova et al., 2019), we collected a set of papers consisted mainly of those relevant to the development and testing of HIV reverse transcriptase (RT) inhibitors, the “HIV-RT-inhibitors corpus.” In the current study, we curated this corpus carefully and enlarged it with publications strictly relevant to the inhibition of HIV RT. The chemical compounds extracted were supposed to be inhibitors of reverse transcriptase based on a specific selection of texts for the test set described below.

We built the models for chemical and protein/gene NER based on CHEMDNER and ChemProt corpora, respectively, and calculated their accuracy using five-fold cross-validation. We performed several computational experiments for predicting SOBIE labels of belonging a token to a chemical and protein/gene named entity (NE) as well as prediction of certain types of chemical entity mention (ABBREVIATION, FORMULA, IDENTIFIER etc.). We evaluated the best way of NER using the features of text developed. The detailed description of text corpora and our computational experiments are provided in the section Materials and Methods.

First, we built a model to predict SOBIE—the labels for the parts of chemical NE: S - Single—for NE that contains one token; B, E—Begin, End—as the labels for the first and the last token of NE, respectively, if NE contains at least two tokens; I—Inside—as a label for the tokens that located between B and E, if NE includes three or more words; O—Out—as a label for the words that does not belong to NE. The results of five-fold cross-validation for annotated corpora CHEMDNER and ChemProt are represented in **Table 2**.

Table 2 displays that the recognition of chemical entity mentions occurs with reasonable accuracy (F1-score is 0.89). We carried out an experiment combining the CHEMDNER and ChemProt to recognize the names of chemical compounds. The volume of training set increased by more than two thousand articles (more than 25%). However, this did not lead to a significant increase in accuracy.

We assumed that the next task might require filtering the names of chemical compounds by their type. Thus, for example, the texts may contain the names of ions (Ca^{2+} , Mn^{2+}), which can be filtered out in case if we are focused on the small drug-like compounds only with molecular weight range from 300 to 700 (Da). Based on this conclusion, we built a model identifying the types of recognized names of chemical compounds. Accuracy was assessed with five-fold cross-validation and is presented in **Table 3**.

From the values of precision, recall, and F1-score displayed in **Table 3**, one can conclude that filtering by the types of NEs of

TABLE 2 | Precision, recall, and F1-score for model that predicts SOBIE for chemical entity mentions.

	Precision	Recall	F1-score
S	0.906	0.8388	0.871
O	0.9908	0.996	0.9934
B	0.861	0.7898	0.8238
I	0.9216	0.8782	0.8994
E	0.8764	0.8032	0.8378
Avg	0.9112	0.8612	0.8851

The average values of precision, recall and F1-score are given in bold.

TABLE 3 | Precision, recall, and F1-score for predicting types of chemical entity mentions.

	Precision	Recall	F1-score
ABBREVIATION	0.9092	0.8946	0.9016
FAMILY	0.8700	0.8074	0.8368
FORMULA	0.9176	0.9030	0.9098
IDENTIFIER	0.8574	0.8954	0.8748
MULTIPLE	0.8334	0.8288	0.8226
SYSTEMATIC	0.9008	0.9394	0.9196
TRIVIAL	0.9200	0.9198	0.9198
Avg	0.8869	0.8841	0.8836

The types of chemical compounds are explained in Supplementary Material (Supplementary Table 1). The average values of precision, recall and F1-score are given in bold.

chemical compound can be used for the selection of the particular type of chemical NE of interest.

The algorithms obtained for recognition of the names of chemical compounds and their types and names of proteins (genes) were then combined and applied on a test set of texts devoted to inhibition of HIV reverse transcriptase. The prediction of genes and proteins names was obtained using the SOBIE labels only. The principle of the combined algorithm was the sequential recognition of the names of chemical compounds, their types, and then the names of proteins.

As a result, the names of proteins and chemical compounds were obtained. The rule for extraction of NEs of HIV RT inhibitors was the presence in the abstract of at least one recognized protein name and one named entity of chemical compound. So, in addition to the names of potential reverse transcriptase inhibitors, it is possible to extract NEs of chemical compounds that can interact with two HIV proteins, for instance, reverse transcriptase and protease.

Evaluation of the results was carried out so that if the recognized chemical compound is not an actual inhibitor of reverse transcriptase, then the number of false positives (FP) increased by one. If an inhibitor was encountered in the text and was not recognized, then the number of false negatives (FN) also increased by one.

Also, we evaluated the accuracy of the recognition of all chemical compounds in the text. The results are shown in **Table 4**.

TABLE 4 | Precision, recall and F1-score obtained when testing the algorithm on a sample of texts devoted to the inhibition of HIV reverse transcriptase.

	Precision	Recall	F1-score
Chemical compounds	0.83	0.94	0.88
Reverse transcriptase inhibitors	0.80	0.97	0.88

Based on the value of the recall metric, we can conclude that, in general, we managed to extract almost all names of reverse transcriptase inhibitors and chemical compounds from the test set. However, there are wrong recognized inhibitors and chemical compounds too. This is in agreement with the overall value of accuracy of NEs recognition in the text according to the results of 5-fold cross-validation. As we have mentioned earlier, we applied filters on classes of chemical compounds to reduce the number of false-positive results. This step allowed us to increase the precision of value to 0.85. We built automatic queries to PubChem database⁵. For all recognized chemical named entities we obtained PubChem identifiers, if PubChem identifier was found.

We also tried to extract the names of chemical compounds from a set of texts dedicated to elite HIV/AIDS controllers. We assumed that if the texts contain proteins (genes) responsible for the non-progression of HIV/AIDS, then they can also indicate chemical compounds that slow down the progression of HIV/AIDS by influencing the protein (gene). Thus, using model (i), we recognized the names of chemical compounds in texts related to the non-progression of HIV/AIDS and then manually checked the presence of chemical compounds that slow down the progression of HIV/AIDS. Unfortunately, there are few texts of the set in which such chemical compounds are mentioned. We provide some examples of extracted chemical compounds in the Discussion section.

Extraction of the Protein and Gene Names

As the next step, we aimed at testing the algorithm for the extraction of protein and gene names from the texts. We built CRF using SOBIE labels similar to the algorithm for the extraction of chemical named entities. The results of protein (gene) recognition based on ChemProt are provided in Table 5.

Despite the fact that the recognition of the named entities of genes and proteins is carried out with a slightly lower accuracy than the recognition of the NEs of chemical compounds, the prediction accuracy still remains reasonable.

We tested the algorithm developed for evaluation a performance of extraction of proteins responsible for HIV/AIDS control and non-progression. To obtain the results, we applied the developed algorithm and the model (ii) to the set of papers, relevant to (1) ECs and (2) the whole cohort of HIV-positive patients retrieved from NCBI PubMed and NCBI PMC databases. The number of proteins and genes extracted from the texts of group (1) and group (2) abstracts

TABLE 5 | Precision, recall, and F1-score for model that predicts SOBIE for names of proteins/genes.

	Precision	Recall	F1-score
S	0.8616	0.828	0.8442
O	0.9732	0.983	0.978
B	0.8314	0.7764	0.803
I	0.8444	0.8078	0.8254
E	0.834	0.7792	0.806
Avg	0.86892	0.83488	0.85132

The average values of precision, recall and F1-score are given in bold.

TABLE 6 | Numbers of proteins (genes) names associated with different velocity of HIV/AIDS progression.

	The number of proteins (genes) unique names retrieved
ECs (group 1 of papers abstracts)	478
HIV-positive (group 2 of papers abstracts)	1,443
Overlap between groups 1 and 2	75
Proteins specific for group 1 only	403
Proteins specific for group 2 only	1,368

is given in Table 6. The full list of proteins extracted for each of the groups represented in Table 6 is given in the **Supplementary Materials**.

As Table 6 displays, there are different protein profiles of the names of genes and proteins extracted from the set of abstracts relevant to the ECs (group 1) and the overall group of HIV-positive patients. For further analysis, we automatically obtained the synonyms of protein (gene) names and UniProt⁶ identifiers for each name of gene or protein extracted from group 1 of abstracts (i.e., articles relevant to ECs). We also identified the main functions of the proteins of this group. The interpretation of our results is given below in the Discussion section.

DISCUSSION

Comparison of the NER Algorithm With Earlier Developed Approaches

We have compared the models obtained for NER of chemical compounds and proteins/genes with those developed earlier by other authors. Earlier NER approaches reached on average F1-score between 77.70 and 88.06% before post-processing (Campos et al., 2015; Khabsa and Giles, 2015; Xu et al., 2015; Korvigo et al., 2018), and for recognizing proteins/genes. Taking into account the results provided, one can conclude that the accuracy of NER for our method is comparable with that of some methods developed earlier. We suggested modifying the text features that lead to an increase in the recognition accuracy: in particular, we

⁵PubChem. Available online at: <https://pubchem.ncbi.nlm.nih.gov/> (accessed November 20, 2020).

⁶UniProt. Available online at: <https://www.uniprot.org/> (accessed November 20, 2020).

expanded the list of non-specific terms used for the recognition of genes and proteins.

We tried to merge two corpora, CHEMDNER and ChemProt, to improve the accuracy of chemical compounds recognition. Despite the assumption that if we increase the number of examples to train the model, the accuracy may increase, this did not happen in our case. A detailed comparison of the precision, recall, and F1-score values for the recognition of chemical NE (i) based on CHEMDNER and (ii) merged CHEMDNER and ChemProt corpora and (ii) recognition of proteins and genes NE based on ChemProt corpus is provided in **Supplementary Figure 1**.

The lower recognition accuracy of chemical compounds based on the combined corpus may be due to ChemProt was primarily aimed at finding relationships between proteins and genes. Therefore, chemical NEs in ChemProt are less common in the texts compared to CHEMDNER.

The algorithms obtained for recognizing the names of chemical compounds and proteins (genes) were examined on the test sets of abstracts related to (i) the inhibition of HIV reverse transcriptase (ii) the identification of proteins associated with HIV control and non-progression.

Extraction of Chemicals Names That Can Be Considered as Potential Medicines for HIV Treatment: A Case Study for HIV Reverse Transcriptase Inhibitors

From the texts relevant to the inhibition of HIV reverse transcriptase, we were able to extract inhibitors that actually exist and are currently used for the therapy. This allows us to conclude that using our method we can extract the names of chemical compounds considered as new inhibitors and can be used to treat HIV infection. The examples of the names of HIV reverse transcriptase inhibitors extracted from the texts of abstracts with their PubChem identifiers and their chemical structures are shown in **Supplementary Figure 4**.

Identification of the Chemical Compounds Responsible for the Velocity of HIV/AIDS Progression: A Case Study for HIV Elite Controllers

Many of extracted compounds from elite controllers test set were parts of HAART or names of amino acids. But among all extracted compounds we were able to detect some that influenced HIV/AIDS progression.

For example, the article by Bermejo et al. (2016) describes the effect of the tyrosine kinase inhibitor dasatinib. During T-cell activation phosphorylation of SAMHD1 allows HIV infection. Dasatinib stopped SAMHD1 phosphorylation, which led to disruption of HIV reverse transcription.

Joshi et al. (2016) reported the relationship between heat shock protein 90 (Hsp90) inhibitors and HIV transcription. It has been shown that administration of Hsp90 inhibitors tanespimycin [17- (allylamino)-17-demethoxygeldanamycin] and AUY922 durably prevented viral rebound in mice.

As was mentioned above (in Results section) there were not plenty recognized chemical compounds that may lead to HIV/AIDS non-progression. But some of them were extracted, it demonstrates the usefulness of the approach developed and the possibility of working out this direction in the future.

Identification of the Proteins and Genes Responsible for the Velocity of HIV/AIDS Progression: Case Study for HIV Elite Controllers

Once we were able to extract the set of protein and gene names from the texts relevant to HIV/AIDS ECs, we aimed to automatically identify the main biological processes and functions associated with them based on Gene Ontology (Gene Ontology Consortium, 2015) terms available from UniProt.

Automated queries to the UniProt database allow us to identify the belonging of a protein or a gene to either organism “homo sapiens” or a virus. There were some NEs associated with HIV, such as “gag-pol protein” or “pol peptide.” Also, we found that 11 names were not associated with any proteins, they represent false-positive results of our algorithm. Therefore, the automated verification using a database or a dictionary of proteins can help filter out the named entities that represent false positive results and therefore improve the recognition accuracy obtained using CRF (Song et al., 2015; Perera et al., 2020). It also helps select the names of protein belonging to the species of interest. We selected only proteins that were found in UniProt database and included the names of a protein extracted as one of the synonyms of names presented in UniProt. Therefore, some of proteins were filtered out because they were not associated with one unique record in UniProt.

As a result of automated processing of the files with gene/protein identifiers and their GO terms, we collected the most important biological processes associated with proteins extracted. They can be associated with HIV infection and can have an impact on the velocity of HIV/AIDS disease progression. We divided them into several groups according to the function most important for HIV-progression (see **Table 7**).

For some proteins, the names of which had been extracted from the texts of publications relevant to the studies of HIV ECs, we found direct associations of these proteins with the HIV progression (Taylor et al., 2000; Oleksyk et al., 2009; Marras et al., 2013; Slavov et al., 2015; Roy et al., 2017; Parodi et al., 2019; Hersberger et al., 2020; Wendel et al., 2020). We provide a few examples of the association between biological processes known for the proteins identified and their possible role in HIV disease progression. For instance, H. Fausther-Bovendo and co-authors reported that the increased expression of NKp44L was observed in CD4+ T cells of HIV-positive patients (Fausther-Bovendo et al., 2009); that leads in an increased sensitivity of NKp44LCD4T cells to the NK lysis activity. The cd85j receptor (LIR-1) was extracted from abstracts and is found to be associated with negative regulation of CD8-positive T cell

TABLE 7 | The main functions of some proteins and genes found in the Uniprot database by the search using named entities extracted from abstracts of publications selected by their relevance to the description of HIV elite controllers.

The list of proteins	
Immune response	Human leukocyte elastase; nkg2a; nkg2d receptor; nkp30; nkp44; antileukoproteinase (ALP), apolipoprotein A-I, c4b-binding protein alpha chain (C4bp), complement C4-B (Basic complement C4), interferon-gamma; Calcitonin gene-related peptide 1 (Alpha-type CGRP); Antileukoproteinase (ALP); C4b-binding protein alpha chain (C4bp); Complement C4-B (Basic complement C4); HLA class I histocompatibility antigen (HLA-B27K protein) (MHC class I antigen) (Major histocompatibility complex); HLA-B alpha chain (B*5703GB) (MHC class I antigen); Interferon beta (IFN-beta) (Fibroblast interferon), Apolipoprotein A-I (Apo-AI); Interferon beta (IFN-beta) (Fibroblast interferon)
Autophagy	Alpha-1A adrenergic receptor; putative peripheral benzodiazepine receptor-related protein; forkhead box protein O3; interferon gamma; microtubule-associated proteins 1A/1B; microtubule-associated protein 1S (MAP-1S); microtubule-associated protein tau (Neurofibrillary tangle protein); Platelet-activating factor acetylhydrolase IB subunit beta; Microtubule-associated proteins 1A; Nicotinamide phosphoribosyltransferase (NAMPTase)
inflammatory response	Adenosine deaminase (EC 3.5.4.4) (Adenosine aminohydrolase); angiotensin-converting enzyme 2; antithrombin-III (ATIII) (Serpin C1); calcitonin; 5'-nucleotidase (5'-NT); Integrin beta-2; human leukocyte antigen b57; interferon-gamma; tumor necrosis factor(nf)- alpha, interleukin (il)-2, C-reactive protein; Interferon gamma (IFN-gamma) (Immune interferon); Platelet factor 4 (PF-4); prostaglandin F2-alpha receptor (PGF receptor) (PGF2-alpha receptor) (Prostanoid FP receptor); prothrombin (EC 3.4.21.5) (Coagulation factor II)
Negative regulation of T cell mediated cytotoxicity	Carcinoembryonic antigen-related cell adhesion molecule 1; leukocyte immunoglobulin-like receptor subfamily B member 1 (CD85 antigen-like family member J); leukocyte immunoglobulin-like receptor subfamily B member 1 (LIR-1); HLA class I histocompatibility antigen, alpha chain G (HLA G antigen)
Negative regulation of CD8-positive, alpha-beta T cell activation	Leukocyte immunoglobulin-like receptor subfamily B member 1 (LIR-1) (cd85j receptor)
Regulation of gene expression	Progonadoliberein-1 (Progonadoliberein I), lhrh, Prostaglandin F2-alpha receptor (PGF receptor); Angiotensin-converting enzyme (ACE); Myc proto-oncogene protein (Proto-oncogene c-Myc); Calcitonin receptor (CT-R); Leukocyte immunoglobulin-like receptor subfamily B member 1 (LIR-1); Core histone macro-H2A.2, C-reactive protein; Estrogen receptor (ER); Pro-epidermal growth factor (EGF); Fibronectin (FN); Interferon gamma (IFN-gamma)

activation according to UniProt data (Table 7). Also, there are data on its role in the control of HIV-1 replication in autologous dendritic cells (Scott-Algara et al., 2008). Also, we found the studies aimed at the identification of the interactions between S100A9 protein (a calcium-binding protein of the S100 family) and cd85j receptor. In particular, it was shown that HIV-1 infection modulates S100A9 expression on the surface of the monocyte-derived dendritic cells. Interaction between S100A9 protein and cd85j receptor, in turn, can have an impact on the anti-HIV activity of human NK (natural killer) cells (Arnold et al., 2013). Vincent Arnold, and co-authors suppose that an exogenous peptide S100A9 can be considered as the potential ligand for the control HIV-1 replication by NK cells (Arnold et al., 2013). Therefore, we can identify the existing novel approaches for HIV infection control that can be useful for the development of novel strategies to cure HIV. It also can help to create new hypotheses about potential mechanisms of HIV control leading to the development of new approaches to HIV treatment.

For some other proteins, there was no direct evidence of their role in HIV infection control and progression. But the analysis of their functions led to understanding that the differences in the expression of these proteins in the CD4+ or CD8+ T cells and some other immune cells may be associated with the velocity of HIV/AIDS progression. For instance, since HIV-1 glycoprotein 41 (gp) 41 prefers to interact with the cell-surfaced human leukocyte elastase (Bristow et al., 2003), one can suggest that the low levels of HLE expression can slow down the

dissemination of HIV particles and therefore have an important role in HIV/AIDS progression.

There are experiments that provide the insights into CD8+ T cell response associated with the function of carcinoembryonic antigen-related cell adhesion molecule 1 (Khairnar et al., 2018). Based on the experiments carried out with lymphocytic choriomeningitis virus it was shown that carcinoembryonic antigen-related cell adhesion molecule 1 is essential for activation of CD8+ T cells. However, such results should be considered with awareness and more experiments are needed to adopt these hypotheses to HIV-1 viremic control.

Based on the text and data mining, we have earlier identified a set of proteins and discussed some molecular mechanisms shared by a novel coronavirus SARS-CoV-2 and HIV-1 (Tarasova et al., 2020). There were a few molecular pathways, including those related to immunology, autophagy, cell cycle regulation, shared by these two viruses if they infect humans. The present study is focused on the possibilities of text mining to extract data from the strictly relevant publications and investigate whether such an approach can address questions related to the aspect of biological studies. Our approach is based on a particular set of proteins that can be associated with the slow HIV/AIDS disease progression.

The two case studies aimed at the extraction of chemical names from the texts relevant to HIV reverse transcriptase inhibition, proteins and genes from the texts relevant to HIV control allow us to determine the advantages and disadvantages of text mining approaches to new information. The main advantage of text mining approaches is the possibility of covering

the huge amount of textual data (Ruusmann and Maran, 2013; Capuzzi et al., 2017, 2018; Kandhro et al., 2017; Azam et al., 2019; Gambardella and di Bernardo, 2019; Guin et al., 2019; Ivanisenko et al., 2019; Alves et al., 2020). Text mining approaches allow retrieving the most recent and important information about chemicals, proteins, and genes associated with HIV treatment including their tissue-specific expression level (Ivanisenko et al., 2019). The main disadvantage is that our approach does not consider the expression level of the proteins extracted; this is due to the incomplete description of expression data in abstracts. On the other hand, we have evaluated the fully automated workflow for the purposes of extraction and analysis of the proteins or genes names that can be associated with the investigation of HIV ECs. Our study demonstrates that automated analysis of protein functions associated with HIV elite controllers allows us to hypothesize about the role of this protein in the HIV/AIDS progression. These observations and hypotheses may help to plan new experiments and develop new methods for HIV/AIDS treatment including the search for novel chemical compounds that can modulate the level of expression of target proteins, and vaccine development. We suppose that the suggested approach can be applied to an analysis of other viral infections, including those that have been affecting humanity for the last years (Basak et al., 2019; Tarasova et al., 2020; Tworowski et al., 2020). In addition it may help provide the possibility to analyze other various pathological processes non-related to viral infections that can involve the changes in gene expression (Kovalenko et al., 2016; Kandhro et al., 2017; Bizzarri et al., 2020) and find possible strategies to combat pathological conditions.

CONCLUSIONS

In our study, we have developed and tested a new approach for automated extraction of named entities representing proteins and chemical compounds from the texts of scientific publications. Our method is based on the conditional random fields algorithm. We have developed a set of text features providing the reasonable accuracy of named entity recognition. We proposed the retrieval of named entities from the set of papers strictly relevant to (i) HIV reverse transcriptase inhibition and (ii) to the control of HIV/AIDS progression to test the ability of the algorithm developed to extract both the names of chemicals and proteins (genes). Our algorithm was tested on the retrieval of data on inhibitors of HIV reverse transcriptase. We were able to identify the HIV RT inhibitors with the precision 0.80 and recall 0.94.

REFERENCES

- Alves, V. M., Capuzzi, S. J., Baker, N., Muratov, E. N., Trospsa, A., and Hickey, A. J. (2020). "Mining complex biomedical literature for actionable knowledge on rare diseases," in *Approaching Complex Diseases Human Perspectives in Health Sciences and Technology*, ed. M. Bizzarri (Cham: Springer International Publishing), 77–94. doi: 10.1007/978-3-030-32857-3_4
- Arnold, V., Cummings, J.-S., Moreno-Nieves, U. Y., Didier, C., Gilbert, A., Barré-Sinoussi, F., et al. (2013). S100A9 protein is a novel ligand for the CD85j receptor and its interaction is implicated in the control of HIV-1 replication by NK cells. *Retrovirology* 10:122. doi: 10.1186/1742-4690-10-122
- Azam, Md. F., Musa, A., Dehmer, M., Yli-Harja, O. P., and Emmert-Streib, F. (2019). Global genetics research in prostate cancer: a text mining and computational network theory approach. *Front. Genet.* 10:70. doi: 10.3389/fgene.2019.00070
- Basak, S. C., Majumdar, S., Nandy, A., Roy, P., Dutta, T., Vracko, M., et al. (2019). Computer-assisted and data driven approaches for surveillance, drug
- Then, we tested the applicability of our algorithm to identify a set of proteins potentially responsible for slow HIV/AIDS disease progression and HIV control. For this purpose, we collected a set of abstracts strictly relevant to the HIV elite controllers, a group of HIV positive patients who did not progress into HIV/AIDS for years in the absence of antiretroviral therapy. The extraction of proteins unique for the studies of HIV elite controllers allows us to identify the set of proteins responsible for the velocity of HIV/AIDS disease progression. Investigation of these proteins and their functions can provide insights into novel approaches for HIV/AIDS treatment.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

NB: provided realization, optimization and testing of the CRF algorithm, extraction of the names of chemical compounds and proteins/genes, computational experiments on the chemical data extraction, and manuscript preparation. OT: study design, preparation of test sets, collection of the abstracts of publication relevant to HIV reverse transcriptase inhibitors, HIV elite controllers, HIV-positive patients from NCBI PubMed and NCBI PMC databases, automated search of the extracted proteins/genes by their name in UniProt database, analysis of the results, manuscript preparation and review, and general supervision. SI: selection of the publications relevant to HIV elite controllers, manuscript review, and useful comments. VP: manuscript review, comments to the study design and to the manuscript content. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Russian Science Foundation Grant No. 19-75-10097.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.618862/full#supplementary-material>

- discovery, and vaccine design for the Zika virus. *Pharmaceuticals* 12:157. doi: 10.3390/ph12040157
- Batista-Navarro, R., Rak, R., and Ananiadou, S. (2015). Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *J. Cheminform.* 7:S6. doi: 10.1186/1758-2946-7-S1-S6
- Bermejo, M., López-Huertas, M. R., García-Pérez, J., Climent, N., Descours, B., Ambrosioni, J., et al. (2016). Dasatinib inhibits HIV-1 replication through the interference of SAMHD1 phosphorylation in CD4+ T cells. *Biochem. Pharmacol.* 106, 30–45. doi: 10.1016/j.bcp.2016.02.002
- Bizzarri, M., Lagana, A. S., Aragona, D., Unfer, V. (2020). Inositol and pulmonary function. Could myo-inositol treatment downregulate inflammation and cytokine release syndrome in SARS-CoV-2? *Eur. Rev. Med. Pharmacol. Sci.* 24, 3426–3432. doi: 10.26355/eurrev_202003_20715
- Bristow, C. L., Mercatante, D. R., and Kole, R. (2003). HIV-1 preferentially binds receptors copatched with cell-surface elastase. *Blood* 102, 4479–4486. doi: 10.1182/blood-2003-05-1635
- Campos, D., Matos, S., and Oliveira, J. L. (2015). A document processing pipeline for annotating chemical entities in scientific documents. *J. Cheminform.* 7:S7. doi: 10.1186/1758-2946-7-S1-S7
- Capuzzi, S. J., Kim, I. S.-J., Lam, W. I., Thornton, T. E., Muratov, E. N., Pozefsky, D., et al. (2017). Chembench: a publicly accessible, integrated cheminformatics portal. *J. Chem. Inf. Model.* 57, 105–108. doi: 10.1021/acs.jcim.6b00462
- Capuzzi, S. J., Thornton, T. E., Liu, K., Baker, N., Lam, W. I., O'Banion, C. P., et al. (2018). Chemotext: a publicly available web server for mining drug-target-disease relationships in PubMed. *J. Chem. Inf. Model.* 58, 212–218. doi: 10.1021/acs.jcim.7b00589
- Carpenter, B. (2007). “LingPipe for 99.9% recall of gene mentions,” in *Proceedings of the 2nd BioCreative Workshop*. Valencia. Available online at: <http://www.alias-i.com/lingpipe/index.html> (accessed December 12, 2020).
- Cash, G. G. (2004). Quantitative structure–activity relationship (QSAR) models of mutagens and carcinogens edited by Romualdo Benigni. *J. Med. Chem.* 47, 5330–5330. doi: 10.1021/jm040146z
- Chen, L., Lu, J., Huang, T., Yin, J., Wei, L., and Cai, Y.-D. (2014). Finding candidate drugs for hepatitis C based on chemical-chemical and chemical-protein interactions. *PLoS ONE* 9:e107767. doi: 10.1371/journal.pone.0107767
- Dai, H.-J., Lai, P.-T., Chang, Y.-C., and Tsai, R. T.-H. (2015). Enhancing of chemical compound and drug name recognition using representative tag scheme and fine-grained tokenization. *J. Cheminform.* 7:S14. doi: 10.1186/1758-2946-7-S1-S14
- Fausther-Bovendo, H., Sol-Foulon, N., Candotti, D., Agut, H., Schwartz, O., Debré, P., et al. (2009). HIV escape from natural killer cytotoxicity: nef inhibits NKP44L expression on CD4+ T cells. *AIDS* 23, 1077–1087. doi: 10.1097/QAD.0b013e32832cb26b
- Gambardella, G., and di Bernardo, D. (2019). A tool for visualization and analysis of single-cell RNA-seq data based on text mining. *Front. Genet.* 10:734. doi: 10.3389/fgene.2019.00734
- Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 43, D1049–1056. doi: 10.1093/nar/gku1179
- Guin, D., Rani, J., Singh, P., Grover, S., Bora, S., Talwar, P., et al. (2019). Global text mining and development of pharmacogenomic knowledge resource for precision medicine. *Front. Pharmacol.* 10:839. doi: 10.3389/fphar.2019.00839
- Hemati, W., and Mehler, A. (2019). LSTMVoter: chemical named entity recognition using a conglomerate of sequence labeling tools. *J. Cheminform.* 11:3. doi: 10.1186/s13321-018-0327-2
- Hersberger, M., Schlaepfer, E., Buehler, M., Bochud, P.-Y., Vernazza, P., Marti-Jaun, J., et al. (2020). Polymorphisms of SOCS-1 are associated with a rapid HIV progression rate. *J. Acquir. Immune Defic. Syndr.* 84, 189–195. doi: 10.1097/QAI.0000000000002319
- Hong, S. K., and Lee, J.-G. (2020). DTranNER: biomedical named entity recognition with deep learning-based label-label transition model. *BMC Bioinform.* 21:53. doi: 10.1186/s12859-020-3393-1
- Huang, M.-S., Lai, P.-T., Lin, P.-Y., You, Y.-T., Tsai, R. T.-H., and Hsu, W.-L. (2020). Biomedical named entity recognition and linking datasets: survey and our recent development. *Brief Bioinform.* 21, 2219–2238. doi: 10.1093/bib/bbaa054
- Ivanisenko, V. A., Demenkov, P. S., Ivanisenko, T. V., Mishchenko, E. L., and Saik, O. V. (2019). A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinform.* 20:34. doi: 10.1186/s12859-018-2567-6
- Iyidogan, P., and Anderson, K. S. (2014). Current perspectives on HIV-1 antiretroviral drug resistance. *Viruses* 6, 4095–4139. doi: 10.3390/v6104095
- Joshi, P., Maidji, E., and Stoddart, C. A. (2016). Inhibition of heat shock protein 90 prevents HIV rebound. *J. Biol. Chem.* 291, 10332–10346. doi: 10.1074/jbc.M116.717538
- Kaewphan, S., Hakala, K., Miekka, N., Salakoski, T., and Ginter, F. (2018). Wide-scope biomedical named entity recognition and normalization with CRFs, fuzzy matching and character level modeling. *Database* 2018, 1–10. doi: 10.1093/database/bay096
- Kandhro, A. H., Shoomboatong, W., Nantasenamat, C., Prachayasittikul, V., and Nuchnoi, P. (2017). The MicroRNA interaction network of lipid diseases. *Front. Genet.* 8:116. doi: 10.3389/fgene.2017.00116
- Khabasa, M., and Giles, C. L. (2015). Chemical entity extraction using CRF and an ensemble of extractors. *J. Cheminform.* 7:S12. doi: 10.1186/1758-2946-7-S1-S12
- Khairnar, V., Duhan, V., Patil, A. M., Zhou, F., Bhat, H., Thoens, C., et al. (2018). CEACAM1 promotes CD8+ T cell responses and improves control of a chronic viral infection. *Nat. Commun.* 9:2561. doi: 10.1038/s41467-018-04832-2
- Kilicoglu, H., Rosembat, G., Fisman, M., and Shin, D. (2020). Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinform.* 21:188. doi: 10.1186/s12859-020-3517-7
- Korvigo, I., Holmatov, M., Zaikovskii, A., and Skoblov, M. (2018). Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. *J. Cheminform.* 10:28. doi: 10.1186/s13321-018-0280-0
- Kovalenko, I. L., Smagin, D. A., Galyamina, A. G., Orlov, Y. L., and Kudryavtseva, N. N. (2016). Changes in the expression of dopaminergic genes in brain structures of male mice exposed to chronic social defeat stress: an RNA-seq study. *Mol. Biol.* 50, 184–187. doi: 10.1134/S0026893316010088
- Krallinger, M., Rabal, O., Leitner, F., Vazquez, M., Salgado, D., Lu, Z., et al. (2015). The ChEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform.* 7:S2. doi: 10.1186/1758-2946-7-S1-S2
- Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *J. Biomed. Inform.* 57, 28–37. doi: 10.1016/j.jbi.2015.07.010
- Lim, Y. X., Ng, Y. L., Tam, J. P., and Liu, D. X. (2016). Human coronaviruses: a review of virus-host interactions. *Diseases* 4:26. doi: 10.3390/diseases4030026
- Marras, F., Nicco, E., Bozzano, F., Di Biagio, A., Dentone, C., Pontali, E., et al. (2013). Natural killer cells in HIV controller patients express an activated effector phenotype and do not up-regulate NKP44 on IL-2 stimulation. *Proc. Natl. Acad. Sci. U. S. A.* 110, 11970–11975. doi: 10.1073/pnas.1302090110
- Oleksyk, T. K., Shrestha, S., Truelove, A. L., Goedert, J. J., Donfield, S. M., Phair, J., et al. (2009). Extended IL10 haplotypes and their association with HIV progression to AIDS. *Genes Immun.* 10, 309–322. doi: 10.1038/gene.2009.9
- Parodi, M., Favoreel, H., Candiano, G., Gaggero, S., Sivori, S., Mingari, M. C., et al. (2019). NKP44-NKP44 ligand interactions in the regulation of natural killer cells and other innate lymphoid cells in humans. *Front. Immunol.* 10:719. doi: 10.3389/fimmu.2019.00719
- Perera, N., Dehmer, M., and Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. *Front. Cell Dev. Biol.* 8:673. doi: 10.3389/fcell.2020.00673
- Rocktäschel, T., Weidlich, M., and Leser, U. (2012). ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* 28, 1633–1640. doi: 10.1093/bioinformatics/bts183
- Rojas-Celis, V., Valiente-Echeverría, F., Soto-Rifo, R., and Toro-Ascuy, D. (2019). New challenges of HIV-1 infection: how HIV-1 attacks and resides in the central nervous system. *Cells* 8:1245. doi: 10.3390/cells8101245
- Roy, A., Banerjee, R., and Basak, S. (2017). HIV progression depends on codon and amino acid usage profile of envelope protein and associated host-genetic influence. *Front. Microbiol.* 8:1083. doi: 10.3389/fmicb.2017.01083
- Ruusmann, V., and Maran, U. (2013). From data point timelines to a well curated data set, data mining of experimental data and chemical structure data from scientific articles, problems and possible solutions. *J. Comput. Aided Mol. Des.* 27, 583–603. doi: 10.1007/s10822-013-9664-4
- Saik, O. V., Ivanisenko, T. V., Demenkov, P. S., and Ivanisenko, V. A. (2016). Interactome of the hepatitis C virus: literature mining with ANDSystem. *Virus Res.* 218, 40–48. doi: 10.1016/j.virusres.2015.12.003

- Scott-Algara, D., Arnold, V., Didier, C., Kattan, T., Pirozzi, G., Barré-Sinoussi, F., et al. (2008). The CD85j+ NK cell subset potently controls HIV-1 replication in autologous dendritic cells. *PLoS ONE* 3:e1975. doi: 10.1371/journal.pone.0001975
- Slavov, S. N., Vilar, F. C., Wagatsuma, V. M. D., Santana, R. C., Machado, A. A., da Fonseca, B. A. L., et al. (2015). Late emergence of A594V and L595W mutations related to ganciclovir resistance in a patient with HCMV retinitis and long-term HIV progression. *Braz. J. Med. Biol. Res.* 48, 777–781. doi: 10.1590/1414-431x20154507
- Song, M., Yu, H., and Han, W.-S. (2015). Developing a hybrid dictionary-based bio-entity recognition technique. *BMC Med. Inform. Decis. Mak.* 15 (Suppl. 1):S9. doi: 10.1186/1472-6947-15-S1-S9
- Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *J. R. Stat. Soc.* 36, 111–133. doi: 10.1111/j.2517-6161.1974.tb00994.x
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131
- Tannenbaum, C., and Sheehan, N. L. (2014). Understanding and preventing drug-drug and drug-gene interactions. *Expert Rev. Clin. Pharmacol.* 7, 533–544. doi: 10.1586/17512433.2014.910111
- Tarasova, O., Ivanov, S., Filimonov, D. A., and Poroikov, V. (2020). Data and text mining help identify key proteins involved in the molecular mechanisms shared by SARS-CoV-2 and HIV-1. *Molecules* 25:2944. doi: 10.3390/molecules25122944
- Tarasova, O., and Poroikov, V. (2018). HIV resistance prediction to reverse transcriptase inhibitors: focus on open data. *Molecules* 23:956. doi: 10.3390/molecules23040956
- Tarasova, O. A., Biziukova, N. Y., Filimonov, D. A., Poroikov, V. V., and Nicklaus, M. C. (2019). Data mining approach for extraction of useful information about biologically active compounds from publications. *J. Chem. Inf. Model.* 59, 3635–3644. doi: 10.1021/acs.jcim.9b00164
- Tarasova, O. A., Urusova, A. F., Filimonov, D. A., Nicklaus, M. C., Zakharov, A. V., and Poroikov, V. V. (2015). QSAR modeling using large-scale databases: case study for HIV-1 reverse transcriptase inhibitors. *J. Chem. Inf. Model.* 55, 1388–1399. doi: 10.1021/acs.jcim.5b00019
- Taylor, J. M., Wang, Y., Ahdieh, L., Chmiel, J. S., Detels, R., Giorgi, J. V., et al. (2000). Causal pathways for CCR5 genotype and HIV progression. *J. Acquir. Immune Defic. Syndr.* 23, 160–171. doi: 10.1097/00042560-200002010-00008
- Tworowski, D., Gorohovski, A., Mukherjee, S., Carmi, G., Levy, E., Detroja, R., et al. (2020). COVID19 drug repository: text-mining the literature in search of putative COVID19 therapeutics. *Nucleic Acids Res.* gkaa969. doi: 10.1093/nar/gkaa969
- Ventura, J. D. (2020). Human Immunodeficiency virus 1 (HIV-1): viral latency, the reservoir, and the cure. *Yale J. Biol. Med.* 93, 549–560.
- Wendel, B. S., Fu, Y., He, C., Hernandez, S. M., Qu, M., Zhang, Z., et al. (2020). Rapid HIV progression is associated with extensive ongoing somatic hypermutation. *J. Immunol.* 205, 587–594. doi: 10.4049/jimmunol.1901161
- Xu, S., An, X., Zhu, L., Zhang, Y., and Zhang, H. (2015). A CRF-based system for recognizing chemical entity mentions (CEMs) in biomedical literature. *J. Cheminform.* 7:S11. doi: 10.1186/1758-2946-7-S1-S11

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Biziukova, Tarasova, Ivanov and Poroikov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Comparative Expression Analysis of Stress-Inducible Candidate Genes in Response to Cold and Drought in Tea Plant [*Camellia sinensis* (L.) Kuntze]

Lidiia S. Samarina^{1*}, Alexandr V. Bobrovskikh^{1,2†}, Alexey V. Doroshkov^{1,2}, Lyudmila S. Malyukova¹, Alexandra O. Matskiv¹, Ruslan S. Rakhmangulov¹, Natalia G. Koninskaya¹, Valentina I. Malyarovskaya¹, Wei Tong³, Enhua Xia³, Karina A. Manakhova¹, Alexey V. Ryndin¹ and Yuriy L. Orlov^{1,4}

¹ Biotechnology Department, Federal Research Centre the Subtropical Scientific Centre of the Russian Academy of Sciences, Sochi, Russia, ² Institute Cytology and Genetics Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, ³ State Key Laboratory of Tea Plant Biology and Utilization, Anhui Agricultural University, Hefei, China, ⁴ Agrarian and Technological Institute, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia

OPEN ACCESS

Edited by:

Joao Carlos Setubal,
University of São Paulo, Brazil

Reviewed by:

Mehtap Sahin Cevik,
Isparta University of Applied
Sciences, Turkey
Robert Kipyegon Koech,
University of Pretoria, South Africa

*Correspondence:

Lidiia S. Samarina
samarinalidia@gmail.com

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 28 September 2020

Accepted: 23 November 2020

Published: 23 December 2020

Citation:

Samarina LS, Bobrovskikh AV, Doroshkov AV, Malyukova LS, Matskiv AO, Rakhmangulov RS, Koninskaya NG, Malyarovskaya VI, Tong W, Xia E, Manakhova KA, Ryndin AV and Orlov YL (2020) Comparative Expression Analysis of Stress-Inducible Candidate Genes in Response to Cold and Drought in Tea Plant [*Camellia sinensis* (L.) Kuntze]. *Front. Genet.* 11:611283. doi: 10.3389/fgene.2020.611283

Cold and drought are two of the most severe threats affecting the growth and productivity of the tea plant, limiting its global spread. Both stresses cause osmotic changes in the cells of the tea plant by decreasing their water potential. To develop cultivars that are tolerant to both stresses, it is essential to understand the genetic responses of tea plant to these two stresses, particularly in terms of the genes involved. In this study, we combined literature data with interspecific transcriptomic analyses (using *Arabidopsis thaliana* and *Solanum lycopersicum*) to choose genes related to cold tolerance. We identified 45 stress-inducible candidate genes associated with cold and drought responses in tea plants based on a comprehensive homologous detection method. Of these, nine were newly characterized by us, and 36 had previously been reported. The gene network analysis revealed upregulated expression in *ICE1*-related cluster of *bHLH* factors, *HSP70/BAM5* connected genes (hexokinases, galactinol synthases, *SnRK* complex, etc.) indicating their possible co-expression. Using qRT-PCR we revealed that 10 genes were significantly upregulated in response to both cold and drought in tea plant: *HSP70*, *GST*, *SUS1*, *DHN1*, *BMV5*, *bHLH102*, *GR-RBP3*, *ICE1*, *GOLS1*, and *GOLS3*. *SnRK1.2*, *HXK1/2*, *bHLH7/43/79/93* were specifically upregulated in cold, while *RHL41*, *CAU1*, *Hydrolase22* were specifically upregulated in drought. Interestingly, the expression of *CIP* was higher in the recovery stage of both stresses, indicating its potentially important role in plant recovery after stress. In addition, some genes, such as *DHN3*, *bHLH79*, *PEI54*, *SnRK1.2*, *SnRK1.3*, and *Hydrolase22*, were significantly positively correlated between the cold and drought responses. *CBF1*, *GOLS1*, *HXK2*, and *HXK3*, by contrast, showed significantly negative correlations between the cold and drought responses. Our results provide valuable information and robust candidate genes for future functional analyses intended to improve the stress tolerance of the tea plant and other species.

Keywords: gene expression regulation, homologs detection, genetic markers, principal component analysis, expression profile, candidate genes, abiotic stress, tea plant

INTRODUCTION

Cold and drought reduce the yield and geographical distribution of most horticultural crops worldwide. Both can lead to decreased water potential of tissues and induce reactive oxygen species accumulation, which causes severe damage to various cellular components (Minhas et al., 2017). Plant responses are complex, particularly in perennial woody crops, and hundreds of genes are involved in them (Chaves et al., 2003; Hao et al., 2018; Xia et al., 2019a). Earlier studies showed that plants have specific and non-specific responses to both stresses (Beck et al., 2007). Cold and drought induce common stress-inducible genes, while one of the stresses specifically induce some genes (Zhou et al., 2019). It is important to identify these common and unique responses under cold and drought stress for understanding the cross-talk mechanisms. To develop cultivars that are tolerant to both cold and drought, it is necessary to reveal the genes that are involved in both stresses and elucidate their response mechanisms to develop genetic markers that can help facilitate breeding programs (Minhas et al., 2017).

The tea plant (*Camellia sinensis* L.) is one of the most important economic crops in China, India, Sri Lanka, Kenya, and certain Caucasian countries (Turkey, Georgia, Russia, and Azerbaijan). This perennial woody evergreen crop is grown in more than 60 countries on five continents, from 49°N in Ukraine to 33°S in South Africa (Turkuzu and Sanlier, 2017). Caucasus tea germplasm collection (44°36'40" N, 40°06'40" E) is located in the border region of the possible tea production and can be the source of the most tolerant cultivars; some genotypes here survive −15–17°C (Tuov and Ryndin, 2011). In most countries, tea plantations are affected by drought and cold stress that significantly reduces the yield and decreases the distribution of the crop in colder areas. Due to out-breeding and its long gestation period, the tea plant requires next-generation breeding strategies to improve its drought and cold tolerance through a deeper understanding of key regulators and their variants for precision introgressions to have better yield and quality under stress conditions. Therefore, efforts are needed to elucidate the global transcriptomic dynamics of multiple tea genotypes in drought and cold stress to critically discern key molecular players (Parmar et al., 2019).

Many transcription factors and metabolite-related genes have been shown to be involved in both the cold and drought responses of plants. For example, the key cold regulators *ICE*, *CBF*, and *DHN* transcription factors participate in both cold and drought and in other abiotic stresses (Liu et al., 2015; Liu S.-C. et al., 2016; Yin et al., 2016; Ban et al., 2017; Hu et al., 2020). The genes involved in the ABA-independent responsive pathway and the *bZIP*-mediated ABA-dependent pathway (Wang et al., 2012; Ban et al., 2017) also participate in tolerance to cold and drought. The overexpression of *CsbZIP6* in *Arabidopsis* resulted in hypersensitivity to several abiotic stresses (Cao et al., 2015). In addition, many other transcription factors (*WRKY*, *bHLH*, *NAC*, *HSP*, *LEA*, *CML*, and others) have been shown to be activated in tea plants in response to cold and drought (Yue et al., 2015; Wang Y.-X. et al., 2016; Chen et al., 2018; Cui et al., 2018; Ma et al., 2019). Recently, Li Y. et al. (2019) revealed that the

genes *LEA2*, *HSP70*, *PRP*, *CIPs*, *PEIs*, *TLPs*, and *ChiA* were more strongly expressed under cold stress in tolerant cultivars than in susceptible cultivars. Recent transcriptomic data on tea plant showed that 12 TF families (*AP2/EREBP*, *bHLH*, *bZIP*, *HD-ZIP*, *HSE*, *MYB*, *NAC*, *WRKY*, zinc-finger protein TFs, *SCL*, *ARR*, and *SPL*) might play crucial roles in tea plant responding to drought (Liu S.-C. et al., 2016). In *Arabidopsis thaliana*, forty three transcription factor families (primarily, *WRKY*, *NAC*, *MYB*, *AP2/ERF*, and *bZIP*) were found to regulate 56% of common genes expressed in drought and cold stress (Sharma et al., 2018).

However, we continue to lack a complex picture of the interactions between the core network and their downstream-regulated target proteins. Additionally, comparison of molecular profiles of an organism under different stresses would make it possible to identify the conserved stress mechanisms (Amrine et al., 2015; Muthuramalingam et al., 2017; Chamani Mohasses et al., 2020). Thus, we have to continue searching for new evolutionarily conserved and species-specific genes related to the stress response. In this study, we combined literature data with interspecific transcriptomic analyses (*A. thaliana* and *Solanum lycopersicum*) to select genes that are related to cold tolerance. We built a network of candidate genes to reveal their interactions with the corresponding homologs for *A. thaliana*. We phenotypically screened a panel of Caucasian tea genotypes for cold and drought tolerance. Further expression analyses of 45 genes were performed in the most tolerant genotype under long-term stress induction and during the following recovery. The cold and drought expression profiles for each gene were compared to analyze overlapping responses in tea plant to both stresses, and correlations between cold and drought were revealed. Our results provide valuable information and robust candidate genes for future functional analyses intended to improve the stress tolerance of the tea plant and other species.

MATERIALS AND METHODS

Candidate Genes Selection

To evaluate the cross-talk of the genetic response between cold and drought, cold responsive genes were selected as described below. The same genes have been tested in response to drought conditions.

We performed the interspecific analysis of transcriptomic data from the NCBI GEO database (ncbi.nlm.nih.gov/geo/, Barrett et al., 2012) for revealing candidate genes with increasing expression during cold. Using the datasets GSE103964, GSE112225, GSE116964 for *A. thaliana* and GSE78154 for *S. lycopersicum* the fold changes of gene expression under cold were calculated and ranks of genes were assigned according to their upregulation quartile (from 1 to 4) (**Supplementary Table 1**). Next, we compared top quartile genes between *A. thaliana* and *S. lycopersicum* using standalone BLAST (Camacho et al., 2009). As a result, nine orthologs were detected as genes with the highest rank in both species, and their nine corresponding orthologs of *Camellia sinensis* were added in experiment.

Further corresponding homologs in tea plant were characterized using BLAST against the Tea Plant Information

TABLE 1 | Candidate cold responsive genes in tea plant.

Source	Gene ID	Description	Trivial name
Best hit of interspecies top ranked genes	TEA003328	Galactinol synthase 1	<i>GOLS1</i>
	TEA006793	Galactinol synthase 3	<i>GOLS3</i>
	TEA030611	Glycine-rich RNA-binding protein 3	<i>GR-RBP3</i>
	TEA021045	Endotransglucosylase	<i>HYDROLASE 22</i>
	TEA020473	Responsive to high light 41	<i>RHL41</i>
	TEA010353	Calcium underaccumulation 1	<i>CAU1</i>
	TEA003997	Pectin methylesterase 41	<i>PME41</i>
	TEA004079	Dehydration response element-binding protein 26	<i>DREB26</i>
Upregulated cold responsive gene (Li Y. et al., 2019)	TEA024722	Aba- and osmotic-stress-inducible	<i>ARCK1, CRK45</i>
	CSA032195	G-type lectin S-receptor-like serine/threonine-protein kinase At4g27290 [<i>Vitis vinifera</i>]	<i>GsSRK</i>
	CSA031147	G-type lectin S-receptor-like serine/threonine-protein kinase RKS1[<i>Theobroma cacao</i>]	<i>GsSRK1</i>
	CSA001565	LRR receptor-like serine/threonine-protein kinase FLS2-like	<i>FLS2</i>
	CSA020614	Receptor-like serine/threonine-protein kinase ALE2 [<i>Nicotiana sylvestris</i>]	<i>RPK2</i>
	CSA000608	Ethylene-responsive transcription factor ERF021[<i>Arabidopsis lyrata</i> subsp. <i>Lyrata</i>]	<i>AP2/ERF-AP21</i>
	CSA000348	Ethylene-responsive transcription factor SHINE 2-like [<i>Cucumis melo</i>]	<i>AP2/ERF-ERF2</i>
	CSA034862	Ethylene response factor 6	<i>AP2/ERF-ERF6</i>
	CSA023474	Bhlh transcription factor bhlh102	<i>bHLH102</i>
	CSA033910	Probable WRKY transcription factor 42	<i>WRKY42</i>
	CSA002423	PREDICTED: zinc finger CCCH domain-containing protein 30 [<i>Ricinus communis</i>]	<i>ZAT30</i>
	CSA003726	Late embryogenesis abundant protein 3L-1 [<i>C. sinensis</i>]	<i>LEA3</i>
	CSA031822	Late embryogenesis abundant protein [<i>C. sinensis</i>]	<i>LEA2</i>
	CSA012537	Heat shock 70 kda protein, mitochondrial-like	<i>HSP70</i>
	CSA014200	36.4 kda proline-rich protein-like [<i>Malus domestica</i>]	<i>PRP</i>
	CSA016010	Putative cold-inducible protein [<i>C. sinensis</i>]	<i>CIP</i>
	CSA001876	Probable pectinesterase/pectinesterase inhibitor 54	<i>PEI54</i>
	CSA035791	Endoglucanase 11-like [<i>Jatropha curcas</i>]	<i>EGase11</i>
	CSA000129	thaumatin-like protein 1b	<i>TLP1</i>
	CSA028426	Peroxidase 73 [<i>Vitis vinifera</i>]	<i>POD73</i>
	CSA006422	Glutathione S-transferase [<i>Camellia japonica</i>]	<i>GST</i>
	CSA010521	Beta-amylase 5 [<i>C. sinensis</i>]	<i>BMV5</i>
	CSA000011	Sucrose synthase 1 [<i>C. sinensis</i>]	<i>SUS1</i>

The genes are retrieved from the TPIA database and the study from Li Y. et al. (2019). For detailed information please see **Supplementary Table 2**.

Archive database (Xia et al., 2019b, **Supplementary Table 2**). The corresponding *A. thaliana* orthologs of *C. sinensis* were also identified from Li Y. et al. (2019) using the best-scored BLAST result. The selected genes (**Table 1**, **Supplementary Table 2**) were further annotated by the blast to the *A. thaliana* TAIR database (Lamesch et al., 2012). Primers were designed using PrimerQuest (eu.idtdna.com/Primerquest) with default parameters and amplicon size between 100 and 250 bp. The quality of the primers was revised using service Multiple Primer Analyzer by Thermofisher Scientific and PCR electrophoresis.

Analysis of Relevance of Selected Genes and Their Interactions

A combined scored method was used to rank the identified genes from 1 to 9 points. In particular, we valued from 2 to 4 if genes have GO terms related to cold response [GO:0009409 Response to cold (“4”), GO:0006979 response to oxidative stress (“3”), GO:0050896 response to a stimulus (“2”)]. Also, we added a score from 1 to 4 if corresponded ortholog was detected in an

upregulated cluster according to *A. thaliana* and *S. lycopersicum* data. Finally, we added 1 point if the gene was presented in related articles. Therefore, genes were ranked (**Supplementary Table 2**) from 1 to 9 points using a combined criterion.

Gene Network Reconstruction and Layout

The data from the literature sources and transcriptome analysis (see **Supplementary Table 2**) were used for the gene network reconstruction. Since most of the data for plant protein-protein interactions were obtained for *A. thaliana*, we identified the best-hit orthologs for *Arabidopsis* (**Supplementary Table 2**, column “AT ID”) and used them as source for building the corresponding gene network.

The network was reconstructed using the String database (https://string-db.org; Szklarczyk et al., 2019) with the following attributes: Textmining/Experiments/Databases interactions and threshold of interaction score = 0.15. For further layout and visualization, we used the Cytoscape (cytoscape.org; Shannon et al., 2003) and algorithm Radial Layout by yFiles.

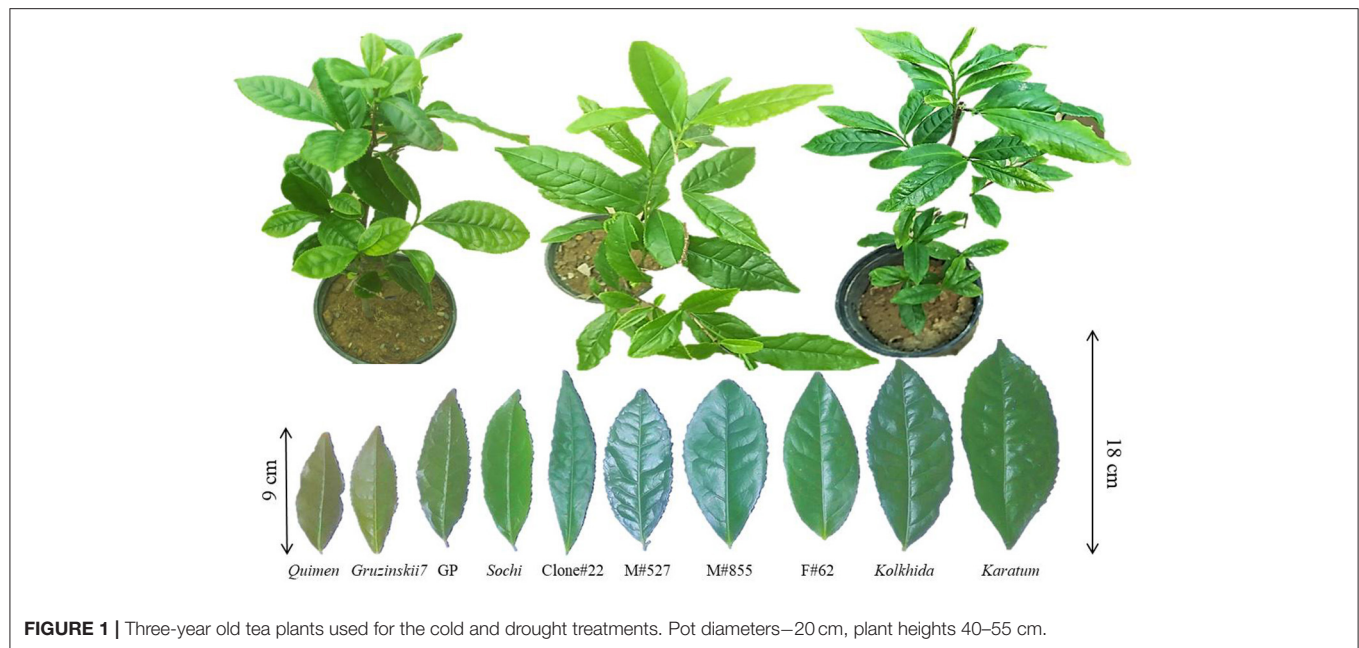


FIGURE 1 | Three-year old tea plants used for the cold and drought treatments. Pot diameters—20 cm, plant heights 40–55 cm.

Plant Material

Three-year-old plants of ten elite tea genotypes obtained by vegetative propagation in FRC SSC RAS (Federal Research Center the Subtropical Scientific Center of the Russian Academy of Sciences) were used for leaf samplings. Ten genotypes of the local breeding were included in this study: *Quimen*, *Gruzinskii7*, GP, *Sochi*, Clone#22, M#527, M#855, Form#62, *Kolkhida*, *Karatum*. Among them, *Quimen*, *Gruzinskii7* were earlier showed to be the cold- and drought-tolerant genotypes; *Kolkhida* and *Karatum* were earlier showed to be cold-susceptible and drought-susceptible genotypes. Other clones and mutant forms showed medium cold-tolerance of drought-tolerance (Gvasaliya, 2015). Plants were grown in 2-liter pots filled with brown forest acidic soil (pH = 5.0) (Figure 1). Only healthy plants were selected for these experiments. Ten plants of each genotype were included in the study. For each assessed parameter, 2nd, 3rd, and 4th mature leaves were used for samplings. Experimental treatments with these plants were replicated twice in the period 2019 to 2020.

Stress Induction and Phenotypical Screening for Tolerance

Control treatment: Before the stress treatments, plants were grown for 3 months in control conditions with the temperature of +22–25°C (with an illumination regime of 16 h of light and 8 h of dark, with the light intensity of 4000 lux with normal irrigation).

Cold treatment: Cold stress was induced in cold chambers HF-506 (Liebherr, Denmark) as follows: decreasing the temperature by 0–+2°C for 10 days to reveal the cold acclimation responses. After that, the temperature was gradually increased to +10°C during 10 days (Recovery-Cold treatment). **Drought treatment:** Drought stress was induced in a laboratory climatic chamber by

gradually decreasing the watering till 15–17% of water content in soil (comparing with control 28–30%) during 10 days (drought treatment) to reveal the drought acclimation response. After that, watering was gradually increased until 28–30% for 10 days (Recovery-Drought treatment). During the treatments, the illumination regime was the same as in the control conditions.

For phenotypical evaluation of the tolerance to stress relative electrical conductivity was measured before the stress induction and after the stress inductions. Relative electrical conductivity was measured using a portable conductivity meter ST300C (Ohaus) to assess the electrolyte leakage indicating the damage of leaf tissues. The leaf sample was immersed in 150 ml of deionized water. The measurement of electrical conductivity was done immediately after immersion (L1) and 2 h later (L2). The relative electrical conductivity (REC, %) was calculated as: $REC = \frac{L1}{L2} * 100$ (Bajji et al., 2001).

Gene Expression Analysis

Total RNA was extracted from the third mature leaf in three biological replicates by the CTAB method (Doyle and Doyle, 1991) with minor modifications. The concentration and quality of RNA were determined using BioDrop μ Lite spectrophotometer and integrity was assessed by agarose gel electrophoresis. RNA samples were treated with DNase I and reverse transcription was performed using the MMLV-RT kit (Biolabmix, Russia). The efficiency of DNaseI treatment and reverse transcription were tested by agarose gel electrophoresis and by qRT-PCR. The results of this verification were evaluated by the presence/absence of a PCR product in RNA samples before and after DNaseI treatment, and by observing the size of PCR fragments in RNA samples before treatment and its cDNA synthesis. Only those samples that confirmed the absence of genomic DNA contamination were included in further analysis of

gene expression. *Actin* (F: 5'-CCATCACCAGAATCCAAGAC-3'; R 5'-GAACCCGAAGGCGAATAGG-3') (Hao et al., 2014) was taken as a reference gene and results were quantified using a Light Cycler 96 analyzer (Roche, Japan). The relative gene expression level was calculated by the Livak and Schmittgen (2001) using the following algorithm: $2^{-\Delta\Delta Cq}$, where:

$$\Delta\Delta Cq = (Cq_{\text{gene of interest}} - Cq_{\text{internal control}})_{\text{treatment}} - (Cq_{\text{gene of interest}} - Cq_{\text{internal control}})_{\text{control}}$$

Statistical Analysis

All analyses were repeated twice with three biological replications in each. Statistical analyses were carried out using XLSTAT software. Student t-test, principal component analysis, and Pearson's correlation tests and Wards-clusterization were performed to evaluate data and confirm the significant differences (at the level $p \leq 0.05$) between the genes expression profiles and respective treatments.

RESULTS

Reconstruction of the Cold Stress Response Gene Regulation Network in Tea Plant for Selection of Priority Targets for Experimental Expression Profiling

A set of 52 genes was involved in the analyses, including nine *de novo* predicted genes from transcriptomic data analyses and 43 from recent articles related to the cold tolerance of *C. sinensis*. The following genes were drawn from the literature: *bHLH* factors (9), *GsSRK* (2), *SnRK1* (3), *HXKs* (3), *ERF* (3), *WRKY* (2), dehydrins (2), late embryogenesis abundant proteins (2), and others (*CBF1*, *ICE1*, *ZAT*, *HSP70*, *PRP*, *CIP*, *PEI54*, *TLP*, *POD*, *GST*, *BMV*, *ALE2*, and *FLS2*). In addition, using interspecies transcriptome analyses we stressed nine orthologs that were highly upregulated in both species (*A. thaliana* and *S. lycopersicum*) using cold treatment: two galactinol synthases (*GOLS1* and *GOLS3*), glycine-rich RNA-binding protein 3 (*GR-RBP3*), xyloglucan endotransglucosylase/hydrolase protein 22 (*XTH22*, *Hydrolase22*), zinc finger protein *RHL41*, histone methylase *SKB1*, pectinesterase inhibitor *PME41*, dehydration response element-binding protein *DREB26*, and protein kinase superfamily protein *ARCK1*. For a better overall understanding of the interactions and to verify our chosen gene set, the gene network was reconstructed using *A. thaliana* data (Figure 2). The core gene network was classified using the three indicated clusters and had 42 genes with 111 edges between them, which indicate their tight interconnection. Interestingly, 30 of 46 genes were upregulated, and seven genes were downregulated.

Phenotypical Selection of Tolerant Genotype Under Cold and Drought Treatments

Cold resulted in increased relative electrical conductivity (REC) that reached 50–60% in most genotypes. Maximum REC was observed in three genotypes: Clone#22, Form#62, and cultivar Kolkhida. The lowest REC was observed in two cold-tolerant

genotypes, Gruzinskii7 and Quimen, at 39 and 31%, respectively. Drought stress resulted in increased REC, which reached 40–49% in most genotypes. The highest REC, 54%, was observed in cv. Kolkhida. The lowest REC, 31%, was observed in cv. Quimen. The recovery stage showed no significant differences among the ten genotypes. Thus, the lowest REC under drought and cold induction was observed in Quimen, indicating the lowest damage of leaf tissues under cold and drought stress (Figure 3). This cultivar showed a similar REC for cold and drought treatment, which produced equal damage to tissues in both stresses, so this cultivar was used as the tolerant one in further gene expression analyses.

Relative Expression Levels of the Studied Genes in Response to Cold and Drought

Of the 45 studied genes, the highest level of expression (hundred-fold) was observed in the four candidate genes in response to a given stress treatment: *HXK2* (Cold), *HSP70* (Cold, RecCold, Drought, and RecDrought), *SUS1*, and *GST* (Cold and RecCold) (Figure 4).

A heat map and hierarchical clustering revealed several clusters characterized by similar gene expression profiles (Figure 4). Cluster 1 combined the two genes *DHN1* and *SnRK1*. 2 with an over 30-fold induced expression in Cold. *DHN1* was also significantly upregulated in Drought and RecCold, indicating its importance in both stress responses.

The other two distant clusters with the most elevated expression were Cluster 3 and Cluster 4, including genes with a 10- to 19-fold upregulation in Cold and RecCold. Cluster 3 included the genes *GOLS3*, *LEA2*, *bHLH7*, and *bHLH93*, with a 13- to 19-fold upregulation under cold stress. Among these, *GOLS3* was also significantly upregulated in RecCold and Drought. Cluster 4 combined five genes (*SnRK1.1*, *SnRK1.3*, *LEA3*, *TLP*, and *FLS2*) that were significantly upregulated under Cold and RecCold, but no elevation in Drought or RecDrought was observed.

The most abundant cluster, Cluster 2, contained 30 genes separated into six sub-clusters. The first subcluster included two genes (*BMV5* and *bHLH102*) with the highest expression level in Drought (7- to 15-fold higher), and significantly induced expression in Cold (3- to 4-fold higher), and no elevated expression in recovery treatments. The second sub-cluster combined eight genes (*CBF1*, *PEI54*, *HXK1*, *bHLH43*, *bHLH79*, *WRKY42*, *PRP*, *GR-RBP*). These genes showed 3- to 9-fold upregulation in Cold. Of these, *WRKY42*, *CBF1*, and *PEI54* were significantly elevated in RecCold and RecDrought. In addition, four were downregulated in Drought and RecDrought: *PRP*, *HXK1*, *bHLH43*, and *bHLH79*. The third sub-cluster included eight genes (*AP-ERF-AP*, *EGASE11*, *CRK45*, *PME*, *DREB26*, *RHL*, *Hydrolase22*, and *CAU1*), which were significantly upregulated in Drought with 2- to 4-fold change, but most were not elevated in Cold. Four genes of the *bHLH* family composed the fourth sub-cluster and were characterized by decreased expression in most treatments: *bHLH12*, *bHLH21*, *bHLH45*, and *bHLH95*. *DHN3*, *POD73*, and *HXK3* combined in the fifth sub-cluster, with about a 2- to 3-fold greater expression under Cold and RecCold but very

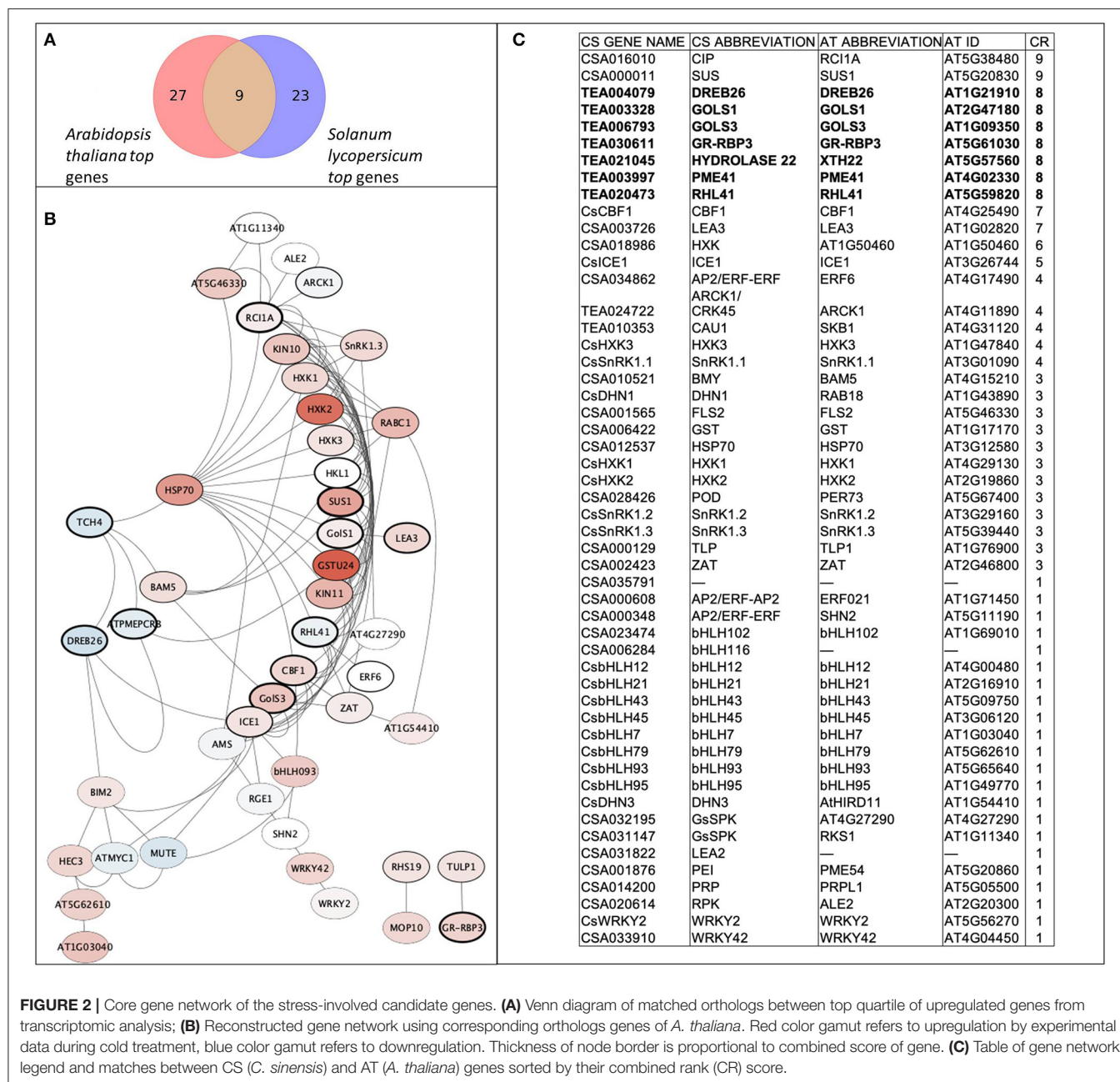


FIGURE 2 | Core gene network of the stress-involved candidate genes. **(A)** Venn diagram of matched orthologs between top quartile of upregulated genes from transcriptomic analysis; **(B)** Reconstructed gene network using corresponding orthologs genes of *A. thaliana*. Red color gamut refers to upregulation by experimental data during cold treatment, blue color gamut refers to downregulation. Thickness of node border is proportional to combined score of gene. **(C)** Table of gene network legend and matches between CS (*C. sinensis*) and AT (*A. thaliana*) genes sorted by their combined rank (CR) score.

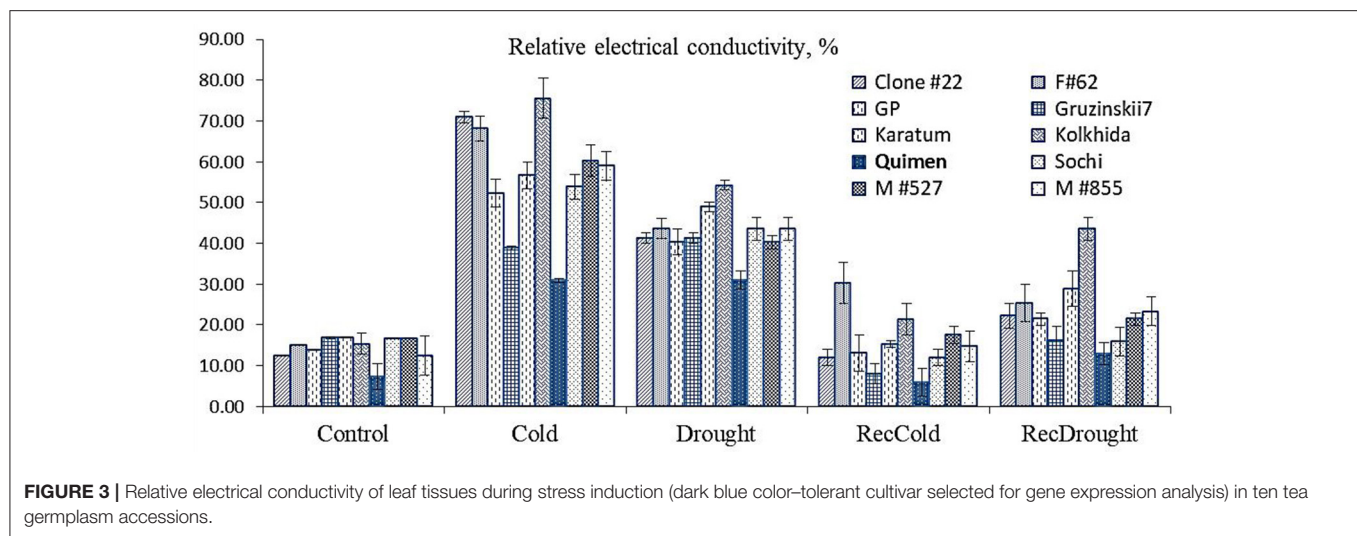
little expression in Drought and RecDrought. The last sub-cluster was formed by *ICE1*, *GOLS1*, *WRKY2*, and *ZAT* and showed 2-fold greater expression in Cold and Drought, as well as being slightly elevated in Recovery treatments.

In summary, the genes significantly upregulated in both Drought and Cold were *HSP70*, *SUS1*, *GST*, *DHN1*, *BMY5*, *bHLH102*, *GR-RBP3*, *ICE1*, *GOLS1*, and *GOLS3*, indicating that they may have important roles in both types of stress response. The genes that were specifically upregulated in Cold were *SnRK1.2*, *HXK1*, *HXL2*, *bHLH43*, *bHLH79*, *bHLH7*, and *bHLH93*. The genes that were specifically upregulated in Drought were *RHL41*, *CAU1*, and *Hydrolase22*. The transcripts of *CIP*

were mostly accumulated in RecCold and RecDrought, and the transcripts of *PME41* were mostly accumulated in RecDrought indicating the possibly important role of these two candidate genes in plant recovery after stress. Generally, the cold response was more active in our study than the drought response. More genes with the highest expression levels were induced in response to cold than to drought.

PCA Analyses and Correlations in Different Responses

Pair comparison of treatments showed that the gene data points were clearly distributed between the two principal components



Cold and RecCold. Most genes were densely grouped and showed similar expression profiles in Cold and RecCold, indicating a systemic response to cold stress. On the other hand, more genes were related to the principal component Cold. The RecCold cluster combined eight genes grouped distantly, which were strongly expressed in the recovery stage: *SnRK1.1*, *SnRK1.3*, *TLP*, *LEA2*, *LEA3*, *FLS2*, *EGase11*, and *CIP*. The genes *SnRK1.2*, *DHN1*, *GOLS3*, and *bHLH7* clustered distantly around the principal component Cold (Figure 5A).

The biplot Drought/RecDrought showed that most genes were densely grouped together with a similar expression pattern during both treatments. However, nine genes were distantly clustered around the RecDrought principal component: *CRK45*, *PME41*, *CBF1*, *CIP*, *PEI54*, *WRKY42*, *DREB26*, *ZAT*, and *ICE1*. Another eight genes were distantly clustered around the principal component Drought: *EGase11*, *RHL41*, *GOLS3*, *BMY5*, *bHLH102*, *DHN1*, *Hydrolase 22*, and *CAU1* (Figure 5B).

Finally, in the Cold/Drought biplot, most data points were clearly divided between the two principal components and showed the different characters of expression in the two stress responses. The two clusters with the greatest distances between Cold and Drought PCs were obtained. The first combined the six genes with the highest expression level in Drought: *GOLS3*, *BMY5*, *bHLH102*, *RHL41*, *CRK45*, and *GR-RBP3*. The second one combined the six genes with the highest expression level in Cold: *SnRK1.2*, *DHN1*, *FLS2*, *LEA2*, *SnRK1.1*, and *bHLH7*. Most of the other genes were also clearly divided between the principal components Cold and Drought (Figure 5C).

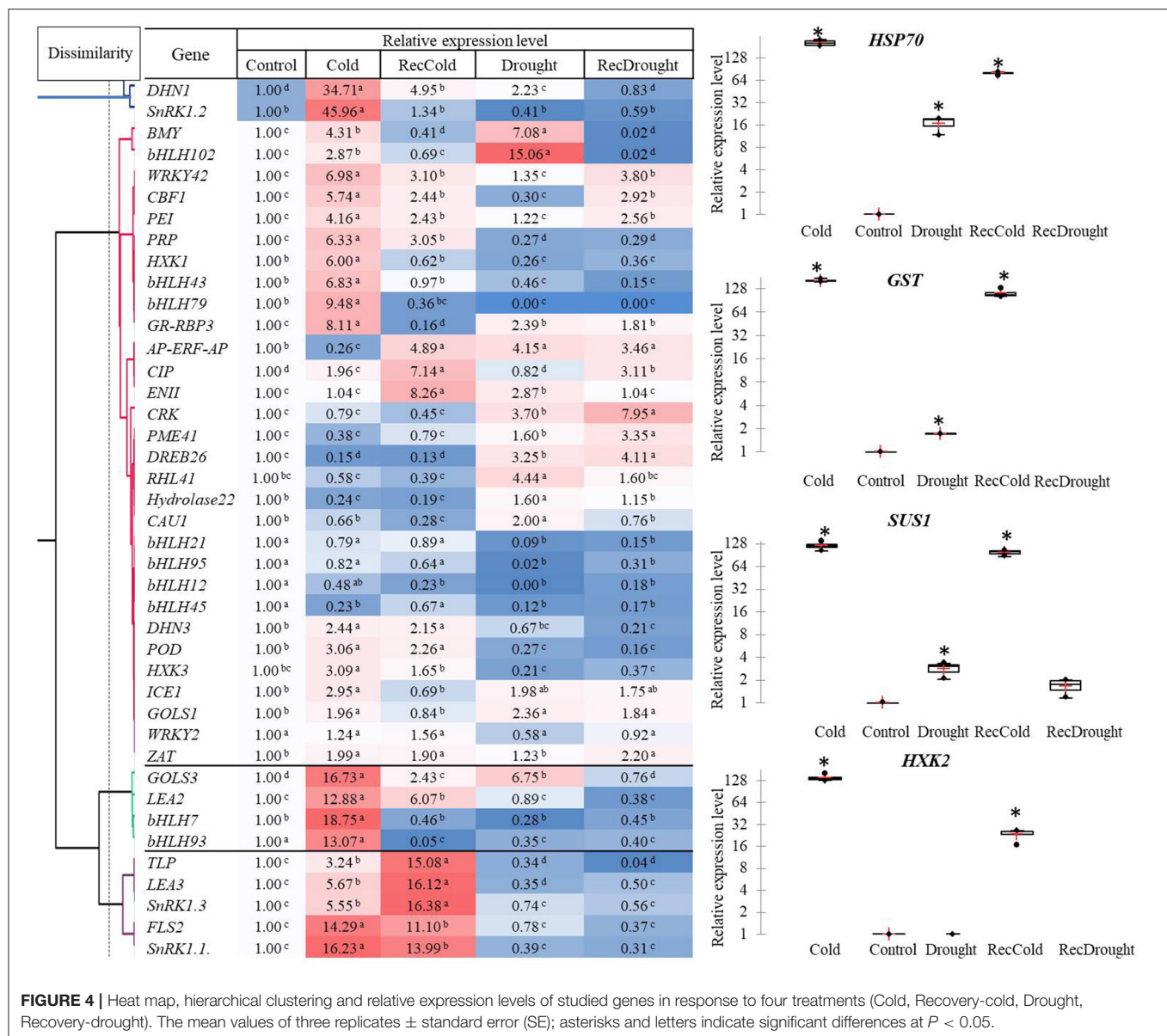
The correlation analyses of responses to Drought, Cold, RecDrought, and RecCold resulted in three large clusters of candidate genes (Figure 6). The first, the largest cluster, included 18 genes with the highest positive and significant correlations between the treatments. This cluster combined three main subclusters. The first included the genes *ICE1* and *bHLH7*, which had a high positive correlation between RecCold and RecDrought. The second sub-cluster combined four genes

that had a high positive correlation between Cold/RecCold and Drought/RecDrought: *POD73*, *bHLH79*, *AP-ERF-AP*, and *LEA3*. The third sub-cluster included genes with high positive correlations between Drought/Cold (*PEI54*, *SnRK1.2*, *SnRK1.3*, and *Hydrolase22*) and Drought/Recovery (*Hydrolase22*, *SnRK1.2*, *CRK45*, *BMY5*, and *bHLH93*).

The second large cluster combined nine genes. Of these, *DHN3* showed a positive correlation between Cold and Drought. *PRP* and *HXX1* showed a high positive correlation between Cold and RecCold. Three genes showed a strong negative correlation between Drought and RecDrought: *DHN1*, *CBF1*, and *GOLS1*. Additionally, *DHN1* was negatively correlated in Cold and RecCold; *CBF1* and *GOLS1* were negatively correlated in Drought and Cold.

The third big cluster combined 16 genes, divided into two big sub-clusters. One sub-cluster included eight genes, of which three showed a significant negative correlation between RecCold and RecDrought: *HXX2*, *DREB26*, and *bHLH45*. However, another three of these genes showed high positive correlations in Drought and RecDrought. Finally, the second small sub-cluster of Cluster 3 included six genes, of which four showed significant negative correlations between Cold and RecCold (*SnRK1.1* and *SUS1*) RecCold and RecDrought (*SUS1*), and Cold and Drought (*HXX3* and *CIP*).

In summary, the following genes were significantly positively correlated between Cold and Drought: *DHN3*, *bHLH79*, *PEI54*, *SnRK1.2*, *SnRK1.3*, and *Hydrolase22*. On the other hand, *CBF1*, *GOLS1*, *HXX2*, and *HXX3* showed significant negative correlations. Many genes were positively correlated between Drought and RecDrought, namely, *POD73*, *bHLH79*, *AP-ERF-ERF*, *LEA3*, *Hydrolase 22*, *SnRK1.2*, *CRK45*, *BMY5*, *bHLH93*, *bHLH95*, *DREB26*, and *HXX2*. Three genes showed negative correlations: *DHN1*, *CBF1*, and *GOLS1*. Six genes were positively correlated between Cold and RecCold: *bHLH79*, *AP-ERF-ERF*, *LEA3*, *PRP*, *HXX1*, and *TLP1*. Three genes were negatively correlated: *SnRK1.1*, *SUS1*, and *DHN1*. Finally, RecCold and RecDrought analyses resulted in four positively correlated genes



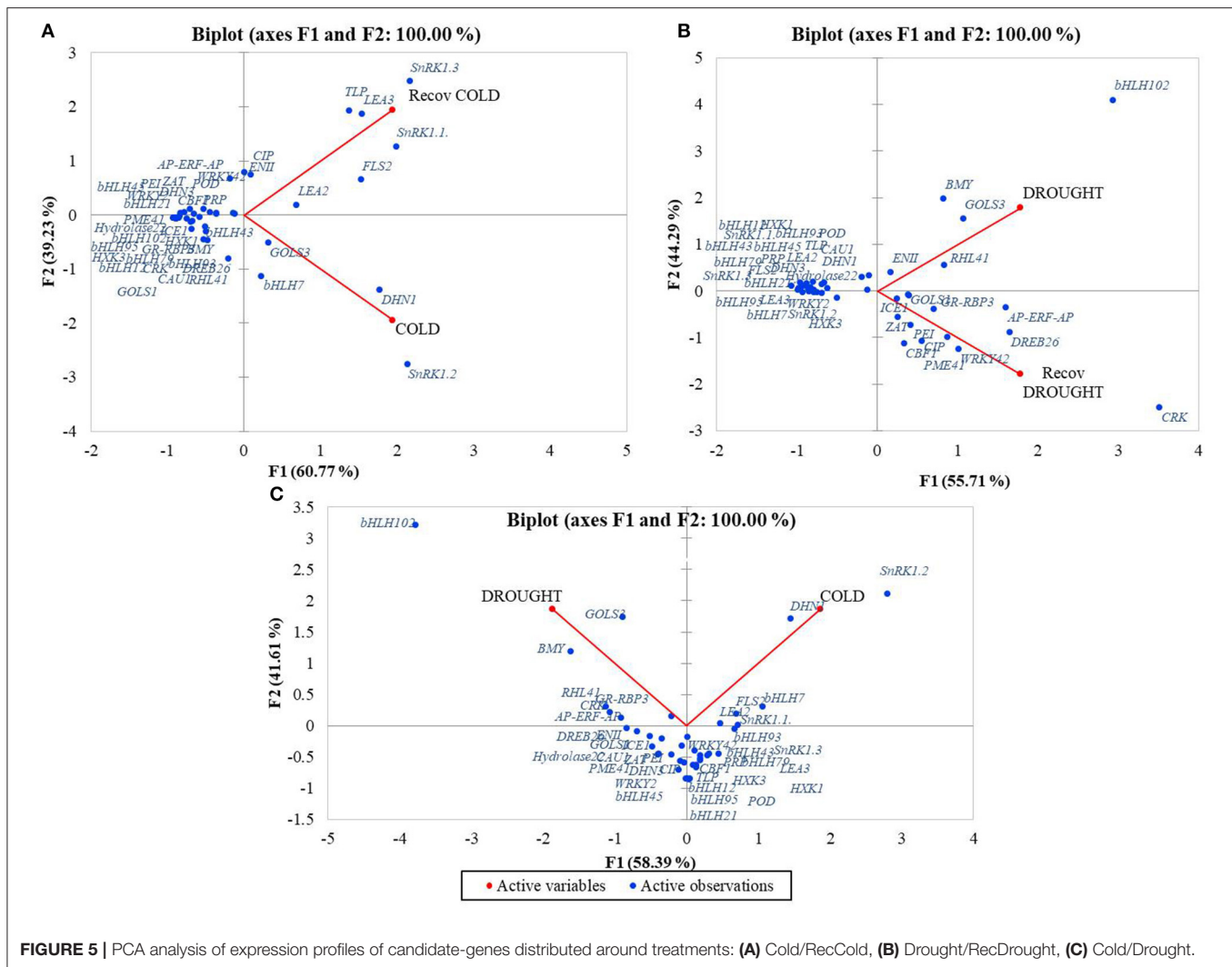
(ICE1, bHLH7, POD73, and CBF1) and four negatively correlated genes (bHLH45, DREB26, HXK2, and SUS1).

DISCUSSION

Reconstruction of the Cold Stress Response Gene Regulation Network in Tea Plant

To develop tolerant genotypes, breeders need reliable sets of informative genetic markers to select donors from germplasm collections. The homolog databases of candidate genes can be an efficient tool for finding these markers with *in silico* searches in model plant species. In our study, we used this approach to identify new possible candidate genes and their homologs in tea plants. We selected possible candidate genes

and built a core network for 42 genes with 111 edges between them, which indicates their tight interconnection. DREB26, GOLS1, GOLS3, GR-RBP3, Hydrolase22, PME41, and RHL41 are commonly found in *A. thaliana* and *S. lycopersicum*. It is known that the evolutionary distance between *A. thaliana* and *S. lycopersicum* is very similar to distance between *A. thaliana* and *C. sinensis* (timetree.org). Thus, we proceeded from the assumption that nine identified genes may have a similar role for *C. sinensis*. Based on the constructed gene network we revealed that eight of the nine candidate genes are linked to the main network of the stress response. So it can be suggested that they belongs to the core part of the stress response, their functions are evolutionarily conservative and these genes can be predicted for the other plant species based on interspecific analysis. The hypothesis of the strong upregulation of galactinol syntases (GOLS1 and GOLS3) and GR-RBP3 and DREB26 in



tea stress responses was confirmed experimentally in our study (Figure 4).

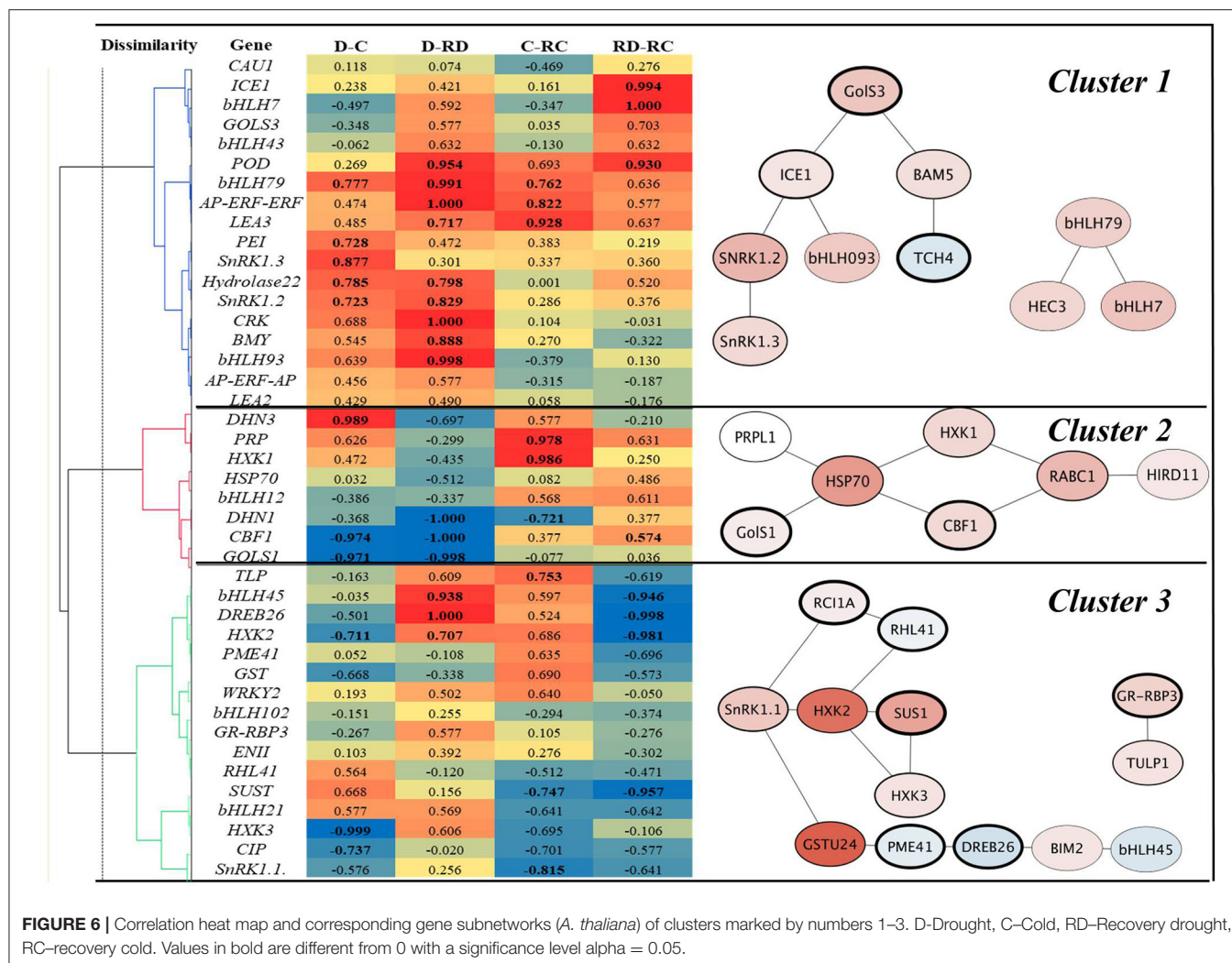
We combined bioinformatics and experimental approaches to test nine new candidate genes that could be relevant for different plant species. However, among their orthologs in *C. sinensis*, only *GOLS3* and *GR-RBP3* were found to be upregulated during cold treatment. On the other hand, *CIP*, which has the highest score according to bioinformatics data (Figure 2), is highly upregulated during recovery. This can indicate a large difference between woody crops and grasses in responses to stress. In addition, well-known regulators such as *DHN1*, *HXK1*, *PEI*, and *CBF1* were confirmed to be highly upregulated during cold treatment. Therefore, experimental testing of well-known regulators with their new target genes for particular genotypes may be a useful and iterative approach for evaluating complex regulatory networks of stress adaptation in plants.

The gene-regulatory networks for cold and drought response remain an open topic for investigation due to the complex nature of genetic interactions and their genotype-specific character. For example, the divergence and specialization of gene networks

involved in trichome development may be connected with the emergence of the plant taxa (Doroshkov et al., 2019). In our study, many regulators were connected to *HSP70* and tightly interconnected among each other. The *ICE1*-related *bHLH* cluster and *WRKY* factors were mostly upregulated, similarly to the *SnRK* complex, hexokinases, and galactinol synthases. However, *XTH22-PME41-DREB26*, and *bHLH12-bHLH45* were downregulated, which may indicate their coordinated repression.

Phenotypical Selection of Tea Under Cold and Drought Treatments

North-Western Caucasus in Russia is the one of the northernmost regions of commercial tea growing in the World. Tea plantations in the region are not of a large scale, but the climate here is colder that is why tea growth without chemical plant protection because there is no pest and diseases. Seeds of tea plant were introduced to Caucasus in nineteenth century from China, Japan, India, Sri Lanka and Indonesia and represent a wide range of hybrid genetic diversity. Domestication of the tea plant in the Caucasus occurred within 150 years,



during which the tea crop moved from the southern regions of Ozurgetti in Georgia (41°55'27" N, 41°59'24" E) to the Northern region in Maykop in Russia (44°36'40" N, 40°06'40" E) (Tuov and Ryndin, 2011). Tea breeding was conducted here from 1950th and as the result many local cultivars were developed, such as *Kolkhida*, *Qimen*, *Gruzinskii7*, *Karatum*, *Sochi*, and many others. Also the set of mutant forms such as M#527, M#855, F#62, Clone #22 and many others were developed by UV and chemical mutagenesis (Gvasaliya, 2015). The genotypes included in our study characterized by high yield and quality in the local conditions. Phenotyping of the tolerance was done using the common approach – the measurement of the relative electrical conductivity (see for example, Ban et al., 2017), that help to assess the electrolyte leakage caused by stress. The results confirmed that genotypes with large and thin leaf (for example, *Karatum*, *Kolkhida*) are less tolerant to cold and drought than the genotypes with small and thick leaf blades (such as *Qimen* and *Gruzinskii7*) (Figure 1). Our results on phenotypical evaluation correspond with the other studies on several plant species in which the drought-resistant genotypes showed tolerance to cold as well (Zheng et al., 2016; Lu et al., 2017; Li X. et al., 2019).

Relative Expression Levels of the Studied Genes in Response to Cold, Drought and Recovery

Genes Upregulated in Response to Both Cold and Drought

In the tolerant genotype the expression levels of the genes *HSP70*, *DHN1*, *GST*, *SUS1*, *bHLH102*, *BM5*, *GR-RBP3*, *ICE1*, *GOLS1*, and *GOLS3* were significantly higher in both Cold and Drought than in control, suggesting shared upstream pathways for signal transduction and regulation under these stimuli.

Among the nine *bHLH* genes included in this study, only *bHLH102* was increasingly expressed in both stress treatments, and we suppose that this new candidate gene can also be an important marker for abiotic stress tolerance in tea. In *A. thaliana*, this gene encodes positive brassinosteroid-signaling protein, and functional validation is necessary in tea plant.

The Hsp70s are highly conserved and widespread and important for protein folding, protein translocation, and the stress response in almost all subcellular compartments (Su and Li, 2008). The *HSP70* genes are upregulated in drought-tolerant Indian tea cultivars that are subjected to water stress

(Maritim et al., 2016). In our study, the highest level of expression for *HSP70* (several hundred-fold) was observed in all experimental treatments (Cold, RecCold, Drought, and RecDrought) (**Figure 3**); however, it was more actively induced by Cold compared to Drought, indicating its possible importance in preventing the dehydration of cell compartments during low temperatures.

Another gene that was upregulated in cold and drought was *GST*. GSTs are a superfamily of enzymes that are notable for their role in phase-II detoxification reactions of quenching reactive molecules by adding glutathione (GSH) and protecting the cell from oxidative damage (Kumar and Trivedi, 2018). In previous work, *GST* and *POD* were upregulated in a tolerant tea cultivar under cold stress (Li Y. et al., 2019), which corresponds with our results. However, these genes were more strongly induced by cold, and we suppose that the cold response is characterized by stronger ROS-scavenging activity than the drought response.

The next gene with a multi-fold change in Cold and RecCold and significant upregulation in Drought was *SUS1*. It encodes sucrose synthase (Sus), a key enzyme of sucrose metabolism. Previous studies reported that the transcription levels of *Sus1* increased after exposure to cold and drought (Dejardin et al., 1999; Stein and Granot, 2019). However, based on the expression profile of *SUS1*, we speculate that sucrose–raffinose conversion is more strongly induced by cold than by drought in tea plant. Also, the bulk degradation of sucrose into glucose and fructose maybe a strategy employed by tea plants to double its osmotic contribution in response to severe drought and cold stress (Zheng et al., 2016).

Another new gene that was significantly overexpressed in response to both drought and cold was *GR-RBP3*, a class-IV GRP (RBP), which is involved in alternative splicing, transcriptional regulation, and developmental processes (Czolpinska and Rurek, 2018). Some GRPs have been described as proteins that mainly enhance plant tolerance to low temperatures. Here, we suppose that they may also be an important genetic marker of both cold and drought tolerance, with a functional role in the tea plant that it is necessary to clarify.

One more gene that was significantly upregulated during drought and cold stress was *BM5*. BM5s degrade starch to soluble sugar, which leads to increased maltose, glucose, fructose, and sucrose levels after further conversion. We suggest that starch degradation is an important mechanism in tea, not only for cold tolerance but also for drought tolerance. This is consistent with the results recently published by Yue et al. (2019), who found that *BM5* genes contain many stress-related cis-acting elements, such as drought stress-related *ABRE*, *DRE1*, *MBS*, and *STRE*; cold stress-related *LTR*; and stress phytohormone-related *ERE* and *TCA*. Taken together, these results suggest that *BM5* genes are involved in the response of tea plants to multiple challenging environmental conditions and may be an important marker for the tea plant.

The last two genes that feature strong upregulation in response to drought and cold are *GolS1* and *GolS3*. GolS is a key enzyme in the synthesis of raffinose family oligosaccharides that function as osmoprotectants in plant cells. *GolS1*- or *GolS2*-overexpressing *Arabidopsis* has high intracellular levels of galactinol and raffinose

in transgenic plants, which correlates with increased tolerance to drought and chilling stress (Panikulangara et al., 2004; Nishizawa et al., 2008; Li Y. et al., 2019). Our results support these findings and confirm that the mechanism of protecting salicylate from attack by hydroxyl radicals mediated by galactinol and raffinose is important for drought and cold defense.

Genes Specifically Upregulated in Drought

The genes specifically upregulated to a higher level in Drought were *RHL41*, *CAU1*, *Hydrolase22*, *CRK45*, *PME41* which suggests that these genes are conservative and may play vital specific roles in response to drought stress.

RHL41, which relates to the zinc-finger protein *Zat12*, is a representative of the small group of genes that respond similarly to many different environmental stresses (Iida et al., 2000; Davletova et al., 2005). A recent study of transgenic plants suggested that *Zat12* plays a role in different stress responses in *Arabidopsis* (Rizhsky and Liang, 2004; Vogel et al., 2005). Some authors have reported that *Zat12* acts as a suppressor of *CBF* transcription (Davletova et al., 2005; Vogel et al., 2005). We observed increased accumulation of *Zat12* (*RHL41*) transcripts during drought, indicating that this gene may have a specific function for drought stress responses in tea plant.

CAU1 encodes an H4R3sme2-type histone methylase and acts as an immediate upstream suppressor of the *CAS* gene (encoding a putative Ca^{2+} binding protein that is proposed to be an external Ca^{2+} sensor). Elevated extracellular calcium decreases *CAU1* protein levels and consequently the methylation level of H4R3sme2 in the *CAS* chromatin, thus derepressing *CAS* expression to close stomata (Fu et al., 2013). Our results indicate the specific activation of *CAU1* under drought. It may be that stomata closure mediated by *CAU1* is an important mechanism of defense against drought in tea plant. This corresponds with previous studies that have reported increased drought tolerance and stomatal closure in *cau1* mutants of *Arabidopsis* (Fu et al., 2013).

Hydrolase22 was also specifically upregulated during drought stress. This gene encodes proteins that maintain the plasticity of the cell wall and increase its thickness by reinforcing the secondary wall with hemicellulose and lignin deposition (Le Gall et al., 2015). We thus consider that the adjustment to the cell wall mediated by this enzyme is an important mechanism in adaptation to drought in tea plant.

Different families of protein kinase had positive regulatory roles in responding to drought stress in tea plant, leading to maintain homeostasis of drought stress and water signal transduction (Liu S.-C. et al., 2016). Our result showed that *CRK45* was upregulated in Drought and RecDrought but not in Cold. It is a member of the membrane-anchored receptor-like protein kinases (RLKs), which recognize extracellular signals at the surface of the cell and activate a downstream signaling pathway by phosphorylating specific target proteins (Tanaka et al., 2012). CRKs make up a large subgroup of the RLKs family and play important roles in plant growth, development, and the stress response (Afzal et al., 2008; Wrzaczek et al., 2010; Tanaka et al., 2012). Thus, negative ABA-signaling mediated by *CRK45*

may play a specific and important role in the drought response of the tea plant.

Increased *PME41* expression was observed in the tea plant in Drought and RecDrought but not in Cold. PME participates in pectin remodeling, which keeps cells from separating, maintains plasma membrane integrity, and prevents cellular leakage. However, distinct genotype-, species- or tissue-dependent mechanisms of temperature control of PME activity have been found (Le Gall et al., 2015). For example, the overexpression of *Arabidopsis* *PME5* and *PMEI3* resulted in softer and harder shoot apical meristem cell walls, respectively (Peaucelle et al., 2011). We suppose that the mechanism of demethylesterification of pectin may be more important for drought defense rather than for cold defense in the tea plant. Further studies with more cultivars are necessary to check the involvement of *PME41* in the cold response of the tea plant.

Genes Specifically Upregulated in Cold

The genes specifically upregulated to a higher level in Cold were *SnRK1.2*, *HXK1*, *HXK2*, *bHLH43*, *bHLH79*, *bHLH7*, and *bHLH93*, which suggests that these genes are conservative and may play vital specific roles in response to cold stress.

HXKs phosphorylate glucose and fructose and participate in sugar signal transduction by modulating the abundances of diverse gene transcripts and integrating stress response substrates, including ABA and ethylene (Yue et al., 2015). In cold stress, HXKs are more induced in tolerant tea cultivars than in susceptible ones (Yue et al., 2015; Li Y. et al., 2019), which is consistent with our results. Another signaling intermediate, *SnRK1*, is involved in Suc, G6P, and T6P sensing and plays an important role in the plant response to sugar starvation (Wang Y. et al., 2019). Yue et al. (2015) found that *CsSnRK1.2* was induced by cold in the tea plant, whereas *CsSnRK1.1* was not elevated, and *CsSnRK1.3* was sharply suppressed. In our study, these three genes were activated in Cold and RecCold, but none was induced in Drought. These results indicate that sugar signal transduction and phosphorylation are more important defense mechanisms for cold tolerance in tea plant than for drought tolerance. However, more genotypes must be examined to confirm this conclusion.

Among the nine studied *bHLH* genes, some were specifically upregulated in response to cold stress. Cui et al. (2018) studied the *bHLH* family and proposed the following stress-related members in tea plant: *CsbHLH007*, *CsbHLH012*, *CsbHLH021*, *CsbHLH043*, *CsbHLH045* (ortholog of *ICE2*), *CsbHLH079*, *CsbHLH093*, and *CsbHLH095*. In our study, some of the genes were specifically upregulated in Cold, namely, *bHLH93*, *bHLH79*, *bHLH43*, and *bHLH7*, and these may play an important specific role in cold defense in the tea plant. We also observed that *CsbHLH012*, *CsbHLH021*, *CsbHLH045*, and *CsbHLH095* were downregulated in tea in Cold and/or Drought or did not differ from the control (Figure 4). This contradiction with Cui et al. (2018) can be explained by the variance in stress conditions: we evaluated long-term stress responses, whereas they evaluated 24 h stress induction (Cui et al., 2018). It may be that the mentioned TFs are more strongly induced by short-term cold stress.

Other genes that were upregulated in both Cold and RecCold were *WRKY42*, *ZAT30*, *POD73*, *LEA2*, *LEA3*, *TLP1*, and *FLS2*. Among them, the LEA proteins protect plant metabolism against abiotic stresses, marshaling properties that include antioxidant activity, metal ion binding, membrane and protein stabilization, hydration buffering, and DNA and RNA interactions (Chen et al., 2019). They also play an important role in stress acclimation (Ling et al., 2016). Liu Y. et al. (2016) investigated a maize *LEA3* gene expressed in *E. coli* and reported enhanced tolerance to low temperature. In rice, the *LEA2*, *LEA3*, and *DHN* groups have been found to show strong responses to osmotic stress (Yu et al., 2016). Our results on the tea plant showed no enhanced expression of *LEA2* and *LEA3* in Drought or RecDrought; however, Cold and RecCold greatly induced expression of both genes, indicating that these two genes can have specific functions on regulating cold tolerance in the tea plant.

WRKY42 and *ZAT30* (*CCCH*) are zinc-finger proteins involved in the ABA-mediated stress response. We observed specific upregulation of *WRKY42* and *ZAT30* during cold and recovery in the tea plant. The *WRKY* genes are involved in stress and hormone signaling (Phukan et al., 2016; Jiang et al., 2017) during the drought stress response (Wang et al., 2016) and cold response (Samarina et al., 2020) in the tea plant. *ZAT30* (*CCCH*) is a zinc finger protein that is involved in developmental processes, responses to cold and osmotic stress (Pi et al., 2018), and participates in signal transduction. Both of these TF families are of particular interest, as they are involved in various biotic/abiotic stress responses and in developmental/physiological processes (Jiang et al., 2015). Maybe further studies are needed to confirm the role of the both genes in tea plant.

TLP is another gene with no elevated expression in Drought but greatly induced in Cold and RecCold. It is a member of the *TLPs*, made up of five pathogenesis-related proteins that are responsive to biotic and abiotic stress. The previous results indicate potential applications of *TLP* for crop improvement through a genetic transformation with applications in both biotic and abiotic stress protection, with strong evidence for a role in the crosstalk between the stress types. Transgenic plants that overexpress the *TLP* gene in different plant crops showed resistance to pathogens and tolerance to salinity and drought (Jesus-Pires et al., 2019). Our data confirm the possibly important role of *TLP1* in the cold stress response and in recovery in tea plant.

FLS2, which encodes receptor-like protein kinase, was also highly upregulated during Cold and RecCold in our study. *FLS2* is representative of the *RLK* family, playing an important role in mediating early flagellin signaling (Lu et al., 2010) upregulated in the cold response and recovery to stress. Our results are consistent with those of Li Y. et al. (2019), who found that *FLS2* exhibited a higher level of expression in tolerant tea cultivars with many-fold change under cold. We also speculate that plant-pathogen-related immunity mediated by *FLS2* may be important specifically for cold tolerance rather than for drought tolerance.

PRP and *GRP*, are covalently linked with pectin or hemicellulose and thus contribute to the strengthening of the cell wall in response to abiotic stress (Hijazi et al., 2014). In our

study, significantly elevated expression of both *GRP* and *PRP* was observed in response to cold stress, which could indicate that the cell wall strengthening through pectin remodeling may be an important mechanism of tea plant cold tolerance. Earlier investigations also showed that one of the specific mechanisms of cold response in plants is the strengthening the cell wall, in contrast to the drought response (Beck et al., 2007).

Genes Upregulated at Recovery Treatments

Interestingly, some stress-inducible genes were seen to have higher transcript abundance during the recovery stages than had been seen in previous stress treatments. In RecCold, these genes were *AP-ERF-AP*, *LEA3*, *GST*, *SnRK1.3*, *SUS1*, *CIP*, *EGASE11*, and *TLP1*. In RecDrought, they were *WRKY42*, *CBF1*, *DREB26*, *ZAT*, *PEI54*, *CIP*, *CRK45*, and *PME41*. This indicates that the recovery of the tea plant after stress is a complex process and is important for defensive responses, and its regulation pathways differ from those for Cold and Drought. Moreover, we observed that RecCold and RecDrought produce very different responses, and *CIP* was the only gene upregulated in both recovery treatments in the tea plant. *CIP* belongs to the dehydrin family, and functional predictions suggest that this protein protects the membranes and prevents macromolecular coagulation or sequestration of calcium ions by association or disassociation with membrane under low-temperature conditions (Liu et al., 2006). We conclude that this gene may have a specific important function in recovery in tea plant.

Among the upregulated transcription factors, *AP-ERF-AP*, *DREB26*, and *CBF1* are representatives of the *AP2/ERF* family and mediate the transcriptional regulation of osmotic stress-responsive genes (Licausi et al., 2013; Parmar et al., 2019). Our results demonstrate that these genes are not only involved in the stress response but also in the recovery of the tea plant after cold and drought stresses. Previous gene expression studies have reported that most *AP2/ERFs* are expressed at low levels under normal conditions, but their expression can be induced or repressed at certain growth stages by hormones and stress stimuli (Xie et al., 2019). The *DREB* subfamily may be a key candidate for future exploration of a means to enhance drought and cold tolerance in tea (Ban et al., 2017; Parmar et al., 2019; Wang et al., 2019; Hu et al., 2020). It has been classified into six subgroups (A1–A6) (Sakuma et al., 2002). Among these, *DREB08* and *DREB26*, the A5 subgroup, encode repressor proteins inhibiting the expression of other *DREB* TFs (Dong and Liu, 2010). This means that they can suppress defense and stress-inducible genes in the absence of stress. In our study, increased expression of *DREB26* was observed during drought stress and also during recovery. This partly contradicts previous results that indicated that transcription levels of *DREB26* were hardly changed under drought and cold in *Vitis vinifera* (Zhao et al., 2014). Further studies of this gene in the tea plant are necessary for a comprehensive understanding of its role in stress responses.

Out of the other studied genes, with the pronounced expression profile during Recovery two genes are related to the cell wall remodeling, these are *EGases* and *PEI*. *EGases* are important cell wall-related proteins that modulate cell wall

extensibility, which mediates cell enlargement and expansion. The *EGase11* gene in the tea plant was significantly upregulated in RecCold, Drought, and RecDrought, indicating cellulase growing activity. This result is not easy to explain, and further investigation is necessary. Earlier studies of these genes reported that increased hydrolases activity is evidence of cell wall degradation (Le Gall et al., 2015). *PEIs* are invertase inhibitor-related proteins and play an important role in the regulation of metabolic enzymes and viscoelastic properties of the cell wall (Wu et al., 2010). In our study, the elevated expression of *PEI54* observed in Cold, RecCold, and Rec Drought indicates that pectin methylesterification in cell walls is activated in these stress treatments. These results showed that the cell wall remodeling activity is enhanced not only during the stress response but also during the recovery in tea plant. The genes encoding the cell wall remodeling enzymes can be further studied more comprehensively in tea plant as they might play a very important role in the responses to abiotic stresses and recovery after stress.

Correlations in Different Responses

Based on expression profiles we tried to find correlations between responses to drought and cold. Highly correlated gene modules with specific expression patterns can help illustrating the framework of stress transcriptome. This analysis provides evidences about common and unique stress mechanism components under cold and drought stress in *C. sinensis*. In *A. thaliana* gene co-expression network analysis revealed 21 and 16 highly inter-correlated gene modules with specific expression profiles under drought and cold stress respectively (Sharma et al., 2018). In oil palm the significant correlations were found between cold-responsive genes and physiological parameters that helped to better understand the regulation networks (Li J. et al., 2019).

In our study, six genes (*DHN3*, *bHLH79*, *PEI54*, *SnRK1.2*, *SnRK1.3*, and *Hydrolase22*) were correlated positively and four genes (*CBF1*, *GOLS1*, *HXX2*, and *HXX3*) were correlated negatively in response to Cold and Drought. This indicated that the mentioned genes have the similar expression character during cold and drought. Under drought induction in tea plant we found twelve genes (*POD73*, *bHLH79*, *AP-ERF-ERF*, *LEA3*, *Hydrolase 22*, *SnRK1.2*, *CRK45*, *BMV5*, *bHLH93*, *bHLH95*, *DREB26*, and *HXX2*) that were positively correlated and three genes (*DHN1*, *CBF1*, and *GOLS1*) that were negatively correlated between Drought and RecDrought. On the other hand, under cold induction in tea plant we found six genes (*bHLH79*, *AP-ERF-ERF*, *LEA3*, *PRP*, *HXX1*, and *TLP1*) that were positively correlated and three genes (*SnRK1.1*, *SUS1*, and *DHN1*) that were negatively correlated between Cold and RecCold. Based on these results it can be speculated that recovery stage after drought is more similar to Drought response than RecCold-to-Cold response.

In general, our results showed that more genes were activated in response to cold rather than drought in tea plant. These results corresponds with the transcriptomic studies reported that much more DEGs were upregulated under cold rather than drought in tea plant (Zheng et al., 2016), apple (Li X. et al., 2019) and in maize (Lu et al., 2017). Cold induces an extensive activation of transcription, drought stress, however, induced

fewer transcriptional changes (only 15% as many), than cold in maize (Lu et al., 2017) suggesting that the more sensitive response to cold rather than drought would be a conserved mechanism in many plant species.

In other studies, an overlap between the expression patterns of stress-responsive genes in several plant species was observed after drought and cold stress induction (Li X. et al., 2019). In apple they found evidence of crosstalk between drought and cold stress signaling, with 377 commonly upregulated and 211 commonly downregulated genes (Li X. et al., 2019). In tomato, only about 10% of the drought-inducible genes were also induced by cold indicating different molecular strategies in their reaction to the two stresses (Zhou et al., 2019). In maize, only 194 DEGs were shared in cold and drought and, nearly 90% among them are regulated in a similar manner by both stresses, indicating that there is a shared network to regulate the cold and drought induced responses (Lu et al., 2017). On the other hand, specific regulations in response to cold or drought were also clearly visible in these crops. Nevertheless, in some plant species, the induction of cold resistance also promotes drought resistance and high-salinity tolerance, which is consistent with an increase in the levels of osmo-regulatory compounds and antioxidant enzyme activities (Hossain et al., 2013).

The effects of drought and cold reported here have arisen from a limited range of potential types and severities of stress. A greater range of treatments for (e.g., timing, severity, frequency) need to be examined in future studies to provide more clues for understanding the adaptation and tolerance mechanisms in tea plant.

CONCLUSION

Using an *in silico* approach combined with an experimental approach, we confirmed the involvement of the nine new genes in the cold and/or drought response of tea plant: *GOLS1*, *GOLS3*, *GR-RBP3*, *HYDROLASE22*, *RHL41*, *CAU1*, *PME41*, *DREB26*, and *CRK45*. We hypothesized that many genes have similar expression profiles between the cold and drought responses of the tea plant. However, of 45 genes studied, only ten were significantly upregulated in response to both cold and drought: *HSP70*, *GST*, *SUS1*, *DHN1*, *BMV*, *bHLH102*, *GR-RBP3*, *ICE1*, *GOLS1*, and *GOLS3*. These genes can be considered as genes of non-specific stress response. *SnRK1.2*, *HXK1/2*, and *bHLH7/43/79/93* were upregulated in response to cold only, and the expression levels of *RHL41*, *CAU1*, and *Hydrolase22* were increased in the drought response. Interestingly, we found that the expression of *CIP* was higher in the recovery stage of both stresses, indicating its potentially important role in plant recovery after stress. In addition, some genes, such as *DHN3*,

bHLH79, *PEI54*, *SnRK1.2*, *SnRK1.3*, and *Hydrolase22*, were significantly positively correlated between the cold and drought responses. *CBF1*, *GOLS1*, *HXK2*, and *HXK3*, by contrast, showed significantly negative correlations between the cold and drought responses. Because overexpression of many new candidate genes can confer stress tolerance, these proteins may play a promising role in agriculture in the context of plant genetic engineering. The isolation, cloning, characterization, and functional validation of novel candidate genes in response to diverse stress conditions are expected to be growth areas of research in coming years. In addition, the identification of the interaction partners of these proteins and the factors affecting these interactions is necessary to understand their role in conferring protection against different stress conditions in tea plants. These results provide valuable information and robust candidate genes for future functional analyses to improve the stress tolerance of the tea plant.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

AUTHOR CONTRIBUTIONS

LS planned and conducted experimental part of the manuscript, analyzed data, and wrote the manuscript. AB and AD performed bioinformatics part of the manuscript. LM participated in statistical analyses and critically reviewed the manuscript. AM, RR, and NK performed phenotyping and gene expression analyses. KM helped in gene-expression protocols improvement. VM, AR, WT, EX, and YO are scientific consultants and critically revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was conducted under financial support of the Russian Science Foundation (Project # 18-76-10001). The plant material for this study was provided by the germplasm conservation program # 0683-2019-0002 of the Ministry of Science and Higher Education of the Russian Federation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.611283/full#supplementary-material>

REFERENCES

- Afzal, A. J., Wood, A. J., and Lightfoot, D. A. (2008). Plant receptor-like serine threonine kinases: roles in signaling and plant defense. *Mol. Plant Microbe Interact.* 21, 507–517. doi: 10.1094/MPMI-21-5-0507
- Amrine, K. C. H., Blanco-Ulate, B., and Cantu, D. (2015). Discovery of core biotic stress responsive genes in *Arabidopsis* by weighted gene co-expression network analysis. *PLoS ONE* 10:e0118731. doi: 10.1371/journal.pone.0118731
- Bajji, M., Kinet, J.-M., and Lutts, S. (2001). The use of the electrolyte leakage method for assessing cell membrane stability as a water stress tolerance test in durum wheat. *Plant Growth Regul.* 36, 61–70. doi: 10.1023/A:1014732714549

- Ban, Q., Wang, X., Pan, C., Wang, Y., Kong, L., Jiang, H., et al. (2017). Comparative analysis of the response and gene regulation in cold resistant and sensitive tea plants. *PLoS ONE* 12:e0188514. doi: 10.1371/journal.pone.0188514
- Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, 991–995. doi: 10.1093/nar/gks1193
- Beck, E. H., Fetting, S., Knake, C., Hartig, K., and Bhattacharai, T. (2007). Specific and unspecific responses of plants to cold and drought stress. *J. Biosci.* 32, 501–510. doi: 10.1007/s12038-007-0049-5
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi: 10.1186/1471-2105-10-421
- Cao, H., Wang, L., Yue, C., Hao, X., Wang, X., and Yang, Y. (2015). Isolation and expression analysis of 18 *CsbZIP* genes implicated in abiotic stress responses in the tea plant (*Camellia sinensis*). *Plant Physiol. Biochem.* 97, 432–442. doi: 10.1016/j.plaphy.2015.10.030
- Chamani Mohasses, F., Solouki, M., Ghareyazie, B., Fahmideh, L., and Mohsenpour, M. (2020). Correlation between gene expression levels under drought stress and synonymous codon usage in rice plant by *in-silico* study. *PLoS ONE* 15:e0237334. doi: 10.1371/journal.pone.0237334
- Chaves, M. M., Maroco, J. P., and Pereira, J. S. (2003). Understanding plant responses to drought—from genes to the whole plant. *Funct. Plant Biol.* 30, 239–264. doi: 10.1071/FP02076
- Chen, J., Gao, T., Wan, S., Zhang, Y., Yang, J., Yu, Y., et al. (2018). Genome-wide identification, classification and expression analysis of the hsp gene superfamily in tea plant (*Camellia sinensis*). *Int. J. Mol. Sci.* 19:2633. doi: 10.3390/ijms19092633
- Chen, Y., Li, C., Zhang, B., Yi, J., Yang, Y., Kong, C., et al. (2019). The role of the late embryogenesis-abundant (*LEA*) protein family in development and the abiotic stress response: a comprehensive expression analysis of potato (*Solanum tuberosum*). *Genes* 10:148. doi: 10.3390/genes10020148
- Cui, X., Wang, Y.-X., Liu, Z.-W., Wang, W.-L., Li, H., and Zhuang, J. (2018). Transcriptome-wide identification and expression profile analysis of the *bHLH* family genes in *Camellia sinensis*. *Funct. Integr. Genomics* 18, 489–503. doi: 10.1007/s10142-018-0608-x
- Czoplinka, M., and Rurek, M. (2018). Plant Glycine-Rich Proteins in stress response: an emerging, still prospective story. *Front. Plant Sci.* 9:302. doi: 10.3389/fpls.2018.00302
- Davletova, S., Schlauch, K., Coutu, J., and Mittler, R. (2005). The zinc-finger protein *Zat12* plays a central role in reactive oxygen and abiotic stress signaling in arabidopsis. *Plant Physiol.* 139, 847–856. doi: 10.1104/pp.105.068254
- Dejardin, A., Sokolov, L. N., and Kleczkowski, L. A. (1999). Sugar/osmoticum levels modulate differential abscisic acid-independent expression of two stress-responsive sucrose synthase genes in Arabidopsis. *Biochem. J.* 344, 503–509. doi: 10.1042/bj3440503
- Dong, C. J., and Liu, J. Y. (2010). The arabidopsis EAR-motif-containing protein RAP2.1 functions as an active transcriptional repressor to keep stress responses under tight control. *BMC Plant Biol.* 10:47. doi: 10.1186/1471-2229-10-47
- Doroshkov, A. V., Konstantinov, D. K., Afonnikov, D. A., and Gunbin, K. V. (2019). The evolution of gene regulatory networks controlling *Arabidopsis thaliana* L. trichome development. *BMC Plant Biol.* 19:53. doi: 10.1186/s12870-019-1640-2
- Doyle, J. J., and Doyle, J. L. (1991). Isolation of plant DNA from fresh tissue. *Focus* 12, 13–15.
- Fu, Y. L., Zhang, G. B., Lv, X. F., Guan, Y., Yi, H. Y., and Gong, J. M. (2013). *Arabidopsis* histone methylase *CAU1/PRMT5/SKB1* acts as an epigenetic suppressor of the calcium signaling gene *cas* to mediate stomatal closure in response to extracellular calcium. *Plant Cell.* 25, 2878–2891. doi: 10.1105/tpc.113.113886
- Gvasaliya, M. V. (2015). *Spontaneous and induced cultivars and forms of tea (Camellia sinensis (L.) Kuntze) in the humid subtropics of Russia and Georgia: prospects for their cultivation and in vitro conservation* (PhD thesis). Krasnodar: Kuban State Agrarian University. 159p.
- Hao, X., Horvath, D. P., Chao, W. S., Yang, Y., Wang, X., and Xiao, B. (2014). Identification and evaluation of reliable reference genes for quantitative real-time PCR analysis in tea plant (*Camellia sinensis* (L.) O. Kuntze). *Int. J. Mol. Sci.* 15, 22155–22172. doi: 10.3390/ijms151222155
- Hao, X., Wang, L., Zeng, J., Yang, Y., and Wang, X. (2018). “Response and adaptation mechanisms of tea plant to low-temperature stress,” in *Stress Physiology of Tea in the Face of Climate Change*, eds W.-Y. Han, L. Xin, and A. Golam Jalal (Singapore: Springer Nature Pte Ltd.), 39–61. doi: 10.1007/978-981-13-2140-5_3
- Hijazi, M., Velasquez, S. M., Jamet, E., Estevez, J. M., and Albenne, C. (2014). An update on post-translational modifications of hydroxyproline-rich glycoproteins: toward a model highlighting their contribution to plant cell wall architecture. *Front. Plant Sci.* 5:395. doi: 10.3389/fpls.2014.00395
- Hossain, M. A., Mostofa, M. G., and Fujita, M. (2013). Cross protection by cold-shock to salinity and drought stress-induced oxidative stress in mustard (*Brassica campestris* L.) seedlings. *Mol. Plant Breed.* 4, 50–70. doi: 10.5376/mpb.2013.04.0007
- Hu, Z., Ban, Q., Hao, J., Zhu, X., Cheng, Y., Mao, J., et al. (2020). Genome-wide characterization of the C-repeat binding factor (*CBF*) Gene family involved in the response to abiotic stresses in tea plant (*Camellia sinensis*). *Front. Plant Sci.* 11:921. doi: 10.3389/fpls.2020.00921
- Iida, A., Kazuoka, T., Torikai, S., Kikuchi, H., and Oeda, K. (2000). A zinc finger protein *RHL1* mediates the light acclimatization response in arabidopsis. *Plant J.* 24, 191–203. doi: 10.1046/j.1365-313x.2000.00864.x
- Jesus-Pires, C., Ferreira-Neto, J. R. C., Bezerra-Neto, J. P., Kido, E. A., de Oliveira Silva, R. L., Pandolfi, V., et al. (2019). Plant thaumatin-like proteins: function, evolution and biotechnological applications. *Curr. Protein Peptide Sci.* 21, 36–51. doi: 10.2174/1389203720666190318164905
- Jiang, J., Ma, S., Ye, N., Jiang, M., Cao, J., and Zhang, J. (2017). *WRKY* transcription factors in plant responses to stresses. *J. Integr. Plant Biol.* 59, 86–101. doi: 10.1111/jipb.12513
- Jiang, Y., Wang, Y., Pang, W., Chen, L., Sun, H., Liang, Y., et al. (2015). Essential protein identification based on essential protein-protein interaction prediction by integrated edge weights. *Methods* 83, 51–62. doi: 10.1016/j.ymeth.2015.04.013
- Kumar, S., and Trivedi, P. K. (2018). Glutathione S-Transferases: role in combating abiotic stresses including arsenic detoxification in plants. *Front. Plant Sci.* 9:751. doi: 10.3389/fpls.2018.00751
- Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* 40, 1202–1210. doi: 10.1093/nar/gkr1090
- Le Gall, H., Philippe, F., Doman, J. M., Gillet, F., Pelloux, J., and Rayon, C. (2015). Cell wall metabolism in response to abiotic stress. *Plants* 4, 112–166. doi: 10.3390/plants4010112
- Li, J., Yang, Y., Iqbal, A., Qadri, R., Shi, P., Wang, Y., et al. (2019). Correlation analysis of cold-related gene expression with physiological and biochemical indicators under cold stress in oil palm. *PLoS ONE* 14:e0225768. doi: 10.1371/journal.pone.0225768
- Li, X., Li, M., Zhou, B., Yang, Y., Wei, Q., and Zhang, J. (2019). Transcriptome analysis provides insights into the stress response crosstalk in apple (*Malus × domestica*) subjected to drought, cold and high salinity. *Sci. Rep.* 9:9071. doi: 10.1038/s41598-019-45266-0
- Li, Y., Wang, X., Ban, Q., Zhu, X., Jiang, C., Wei, C., et al. (2019). Comparative transcriptomic analysis reveals gene expression associated with cold adaptation in the tea plant *Camellia sinensis*. *BMC Genomics* 20:624. doi: 10.1186/s12864-019-5988-3
- Licausi, F., Ohme-Takagi, M., and Perata, P. (2013). *APETALA2/ethylene* responsive factor (*AP2/ERF*) transcription factors: mediators of stress responses and developmental programs. *New Phytol.* 199, 639–649. doi: 10.1111/nph.12291
- Ling, H., Zeng, X., and Guo, S. (2016). Functional insights into the late embryogenesis abundant (*LEA*) protein family from *Dendrobium officinale* (*Orchidaceae*) using an *Escherichia coli* system. *Sci. Rep.* 6:39693. doi: 10.1038/srep39693
- Liu, H., Yu, C., Li, H., Ouyang, B., Wang, T., Zhang, J., et al. (2015). Overexpression of *ShDHN*, a dehydrin gene from *Solanum habrochaites* enhances tolerance to multiple abiotic stresses in tomato. *Plant Sci.* 231, 198–211. doi: 10.1016/j.plantsci.2014.12.006
- Liu, M., Lu, C., Shen, X., and Yin, W. (2006). Characterization and function analysis of a cold-induced *AmCIP* gene encoding a dehydrin-like protein in *Ammopiptanthus mongolicus*. *DNA Sequence* 17, 342–349. doi: 10.1080/10425170600908484

- Liu, S.-C., Jin, J.-Q., Ma, J.-Q., Yao, M.-Z., Ma, C.-L., Li, C.-F., et al. (2016). Transcriptomic analysis of tea plant responding to drought stress and recovery. *PLoS ONE* 11:e0147306. doi: 10.1371/journal.pone.0147306
- Liu, Y., Liang, J., Sun, L., Yang, X., and Li, D. (2016). Group 3 LEA protein, *ZmLEA3*, is involved in protection from low temperature stress. *Front. Plant Sci.* 7:1011. doi: 10.3389/fpls.2016.01011
- Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using realtime quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262
- Lu, D., Wu, S., Gao, X., Zhang, Y., Shan, L., and He, P. (2010). A receptor-like cytoplasmic kinase, *BIK1*, associates with a flagellin receptor complex to initiate plant innate immunity. *PNAS* 107, 496–501. doi: 10.1073/pnas.0909705107
- Lu, X., Zhou, X., Cao, Y., Zhou, M., McNeil, D., Liang, S., et al. (2017). RNA-seq Analysis of Cold and Drought Responsive Transcriptomes of *Zea mays* ssp. *mexicana* L. *Front. Plant Sci.* 8:136. doi: 10.3389/fpls.2017.00136
- Ma, Q., Zhou, Q., Chen, C., Cui, Q., Zhao, Y., Wang, K., et al. (2019). Isolation and expression analysis of *CsCML* genes in response to abiotic stresses in the tea plant (*Camellia sinensis*). *Sci. Rep.* 9:e8211. doi: 10.1038/s41598-019-44681-7
- Maritim, T., Kamunya, S., Mwenda, C., Mireji, P., Muoki, R., Wamalwa, M., et al. (2016). Transcriptome-based identification of water-deficit stress responsive genes in the tea plant, *Camellia sinensis*. *J. Plant Biotechnol.* 43, 302–310. doi: 10.5010/JPB.2016.43.3.302
- Minhas, P. S., Rane, J., and Pasala, R. K. (2017). *Abiotic Stress Management for Resilient Agriculture*. Singapore: Springer. doi: 10.1007/978-981-10-5744-1
- Muthuramalingam, P., Krishnan, S. R., Pothiraj, R., and Ramesh, M. (2017). Global transcriptome analysis of combined abiotic stress signaling genes unravels key players in *Oryza sativa* L.: an *in silico* approach. *Front. Plant Sci.* 8:759. doi: 10.3389/fpls.2017.00759
- Nishizawa, A., Yabuta, Y., and Shigeoka, S. (2008). Galactinol and raffinose constitute a novel function to protect plants from oxidative damage. *Plant Physiol.* 147, 1251–1263. doi: 10.1104/pp.108.122465
- Panikulangara, T. J., Eggers-Schumacher, G., Wunderlich, M., Stransky, H., and Schöffl, F. (2004). Galactinol synthase1 a novel heat shock factor target gene responsible for heat-induced synthesis of raffinose family oligosaccharides in arabidopsis. *Plant Physiol.* 136, 3148–3158. doi: 10.1104/pp.104.042606
- Parmar, R., Seth, R., Singh, P., Singh, G., Kumar, S., and Sharma, R. K. (2019). Transcriptional profiling of contrasting genotypes revealed key candidates and nucleotide variations for drought dissection in *Camellia sinensis* (L.) O. Kuntze. *Sci. Rep.* 9, 1–12. doi: 10.1038/s41598-019-43925-w
- Peaucelle, A., Braybrook, S. A., Le Guillou, L., Bron, E., Kuhlemeier, C., and Höfte, H. (2011). Pectin-induced changes in cell wall mechanics underlie organ initiation in Arabidopsis. *Curr. Biol.* 21, 1720–1726. doi: 10.1016/j.cub.2011.08.057
- Phukan, U. J., Jeena, G. S., and Shukla, R. K. (2016). WRKY Transcription Factors: molecular regulation and stress responses in plants. *Front. Plant Sci.* 7:760. doi: 10.3389/fpls.2016.00760
- Pi, B., He, X., Ruan, Y., Jang, J.-C., and Huang, Y. (2018). Genome-wide analysis and stress responsive expression of *CCCH* zinc finger family genes in *Brassica rapa*. *BMC Plant Biol.* 18:373. doi: 10.1186/s12870-018-1608-7
- Rizhsky, L., and Liang, H. (2004). When defense pathways collide. The response of arabidopsis to a combination of drought and heat stress. *Plant Physiol.* 134, 1683–1696. doi: 10.1104/pp.103.033431
- Sakuma, Y., Liu, Q., Dubouzet, J. G., Abe, H., Shinozaki, K., and Yamaguchi-Shinozaki, K. (2002). DNA-binding specificity of the *ERF/AP2* domain of arabidopsis *DREBs*, transcription factors involved in dehydration- and cold-inducible gene expression. *Biochem. Biophys. Res. Commun.* 290, 998–1009. doi: 10.1006/bbrc.2001.6299
- Samarina, L. S., Malyukova, L. S., Efremov, A. M., Simonyan, T. A., Matskiv, A. O., Koninskaya, et al. (2020). Physiological, biochemical and genetic responses of Caucasian tea (*Camellia sinensis* (L.) Kuntze) genotypes under cold and frost stress. *PeerJ* 8:e9787. doi: 10.1717/peerj.9787
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Sharma, R., Singh, G., Bhattacharya, S., and Singh, A. (2018). Comparative transcriptome meta-analysis of *Arabidopsis thaliana* under drought and cold stress. *PLoS ONE* 13:e0203266. doi: 10.1371/journal.pone.0203266
- Stein, O., and Granot, D. (2019). An overview of sucrose synthases in plants. *Front. Plant Sci.* 10:95. doi: 10.3389/fpls.2019.00095
- Su, P. H., and Li, H. M. (2008). *Arabidopsis* stromal 70-kD heat shock proteins are essential for plant development and important for thermotolerance of germinating seeds. *Plant Physiol.* 146, 1231–1241. doi: 10.1104/pp.107.114496
- Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, 607–613. doi: 10.1093/nar/gky1131
- Tanaka, H., Osakabe, Y., Katsura, S., Shinji Mizuno, S., Maruyama, K., Kusakabe, K., et al. (2012). Abiotic stress-inducible receptor-like kinases negatively control ABA signaling in Arabidopsis. *Plant J.* 70, 599–613. doi: 10.1111/j.1365-3113.2012.04901.x
- Tuov, M., and Ryndin, A. (2011). Perspective tea hybrids in subtropics of the Russian Federation. *Subtrop. Ornamental Hortic.* 44, 101–109. Available online at: <https://www.elibrary.ru/item.asp?id=17656214&>
- Turksoz, D., and Sanlier, N. (2017). L-theanine, unique amino acid of tea, and its metabolism, health effects, and safety. *Crit. Rev. Food Sci. Nutr.* 57, 1681–1687. doi: 10.1080/10408398.2015.1016141
- Vogel, E. K., McCollough, A. W., and Machizawa, M. G. (2005). Neural measures reveal individual differences in controlling access to working memory. *Nature* 438, 500–503. doi: 10.1038/nature04171
- Wang, P., Chen, X., Guo, Y., Zheng, Y., Yue, C., Yang, J., et al. (2019). Identification of *CBF* transcription factors in tea plants and a survey of potential *CBF* target genes under low temperature. *Int. J. Mol. Sci.* 20, 5137. doi: 10.3390/ijms20205137
- Wang, Y., Jiang, C.-J., Li, Y.-Y., Wei, C.-L., and Deng, W.-W. (2012). *CsICE1* and *CsCBF1*: two transcription factors involved in cold responses in *Camellia sinensis*. *Plant Cell Rep.* 31, 27–34. doi: 10.1007/s00299-011-1136-5
- Wang, Y., Shu, Z., Wang, W., Jiang, X., Li, D., Pan, J., et al. (2016). *CsWRKY2*, a novel WRKY gene from *Camellia sinensis*, is involved in cold and drought stress responses. *Biol. Plant.* 60, 443–445. doi: 10.1007/s10535-016-0618-2
- Wang, Y., Yan, H., Qiu, Z., Hu, B., Zeng, B., Zhong, C., et al. (2019). Comprehensive analysis of *SnRK* gene family and their responses to salt stress in *Eucalyptus grandis*. *Int. J. Mol. Sci.* 20:2786. doi: 10.3390/ijms20112786
- Wang, Y.-X., Liu, Z.-W., Wu, Z.-J., Li, H., and Zhuang, J. (2016). Transcriptome-wide identification and expression analysis of the NAC gene family in tea plant [*Camellia sinensis* (L.) O. Kuntze]. *PLoS ONE* 11:e0166727. doi: 10.1371/journal.pone.0166727
- Wrzaczek, M., Brosché, M., Salojärvi, J., Kangasjärvi, S., Idänheimo, N., Mersmann, S., et al. (2010). Transcriptional regulation of the *CRK/DUF26* group of Receptor-like protein kinases by ozone and plant hormones in arabidopsis. *BMC Plant Biol.* 10:95. doi: 10.1186/1471-2229-10-95
- Wu, H.-C., Bulgakov, V. P., and Jinn, T.-L. (2010). Pectin methylesterases: cell wall remodeling proteins are required for plant response to heat stress. *Front. Plant Sci.* 9:1612. doi: 10.3389/fpls.2018.01612
- Xia, E.-H., Li, F.-D., Tong, W., Li, P.-H., Wu, Q., Zhao, H.-J., et al. (2019b). Tea plant information archive: a comprehensive genomics and bioinformatics platform for tea plant. *Plant Biotechnol. J.* 17, 1938–1953. doi: 10.1111/pbi.13111
- Xia, E.-H., Tong, W., Wu, Q., Wei, S., Zhao, J., Zhang, Z. Z., et al. (2019a). Tea plant genomics: achievements, challenges and perspectives. *Hortic. Res.* 7:7. doi: 10.1038/s41438-019-0225-4
- Xie, Z., Nolan, T. M., Jiang, H., and Yin, Y. (2019). *AP2/ERF* transcription factor regulatory networks in hormone and abiotic stress responses in arabidopsis. *Front. Plant Sci.* 10:228. doi: 10.3389/fpls.2019.00228
- Yin, H., Ma, Y., Deng, Y., Xu, Z., Liu, J., Zhao, J., et al. (2016). Genome shuffling of *Saccharomyces cerevisiae* for enhanced glutathione yield and relative gene expression analysis using fluorescent quantitation reverse transcription polymerase chain reaction. *J. Microbiol. Methods* 127, 188–192. doi: 10.1016/j.mimet.2016.06.012
- Yu, J., Lai, Y., Wu, X., Wu, G., and Guo, C. (2016). Overexpression of *OsEm1* encoding a group I LEA protein confers enhanced drought tolerance in rice. *Biochem. Biophys. Res. Commun.* 478, 703–709. doi: 10.1016/j.bbrc.2016.08.010
- Yue, C., Cao, H., Lin, H., Hu, J., Ye, Y., Li, J., et al. (2019). Expression patterns of alpha-amylase and beta-amylase genes provide insights into the molecular mechanisms underlying the responses of tea plants (*Camellia sinensis*) to stress and postharvest processing treatments. *J. Plant Biol.* 250, 281–298. doi: 10.1007/s00425-019-03171-w

- Yue, C., Cao, H. L., Wang, L., Zhou, Y. H., Huang, Y. T., Hao, X. Y., et al. (2015). Effects of cold acclimation on sugar metabolism and sugar-related gene expression in tea plant during the winter season. *Plant Mol. Biol.* 88, 591–608. doi: 10.1007/s11103-015-0345-7
- Zhao, T., Xia, H., Liu, J., and Ma, F. (2014). The gene family of dehydration responsive element-binding transcription factors in grape (*Vitis vinifera*): genome-wide identification and analysis, expression profiles, and involvement in abiotic stress resistance. *Mol. Biol. Rep.* 41, 1577–1590. doi: 10.1007/s11033-013-3004-6
- Zheng, C., Wang, Y., Ding, Z., and Zhao, L. (2016). Global transcriptional analysis reveals the complex relationship between tea quality, leaf senescence and the responses to cold-drought combined stress in *Camellia sinensis*. *Front. Plant Sci.* 7:1858. doi: 10.3389/fpls.2016.01858
- Zhou, R., Yu, X., Zhao, T., Ottosen, C.-O., Rosenqvist, E., and Wu, Z. (2019). Physiological analysis and transcriptome sequencing reveal the effects of combined cold and drought on tomato leaf. *BMC Plant Biol.* 19:377. doi: 10.1186/s12870-019-1982-9
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Copyright © 2020 Samarina, Bobrovskikh, Doroshkov, Malyukova, Matskiv, Rakhmangulov, Koninskaya, Malyarovskaya, Tong, Xia, Manakhova, Ryndin and Orlov. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Sea Urchin as a Universal Model for Studies of Gene Networks

Leonid Adonin^{1,2,3*}, Anatoliy Drozdov⁴ and Nickolai A. Barlev^{1,3,5*}

¹Moscow Institute of Physics and Technology, Dolgoprudny, Russia, ²Institute of Environmental and Agricultural Biology (X-BIO), Tyumen State University, Tyumen, Russia, ³Orekhovich Institute of Biomedical Chemistry, Moscow, Russia, ⁴Zhirmunsky National Scientific Center of Marine Biology, Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russia, ⁵Institute of Cytology, Russian Academy of Sciences, Saint-Petersburg, Russia

The purple sea urchin *Strongylocentrotus purpuratus* has been used for over 150 years as a model organism in developmental biology. Using this model species, scientists have been able to describe, in detail, the mechanisms of cell cycle control and cell adhesion, fertilization, calcium signaling, cell differentiation, and death. Massive parallel sequencing of the sea urchin genome enabled the deciphering of the main components of gene regulatory networks during the activation of embryonic signaling pathways. This knowledge helped to extrapolate aberrations in somatic cells that may lead to diseases, including cancer in humans. Furthermore, since many, if not all, developmental signaling pathways were shown to be controlled by non-coding RNAs (ncRNAs), the sea urchin organism represents an attractive experimental model. In this review, we discuss the main discoveries in the genetics, genomics, and transcriptomics of sea urchins during embryogenesis with the main focus on the role of ncRNAs. This information may be useful for comparative studies between different organisms, and may help identify new regulatory networks controlled by ncRNAs.

Keywords: sea urchin, gene expression, cell signaling, long non-coding RNA, genomics

INTRODUCTION

The first use of animals for experimental purposes dates back to Ancient Greece. “Generation of Animals” of Aristotle (1942) describes the first systematic study of embryonic development as a phenomenon, which recognizes the key questions about the emergence and relations between hierarchically organized parts of an organism.

Model organisms help in the testing of novel biological hypotheses, which come from *in cellulo* observations and need to be tested at the whole organism level. Hence, model organisms represent a very important tool in modern biology. Currently, the list of model organisms includes over 100 species of animals, plants, protozoa, and viruses. The most popular model species include the frog, zebrafish, chicken, mouse, fruit fly, and nematode (Rzepnikowska et al., 2017; Kuo et al., 2018; Chatterjee and Deng, 2019; Marques et al., 2019; Nielsen, 2019; Tang et al., 2019). All these species are used by researchers in a wide range of molecular biological applications, but, unfortunately, none of them are versatile enough to satisfy various experimental needs.

In this review, we focus on the main discoveries in genetics and genomics that were made using a popular model object – the purple sea urchin *Strongylocentrotus purpuratus*, which

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Anna Kudryavtseva,
Engelhardt Institute of Molecular
Biology (RAS), Russia
Alexander M. Ishov,
University of Florida, United States

*Correspondence:

Leonid Adonin
leo.adonin@gmail.com
Nickolai A. Barlev
nick.a.barlev@gmail.com

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 08 November 2020

Accepted: 10 December 2020

Published: 20 January 2021

Citation:

Adonin L, Drozdov A and
Barlev NA (2021) Sea Urchin as a
Universal Model for Studies of
Gene Networks.
Front. Genet. 11:627259.
doi: 10.3389/fgene.2020.627259

has been used in biology for over 150 years (Stimpson, 1857). The species was chosen as a model object for several objective reasons: sea urchins are easy to propagate in the laboratory; it is easy to get synchronous embryo cultures and induce rapid embryogenesis; the embryo is transparent and has a simple structure. Genome sequencing and the description of complex gene regulatory networks during the sea urchin embryogenesis made this model object indispensable for the study of gene expression regulation.

Echinoderms are a sister group of the Chordate phylum. This group has branched out from Chordate before the Cambrian period (more than 500 million years ago, **Figure 1**; McClay, 2011). Other studies, based on the multigene and multiprotein studies, indicate a more accurate time of divergence (Hedges, 2002; Zaidel-Bar, 2009).

The first studies, which describe the normal embryogenesis of a sea urchin, date back to the middle of the 19th century. Since then, sea urchin embryos have become a popular model in developmental biology. The normal life cycle of sea urchins is shown in **Figure 2**. At present, certain echinoid species (e.g., *S. purpuratus*, *Strongylocentrotus droebachiensis*, *Strongylocentrotus intermedius*, *Hemicentrotus pulcherrimus*, *Lytechinus variegatus*, *Paracentrotus lividus*, and *Mesocentrotus franciscanus*) are widely used as experimental models in developmental biology (McClay, 2011). The early stage of embryogenesis of the purple sea urchin was used for studying intercellular communication and cell adhesion (Horstadius, 1939; Giudice, 1962; McClay and Fink, 1982), cell cycle control mechanisms (Evans et al., 1983), calcium signaling (Whitaker, 2006), fertilization (Briggs and Wessel, 2006), cell differentiation (Giudice, 1973), and cell survival and death (Chiarelli et al., 2016).

The genome size of the purple sea urchin is only a quarter of the human genome, despite it having about the same number of genes. Genome analysis has shown that most of its genes are common to representatives of deuterostomes, which has

in turn uncovered an unexpectedly close relation to humans among all used invertebrate model species (Davidson, 2006; Sea Urchin Genome Sequencing Consortium et al., 2006; Cameron, 2014). For example, the sea urchin genome contains orthologs of human disease-associated genes, which are expressed in sea urchin embryogenesis.

The sea urchin genome was shown to contain more than 400 genes, whose products are involved in the regulation of cell homeostasis. Most of these genes display a remarkable conservation of their sequences during the evolution (Goldstone et al., 2006; Rast et al., 2006). The sea urchin genome contains 65 genes of the ATP-binding cassette transporter superfamily (Hamdoun et al., 2004; Goldstone et al., 2006), while, in humans, only 48 members of this family are known to date (Dean and Annilo, 2005). Mutations in these genes lead to several pathologies in humans, including degeneration of the retina, cystic fibrosis, neurological diseases, cholesterol transport disorders, anemia, and many others (Dean et al., 2001). The sea urchin as an experimental model is also frequently used in toxicology and in environmental human health science since it allows an accurate estimate of cancer risk before any epidemiologic evidence is available (Bellé et al., 2007).

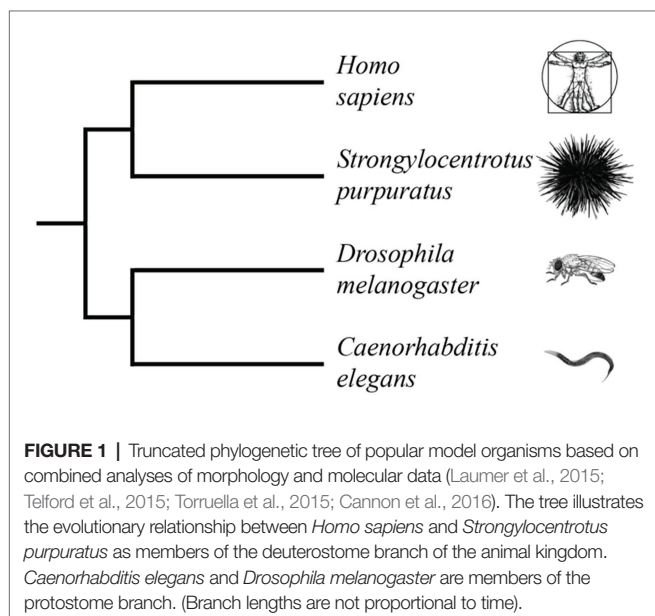
Furthermore, sea urchin embryos are used by scientists as a convenient object for elucidating common cellular molecular mechanisms involved in human health and disease. In particular, the sea urchin is used as a model system for studying neurodegenerative disorders that can cause dementia and memory loss (Nakajima et al., 2004).

In the past, it was hypothesized that certain signaling pathways involved in the embryo's morphogenesis could be aberrantly activated during tumorigenesis. Unfortunately, to develop this idea further, scientists did not have an appropriate human experimental model. This is due to various ethical aspects that restricted human embryo studies (De Wert et al., 2002; De Wert and Mummery, 2003; Holm, 2003; Lo and Parham, 2009). Therefore, an early embryogenesis of sea urchins could be a good model for cancer research.

The normal processes of cell proliferation and differentiation are controlled by several developmental gene regulatory networks. Disunity in these networks leads to the initiation and progression of tumors (Hanahan and Weinberg, 2000; Reya et al., 2001; Pires-daSilva and Sommer, 2003).

Notch, Wnt, and Hedgehog (Hh) signaling pathways are highly conserved from sea urchins to humans. The current model of these pathways, including general components shown on **Figure 3** and their main roles in embryogenesis and cancer, is described on **Table 1**.

The Hh pathway plays a crucial role in many fundamental processes of metazoan organisms, including tissue homeostasis and their embryonic development. In the development of the sea urchin, this pathway controls the establishment of the left–right asymmetry in embryos (Warner et al., 2016). According to the previously proposed model, this signaling pathway, similar to vertebrates, provides an asymmetrical expression of Nodal, which is an important cytokine of the TGF beta superfamily. Moreover, in humans, the aberrant activation of this signaling pathway is increasingly associated with various cancers.



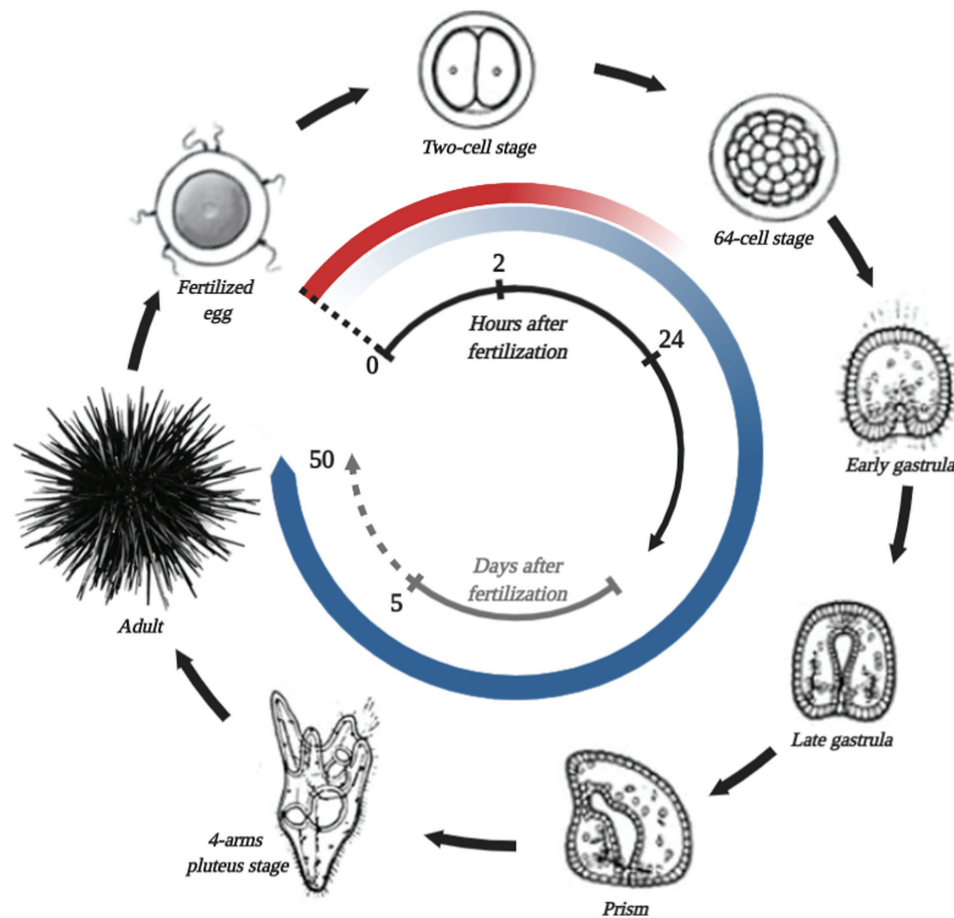


FIGURE 2 | Simplified *S. purpuratus* life cycle, stages of which are connected by black arrows. The name of each stage is shown in the picture. The beginning of the life cycle is fertilization, which is marked by a black dotted line. In the center, the black and gray time scale represents hours and days after fertilization, correspondingly. The red circular gradient line represents the degradation of the general maternal transcripts. The circular blue gradient shows the beginning of zygotic genome activation and increases in transcribed gene numbers.

For example, Hh was shown to control proliferation, malignancy, and metastasis (Sari et al., 2018). In particular, while the Hh signaling pathway is mainly repressed during mammary embryonic development, overexpression of some components (PTCH1, GLI1/2) of the Hh are up-regulated in tumor stem cells of human breast cancer (Liu et al., 2006).

The key components of Notch signaling are present in all metazoan organisms (Gordon et al., 2008). The canonical Notch pathway begins when a ligand of the Delta/Serrate/LAG-2 (DSL) family binds to the transmembrane receptor protein, Notch (Fehon et al., 1990). The schema of all stages of the process is shown in **Figure 3** (part NOTCH). It is well established that the Delta/Notch signaling pathway is intimately involved in mesoderm formation. Subsequently, deregulation of this pathway leads to the elimination of mesoderm derivatives during the embryogenesis of sea urchins (Sherwood and McClay, 1999, 2001; Sweet et al., 2002; Croce and McClay, 2010). In humans, members of the Notch signaling pathway play a key role in embryonic vasculature development

(Patel et al., 2005). Notch can in fact be either oncogenic or tumor suppressive depending on the tissue and cellular context. In addition, this pathway is one of the most activated in cancer cells and contributes to metastasis (Venkatesh et al., 2018). For example, the development of squamous cell carcinomas in various epithelial tissues is directly related to mutations in members of the Notch family. These mutations represent the most common cause of misregulation of this signaling pathway (Nowell and Radtke, 2017).

The Wnt signaling pathway regulates the embryogenesis and homeostasis of multicellular organisms. In sea urchin embryos, the Wnt signaling pathway contributes to the activation of the endomesodermal gene regulatory network, whose genes start their expression on the 16-cell stage of embryos (Kumburegama and Wikramanayake, 2008). Also, this pathway regulates the formation of the animal-vegetal (A-V) axis in sea urchin and sea anemone embryos.

In humans, cancer, obesity, and diabetes are the result of the Wnt pathway dysregulation (Langton et al., 2016;

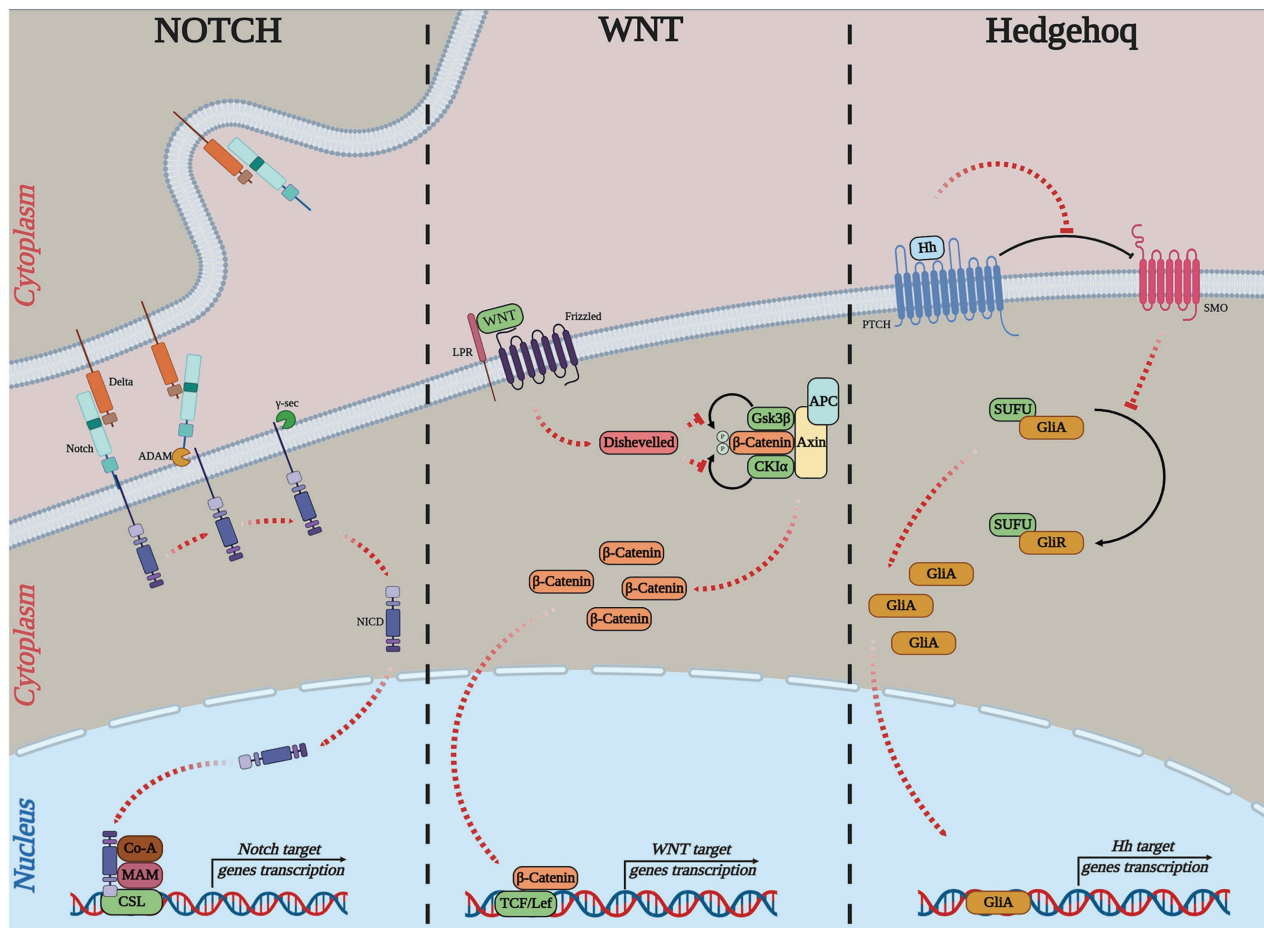


FIGURE 3 | Simple outline of the current models of the canonical Notch, WNT, and Hh pathways. NOTCH: Delta, delta-like ligand, Notch, ADAM, ADAM-family metalloprotease; γ -sec, γ -secretase; NICD, notch intracellular domain; Co-A, transcription coactivator; MAM, conserved and essential nuclear factor mastermind; CSL, DNA-binding transcription factor. WNT: Frizzled; WNT, wingless-type MMTV integration site; LRP, low-density lipoprotein receptor-related protein; APC, adenomatous polyposis coli; Dishevelled, cytoplasmic phosphoprotein; GSK3 β , glycogen synthase kinase-3; CK1 α , casein kinase 1 alpha; TCF, T-cell-specific transcription factor; LEF, lymphoid enhancer-binding factor. Hedgehog: Hh, hedgehog; PTCH, patched; SMO, smoothened; SUFU, suppressor of fused; GliA, Gli-family zinc finger.

Mzoughi et al., 2017). It is shown that the pathway is involved in the regulation of the metabolism of cancer cells, which facilitates tumor progression (Lu et al., 2010; Wang and Kunze, 2015; Poli et al., 2018; Sun et al., 2018).

In the comparison analysis, it was shown that the sea urchin genome contains about 90% of the described homologous components of Wnt signal transduction pathways (Croce et al., 2006; Robert et al., 2014). However, from 13 known Wnt subfamilies, *S. purpuratus* has only 11: it is missing only Wnt2 and Wnt11 homologs. Meanwhile, last year Croces' group identified a gene encoding Wnt2 ortholog in the genome of a related sea urchin *P. lividus*. However, they found no evidence of a bona fide wnt2 gene in *S. purpuratus* when they reanalyzed its genome (Robert et al., 2019). A consortium of scientists was able to find only about half of the Wnt transcriptional target genes that were reported in the literature.

The purple sea urchin can also be considered as a model of gene expression in the normal developmental processes and is used now as an *in vivo* model to evaluate the Epithelial/Mesenchymal Transition (EMT; Romancino et al., 2017). In humans, reactivated EMT drives organ fibrosis and tumor progression (Lim and Thiery, 2012; Wu et al., 2012; Nieto et al., 2016). The process of EMT is regulated by a cohort of specific transcription factors that includes Zeb1/Zeb2, Snail, Slug, and Twist (Tulchinsky et al., 2019). Together with chromatin-modifying enzymes, these factors exert both repressive and activating functions. For example, Zeb1, by binding to the E-box consensus site in the DNA, inhibits the transcription of the CDH1 gene, whose product plays a critical role in forming cell-cell junctions. On the other hand, when Zeb1 complexes with Yap1, a member of the Hippo pathway, it becomes a transcriptional activator to control the transcription of CTGF and AXL genes (Lehmann et al., 2016).

TABLE 1 | Major Notch, WNT, and Hedgehog (Hh) pathways roles in multicellular organisms' embryo development and cancer.

	NOTCH	WNT	Hedgehog
Functions in development	The Notch pathway is a major determinant of cell fate across all metazoans (Artavanis-Tsakonas and Muskavitch, 2010; Bray, 2016; Henrique and Schweisguth, 2019; Lloyd-Lewis et al., 2019).	The Wnt signaling pathway regulates many cell functions, including proliferation, migration, apoptosis, and differentiation. It also plays a key role in controlling body axis formation. It is essential during embryonic development and also in the homeostasis of several adult tissues including the GI tract (Flanagan et al., 2015, 2017), liver, breast, and skin (Nusse and Clevers, 2017).	The Hedgehog signaling pathway plays a significant role in the normal embryonic development of invertebrates and vertebrates (Skoda et al., 2018). The Hh genes are played in organization of the polarity of the organism and the development of many tissues and organs. The pathway is involved in the maintenance of somatic stem cells and pluripotent cells important for tissue repair (Beachy et al., 2004; Karhadkar et al., 2004; Zhou et al., 2006; Stecca et al., 2007; Lowry et al., 2008).
Role in cancer	Notch plays an oncogenic role: it is overexpressed in breast cancer (Kontomanolis et al., 2018), gastric cancer (Zhou et al., 2013), pancreatic cancer (Ma et al., 2013), and colorectal cancer (Vinson et al., 2016). Notch acts as a tumor suppressor gene: its expression is downregulated in skin cancer (Lefort et al., 2007), liver cancer (Viatour et al., 2011), non-small cell lung cancer (Konishi et al., 2010), and some breast cancers (Parr et al., 2004).	Mutations of Wnt pathway members cause cancer development in humans (Segditsas and Tomlinson, 2006). It is known that Wnt signaling is deregulated in gastric tumors (Clements et al., 2002; Flanagan et al., 2017). The WNT pathway plays critical roles in epithelial ovarian cancer development (Nguyen et al., 2019), colorectal cancer (Wang et al., 2018), and thyroid carcinogenesis (Ely et al., 2018).	Hh signaling is involved in the development of pancreatic, and esophageal cancer (Bailey et al., 2009), gastric, and prostate cancer (Sheng et al., 2004), as well as basal cell carcinoma (Gutzmer and Solomon, 2019) and medulloblastoma (Gordon et al., 2018).

THE POSTGENOMIC ERA IN SEA URCHIN-RELATED RESEARCH

Sea Urchin Genome Sequencing and Analysis

As mentioned before, the sea urchin genome reveals striking similarities to humans and shares with the latter a lot of common gene regulatory pathways.

The first assembly and annotation of the sea urchin genome results were published in 2006 (Sea Urchin Genome Sequencing Consortium et al., 2006) and initiated an active exploration of its genomics and transcriptomics. Over the past 14 years, researchers refined the assembly and annotation of the sea urchin genome. Additional genomic and transcriptomic resources were created, for example, EchinoBase.¹

The primary assessment of the purple sea urchin genome was estimated as ~800 Mb in size (Hinegardner, 1971). After the deep-sequencing refinement, the purple sea urchin genome was predicted to contain 23,300 genes (Sea Urchin Genome Sequencing Consortium et al., 2006). The current genome analysis revealed 33,491 genes (and 556 pseudogenes) that encode 38,439 proteins.²

A comparative analysis of the sea urchin genome with vertebrates revealed an unprecedented complexity relative to other animals in terms of their innate immune recognition receptors (Rast et al., 2006). The SUGSC research team assumed that about 4–5% of all the sea urchin genes identified are involved directly in the immune functions (Sea Urchin Genome Sequencing Consortium et al., 2006).

Around 222 members of the Toll-like receptor family and 203 genes of the NACHT domain–LRR family were described

in the sea urchin genome in addition to genes from a large family of cysteine-rich receptor proteins (Rast et al., 2006). The sea urchin immune system showed the presence of a complement system similar to the chordate (Hibino et al., 2006).

Since transcriptional networks are regulated by transcription factors, it is important to mention the work of Howard-Ashby's team that described the main families of genes coding for transcription factors in the *S. purpuratus* genome (bHLH, Nuclear Receptor, Basic Leucine Zipper, T-box, Smad, Sox, and other smaller families). The number of genes encoding transcription factors of each family in the sea urchin is comparable to that found in the *Drosophila* genome, but it is almost twice less than the number of such genes found in the human genome. A similar result was obtained when analyzing the situation with HomeoBox genes (Howard-Ashby et al., 2006). The evolution of new genes during the evolution since branching Echinodermata from the common Deuterostome branch is associated with the adaptation process, increasing the level of complexity and/or changing the key cellular mechanisms.

During the course of genome analysis, the SUGSC team identified more than 1,200 genes involved in signal transduction. The *S. purpuratus* genome contains 353 protein kinases (Sea Urchin Genome Sequencing Consortium et al., 2006) and 14 lipid kinases (Bradham et al., 2006). The number of sea urchin protein kinases is higher than that described in the *Drosophila* genome (about 230 members) but less than in the human kinome (518). Although most of the sea urchin kinase subfamilies are often represented by only a single member, their diversity is surprisingly high and corresponds to approximately 97% of the whole human kinome. Only four subfamilies of kinases are missing (Axl, FastK, H11, and NKF3) in the sea urchin compared to the human kinome, whereas the fruit fly kinome lacks 20 of those, and the kinome of worms misses 32 subfamilies (Bradham et al., 2006). Importantly, it has been shown that

¹<http://legacy.echinobase.org/Echinobase/>

²<https://www.ncbi.nlm.nih.gov/genome/86>

approximately 88% of described kinases are expressed during embryogenesis (Bradham et al., 2006; Byrum et al., 2006).

To follow the compilation of similarities in the gene ontology of the sea urchin and of humans, it is important to note that they share common mechanisms of cell cycle control. Perhaps, not surprisingly, a number of genes involved in cell cycle control and DNA metabolism have been described for the sea urchin, although its number is lower compared to the human genome. In addition, a few cases of echinoderm-specific gene diversifications have been described (Fernandez-Guerra et al., 2006). Notably, the sea urchin genome contains orthologs of almost all cyclin-dependent kinases, except CDK3. Members of the NIMA-related kinases family (NEK proteins) are, judging by their complexity, close to vertebrates, whereas the complexity of Polo and Aurora mitotic kinase families are close to those found in the worm (Fernandez-Guerra et al., 2006).

Furthermore, a number of known genes involved in DNA replication, repair, and the mitotic checkpoint were also found in the sea urchin. Interestingly, the sea urchin has a single p63/p73 hybrid homologous to the p53, p63, and p73 members of the p53 family of tumor suppressors (Belyi et al., 2010). In addition, the sea urchin contains two homologs of the pRB tumor suppressor and also one homolog of the p21/p27 family of CDK inhibitors (Fernandez-Guerra et al., 2006). Furthermore, the sea urchin genome shares four families of RAS GTPases with humans: Ras, Rho, Rab, and Atr, although 90% of all small GTPases are expressed during embryogenesis (Beane et al., 2006).

Gene Regulatory Networks

As early as in the pre-genome era, common features, and concepts, of the gene regulatory network (GRN) were described by researchers who used the *S. purpuratus* sea urchin as a model organism. Genome sequencing and annotation made it possible to structure the information, which led to the creation of one of the most complete networks for the regulation of genes during the early embryogenesis (Davidson et al., 2002; Sea Urchin Genome Sequencing Consortium et al., 2006; Oliveri et al., 2008; Peter and Davidson, 2010, 2017; Peter et al., 2012; Martik et al., 2016).

The key step toward the understanding of basic mechanisms of embryogenesis and global GRN was the deep sequencing of RNA. This allowed the accumulation of data on gene expression networking during the embryogenesis of *S. purpuratus* (Rafiq et al., 2014; Tu et al., 2014; Barsi et al., 2015; Gildor et al., 2016; Israel et al., 2016; Janies et al., 2016; Pérez-Portela et al., 2016). The described schemes of the GRN became the best tool for the analysis of the development of the genetic control (Peter and Davidson, 2017).

In the study by Tu et al. (2014), the expression profiles of more than 16,000 genes were measured during embryogenesis. For a clearer presentation of the expression profiles of embryonic genes, the authors performed a cluster analysis. The clusters were grouped into four main groups according to their overall dynamics: “off,” “on,” “transient,” and “other” (Tu et al., 2014). They showed that complex expression patterns of many genes underlie embryonic development, especially in the early stages preceding gastrulation.

A study of Rafiq (2014) sets a basis for understanding the genomic regulatory control of a major morphogenetic process – skeletal morphogenesis for embryogenesis. The authors have identified 420 transcripts whose expression levels in primary mesenchymal cells (PMC) were significantly different from other samples. Most of these genes are transcribed at relatively low levels at the stage of mesenchymal blastula. They were targeted at Ets1 and Alx1, key transcription factors that provide regulatory inputs at the top of the PMC regulatory differentiation network.

It was shown that more than half of the identified transcripts received essential inputs from Ets1 and/or Alx1, most of which were positive. All these data point to their key role in the cell-specific identity of PMCs (Rafiq et al., 2014). Additionally, the authors described about 200 transcripts that were not significantly affected by Ets1 or Alx1 knockdown.

The name of the GRN concept implies that exons play a major role in gene cascades. However, The Human Genome Project and the subsequent deciphering of a large number of genomes made it obvious that the bulk of the genome consists of sets of repetitive DNA (Lander et al., 2001). Coding DNA fragments (exons) occupy no more than 2–3% of the genome (Carey, 2017). The major components of the genome are represented by two groups of repetitive DNA sequences: tandem repeats and dispersed repeats or transposons [transposable elements, (TE); Paço et al., 2019]. Almost all eukaryotic genomes contain TE, for example, it occupies about half of the human genome (Wang and Kunze, 2015).

Non-coding RNA in the Sea Urchin

It is well established that almost all of the human genome is transcribed into RNA. Surprisingly, most of the transcribed RNA does not code for proteins and is called non-coding RNA (ncRNA; Lander et al., 2001; Carninci et al., 2005; ENCODE Project Consortium et al., 2007; Djebali et al., 2012; Laurent et al., 2015; Carey, 2017), including microRNA (miRNA, 22–25 bp) and long non-coding RNA (lncRNA, >200 bp). ncRNAs are involved in the transcription regulation of genes and other ncRNA (Li and Liu, 2019). The primary source of all kinds of ncRNA in the genome is transposons (Hadjigargyrou and Delihias, 2013).

MicroRNAs control gene expression via multiple modes (Cai et al., 2009; Steitz and Vasudevan, 2009). In general, the 5' proximal “seed” region (nucleotide 2–8) of miRNAs exhibits imperfect complementarity to the 3'UTR of the target mRNA (Lewis et al., 2005). Subsequently, this newly formed double-stranded RNA is destroyed by dsRNase, RISC. However, a few cases have been reported when miRNAs regulated the expression by binding the 5'UTR of mRNAs, thereby interfering with the binding of translation initiation factor, eIF4 (Lee et al., 2009; Brümmer and Hausser, 2014).

To date, several models have been proposed describing the consequences of the interaction between the miRNA complex and their targets. The miRNA-dependent gene silencing can be achieved at three stages, including pre-translational, co-translational, and post-translational steps (Finnegan and Matzke, 2003; Garneau et al., 2007; Grewal and Elgin, 2007; Fabian et al., 2010; Carroll et al., 2014).

The lncRNA-driven transcriptional regulation is more complex and includes several mechanisms: (1) lncRNA can recruit a

regulatory protein complex to a gene or an entire chromosome; (2) the binding of a transcriptional factor is inhibited by lncRNA; (3) transcription of lncRNAs regulates the transcription of adjacent protein-coding genes; and (4) the heterochromatic or euchromatic organization of regions in close proximity stabilizes these territories and controls the spreading of post-translational modifications to nearby chromatin (Long et al., 2017). Furthermore, lncRNAs play a key role in stem cell differentiation, immune response, epigenetic regulation, inflammation-related diseases, and tumor development (Briggs et al., 2015; Huarte, 2015; Boon et al., 2016; Chen et al., 2017).

Description and analysis of ncRNA regulatory networks could provide new insights into gene transcription regulation not only during the embryonic development, but also in cancer, when specific developmental programs are aberrantly reactivated. Thus, understanding the complexity of these regulatory networks will make it possible to determine the consequences of their disruption in the course of various diseases.

REGULATORY NETWORKS BASED ON NON-CODING RNAs

After the publication of the sea urchin genome annotation, miRNAs were identified (Peterson et al., 2009; Wheeler et al., 2009; Campo-Paysaa et al., 2011). These studies revealed that a few, very conserved, miRNAs are also present in humans (Song et al., 2012). The authors had cloned and sequenced small RNAs (18–40 nucleotides) from different embryo stages, ranging from unfertilized eggs to larva pluteus. Around 49 miRNAs were identified, in which three of these were novel miRNAs (not annotated in miRBase previously). Most of the miRNAs are present in the egg and have dynamic accumulation profiles, with the majority of these being upregulated during gastrulation.

To test the function of miRNA during the embryonic development, authors decided to suppress the dsRNA processing enzyme, Dicer, with Dicer morpholino antisense oligonucleotides (MASO). The majority of injected embryos successfully developed to the stage of blastula. However, at the stage of 48 h p.f.,

embryos that expressed Dicer MASO failed to enter the gastrulation stage (Song et al., 2012). Developmental defects varied from general retardation to cell death.

One interesting example of opposing functions exerted by one microRNA is miR-31. Importantly, it is expressed during the embryogenesis of *S. purpuratus* and suppresses several components of PMC GRN (Pmar1, Alx1, Snail, and VegfR7). Knockdown of miR-31 causes a disturbance of the function of PMC that forms the embryonic skeleton (Stepicheva and Song, 2015). Meanwhile, in humans, miR-31 is considered as a tumor suppressor. Yet, it can affect several signaling pathways that have opposite effects on the proliferation: RAS/MARK and PI3K/AKT stimulate growth, whereas RB/E2F inhibits it. Unfortunately, specific molecular mechanisms that regulate miR-31 in the sea urchin are not known at the moment.

The latest analysis of the sea urchin genome showed that known transposons³ occupy about 15% of the genome, including the major class DNA transposons (Figure 4; Lebedev et al., 2019). The percentage of genome occupied by transposons is higher than that of the worm *Caenorhabditis elegans* but less than that of fruit fly or human (Kazazian and Moran, 2017).

The development of new sequencing techniques, such as RNA-Seq, has greatly advanced our understanding and knowledge of new RNAs. In one of the latest studies on this, Hezroni and co-authors, using a PLAR-algorithm, predicted more than 5,000 new sequences of lncRNA in sea urchin transcriptomes (Hezroni et al., 2015).

Genes that code for lincRNAs are more species-specific and less conserved than the gene encoding proteins. In the genome of the sea urchin, syntenic (homologous genes situated on the same chromosomes but in different species) lncRNA genes were identified for more than 2,000 human lincRNA genes. This suggests that the sea urchin likely contains a lot of conserved functional vertebrate lincRNA homologs (Hezroni et al., 2015). Of all detected syntenic lincRNAs, only 18 were found in other amniotes.

One example is LINC00261, located downstream of the *Foxa2* gene, which codes for a transcription factor. In all

³<https://www.girinst.org/repbase>

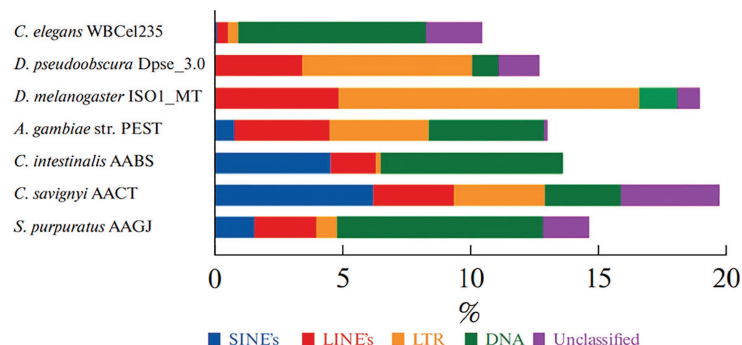


FIGURE 4 | The percent content of transposable elements (TE) in genomes of the sea urchin *S. purpuratus* and other invertebrates. TE classes are marked with a color: nonLTR SINE – blue; LINE – red; LTR TE – yellow; DNA transposons – green; non-annotated repeats – lilac (from Lebedev et al., 2019).

vertebrates, this lincRNA is expressed in endodermal tissues, and in sea urchins, it is expressed in the gut. LINC00261 plays the tumor suppressive role being involved in the regulation of DNA damage (Shahabi et al., 2019).

Another syntenic lincRNA (partially annotated as LINC01122 and LOC101927285 in humans) is expressed in the brain and reproductive tissues across vertebrates. In sea urchins, it is expressed in the adult ovary. Unfortunately, the specific functions of both orthologs are unknown. However, this region is located between the *Fanc1* and *Bcl11a* loci and one can assume that ncRNA expressed from this locus, at least in humans, may participate in the regulation of DNA damage response and/or apoptosis possibly through the p53 regulatory network.

Well-known embryogenesis signaling pathways conserved between different species can serve as the starting point for understanding the complex network of development organization. In recent years, ncRNAs have rapidly emerged as crucial regulators of main signaling pathways in embryo development and cancer (Fu et al., 2019). Regulation takes place at different levels: from the transcriptional to the post-transcriptional and translational levels, for example, lncRNA is involved in the regulation of WNT, Notch, and other signaling pathways in cancer (Li and Kang, 2014; Trimarchi et al., 2014; Yuan et al., 2014; Ong et al., 2017; Peng et al., 2017; Shen et al., 2017).

FUTURE OF THIS MODEL ORGANISM FOR GENE EXPRESSION STUDIES

Embryogenesis is regulated by complicated mechanisms to ensure that different types of cells and tissues develop from one cell. All processes of embryogenesis are strictly coordinated by signaling pathways. Next-generation sequencing and bioinformatics methods have made it possible to describe the main components of these pathways (Figure 3) and the branched gene regulatory networks.

Studying the regulatory networks of development and its organization on all levels requires experimental models. It seems to us that this model object – the purple sea urchin *S. purpuratus* – is the most suitable system for studying the regulation system based on ncRNAs. Further deep bioinformatic analysis of the genome, and the transcriptomic profiling of embryogenesis

stages of this model object, will promote important discoveries in gene networking.

Spontaneous alterations in these coordinated gene expression programs can lead to the development of an unhealthy embryo. Furthermore, reactivation of these pathways in somatic cells can cause many diseases, including cancer in humans (Table 1).

Interestingly, the history of modern cancer research begins with the sea urchin: in the first decade of the 20th century, the German biologist Boveri discovered that improper fertilization of sea-urchin eggs with two sperm rather than one led to chromosomal aberrations and to the failure of proper development (Laubichler and Davidson, 2008). Furthermore, purple sea urchins and some other urchin species (*L. variegatus*, *M. franciscanus*) retain the ability to regenerate lost or damaged tissues with age (Bodnar and Coffman, 2016).

However, cancers have not been detected in sea urchins. In fact, the life span of some sea urchin species reaches 100 years (Ebert, 2010; Kober and Bernardi, 2013). How such genomic stability is achieved and what the regulatory transcription mechanisms involved in the longevity of these organisms are require further investigation. Thus, sea urchins can provide insights into the processes in cases of serious human diseases associated with the regulation of transcription.

AUTHOR CONTRIBUTIONS

LA drafted the manuscript. AD and NB supervised the preparation of the draft. NB helped with writing, proofreading, and editing the final version of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

LA and NB appreciate the support of RSF grant # 20-15-00189. NB appreciates the support of the grant from the Russian Government Program for the Recruitment of the leading scientists into the Russian Institutions of Higher Education 14.W03.31.0029. LA appreciates the Ministry of Science and Higher Education of the Russian Federation (agreement # 075-00337-20-03, project FSMG-2020-0004).

REFERENCES

- Aristotle. (1942). *Generation of animals*. A. L. Peck. Loeb Classical Library 366. Cambridge, MA: Harvard University Press.
- Artavanis-Tsakonas, S., and Muskavitch, M. A. (2010). Notch: the past, the present, and the future. *Curr. Top. Dev. Biol.* 92, 1–29. doi: 10.1016/S0070-2153(10)92001-2
- Bailey, J. M., Mohr, A. M., and Hollingsworth, M. A. (2009). Sonic hedgehog paracrine signaling regulates metastasis and lymphangiogenesis in pancreatic cancer. *Oncogene* 28, 3513–3525. doi: 10.1038/onc.2009.220
- Barsi, J. C., Li, E., and Davidson, E. H. (2015). Geometric control of ciliated band regulatory states in the sea urchin embryo. *Development* 142, 953–961. doi: 10.1242/dev.117986
- Beachy, P. A., Karhadkar, S. S., and Berman, D. M. (2004). Tissue repair and stem cell renewal in carcinogenesis. *Nature* 432, 324–331. doi: 10.1038/nature03100
- Beane, W. S., Gross, J. M., and McClay, D. R. (2006). RhoA regulates initiation of invagination, but not convergent extension, during sea urchin gastrulation. *Dev. Biol.* 292, 213–225. doi: 10.1016/j.ydbio.2005.12.031
- Bellé, R., Le Bouffant, R., Morales, J., Cosson, B., Cormier, P., and Mulner-Lorillon, O. (2007). L'embryon d'oursin, le point de surveillance de l'ADN endommagé de la division cellulaire et les mécanismes à l'origine de la cancérisation [Sea urchin embryo, DNA-damaged cell cycle checkpoint and the mechanisms initiating cancer development]. *J. Soc. Biol.* 201, 317–327. doi: 10.1051/jbio:2007030
- Belyi, V. A., Ak, P., Markert, E., Wang, H., Hu, W., Puzio-Kuter, A., et al. (2010). The origins and evolution of the p53 family of genes. *Cold Spring Harb. Perspect. Biol.* 2:a001198. doi: 10.1101/cshperspect.a001198
- Bodnar, A. G., and Coffman, J. A. (2016). Maintenance of somatic tissue regeneration with age in short- and long-lived species of sea urchins. *Aging Cell* 15, 778–787. doi: 10.1111/acel.12487

- Boon, R. A., Jaé, N., Holdt, L., and Dimmeler, S. (2016). Long noncoding RNAs: from clinical genetics to therapeutic targets? *J. Am. Coll. Cardiol.* 67, 1214–1226. doi: 10.1016/j.jacc.2015.12.051
- Bradham, C. A., Foltz, K. R., Beane, W. S., Arnone, M. I., Rizzo, F., Coffman, J. A., et al. (2006). The sea urchin kinome: a first look. *Dev. Biol.* 300, 180–193. doi: 10.1016/j.ydbio.2006.08.074
- Bray, S. J. (2016). Notch signalling in context. *Nat. Rev. Mol. Cell Biol.* 17, 722–735. doi: 10.1038/nrm.2016.94
- Briggs, E., and Wessel, G. M. (2006). In the beginning...animal fertilization and sea urchin development. *Dev. Biol.* 300, 15–26. doi: 10.1016/j.ydbio.2006.07.014
- Briggs, J. A., Wolvetang, E. J., Mattick, J. S., Rinn, J. L., and Barry, G. (2015). Mechanisms of long non-coding RNAs in mammalian nervous system development, plasticity, disease, and evolution. *Neuron* 88, 861–877. doi: 10.1016/j.neuron.2015.09.045
- Brümmer, A., and Hausser, J. (2014). MicroRNA binding sites in the coding region of mRNAs: extending the repertoire of post-transcriptional gene regulation. *Bioessays* 36, 617–626. doi: 10.1002/bies.201300104
- Byrum, C. A., Walton, K. D., Robertson, A. J., Carbonneau, S., Thomason, R. T., Coffman, J. A., et al. (2006). Protein tyrosine and serine-threonine phosphatases in the sea urchin, *Strongylocentrotus purpuratus*: identification and potential functions. *Dev. Biol.* 300, 194–218. doi: 10.1016/j.ydbio.2006.08.050
- Cai, Y., Yu, X., Hu, S., and Yu, J. (2009). A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* 7, 147–154. doi: 10.1016/S1672-0229(08)60044-3
- Cameron, R. A. (2014). Tools for sea urchin genomic analysis. *Methods Mol. Biol.* 1128, 295–310. doi: 10.1007/978-1-62703-974-1_20
- Campo-Paysaa, F., Sémon, M., Cameron, R. A., Peterson, K. J., and Schubert, M. (2011). microRNA complements in deuterostomes: origin and evolution of microRNAs. *Evol. Dev.* 13, 15–27. doi: 10.1111/j.1525-142X.2010.00452.x
- Cannon, J. T., Vellutini, B. C., Smith, J. 3rd, Ronquist, F., Jondelius, U., and Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature* 530, 89–93. doi: 10.1038/nature16520
- Carey, N. (2017). *Junk DNA: A journey through the dark matter of the genome*. Columbia University Press.
- Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., et al. (2005). The transcriptional landscape of the mammalian genome. *Science* 309, 1559–1563. doi: 10.1126/science.1112014
- Carroll, A. P., Goodall, G. J., and Liu, B. (2014). Understanding principles of miRNA target recognition and function through integrated biological and bioinformatics approaches. *Wiley Interdiscip. Rev. RNA* 5, 361–379. doi: 10.1002/wrna.1217
- Chatterjee, D., and Deng, W. -M. (2019). Drosophila model in cancer: an introduction. *Adv. Exp. Med. Biol.* 1167, 1–14. doi: 10.1007/978-3-030-23629-8_1
- Chen, X., Yan, C. C., Zhang, X., and You, Z. H. (2017). Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief. Bioinform.* 18, 558–576. doi: 10.1093/bib/bbw060
- Chiarelli, R., Martino, C., Agnello, M., Bosco, L., and Roccheri, M. C. (2016). Autophagy as a defense strategy against stress: focus on *Paracentrotus lividus* sea urchin embryos exposed to cadmium. *Cell Stress Chaperones* 21, 19–27. doi: 10.1007/s12192-015-0639-3
- Clements, W. M., Wang, J., Sarnaik, A., Kim, O. J., MacDonald, J., Fenoglio-Preiser, C., et al. (2002). Beta-catenin mutation is a frequent cause of Wnt pathway activation in gastric cancer. *Cancer Res.* 62, 3503–3506.
- Croce, J. C., and McClay, D. R. (2010). Dynamics of delta/notch signaling on endomesoderm segregation in the sea urchin embryo. *Development* 137, 83–91. doi: 10.1242/dev.044149
- Croce, J. C., Wu, S. Y., Byrum, C., Xu, R., Duloquin, L., Wikramanayake, A. H., et al. (2006). A genome-wide survey of the evolutionarily conserved Wnt pathways in the sea urchin *Strongylocentrotus purpuratus*. *Dev. Biol.* 300, 121–131. doi: 10.1016/j.ydbio.2006.08.045
- Davidson, E. H. (2006). The sea urchin genome: where will it lead us? *Science* 314, 939–940. doi: 10.1126/science.1136252
- Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Caletani, C., Yuh, C. -H., et al. (2002). A genomic regulatory network for development. *Science* 295, 1669–1678. doi: 10.1126/science.1069883
- De Wert, G., Berghmans, R. L., Boer, G. J., Andersen, S., Brambati, B., Carvalho, A. S., et al. (2002). Ethical guidance on human embryonic and fetal tissue transplantation: a European overview. *Med. Health Care Philos.* 5, 79–90. doi: 10.1023/A:1014213125573
- De Wert, G., and Mummery, C. (2003). Human embryonic stem cells: research, ethics and policy. *Hum. Reprod.* 18, 672–682. doi: 10.1093/humrep/deg143
- Dean, M., and Annino, T. (2005). Evolution of the ATP-binding cassette (ABC) transporter superfamily in vertebrates. *Annu. Rev. Genomics Hum. Genet.* 6, 123–142. doi: 10.1146/annurev.genom.6.080604.162122
- Dean, M., Rzhetsky, A., and Allikmets, R. (2001). The human ATP-binding cassette (ABC) transporter superfamily. *Genome Res.* 11, 1156–1166. doi: 10.1101/gr.184901
- Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., et al. (2012). Landscape of transcription in human cells. *Nature* 489, 101–108. doi: 10.1038/nature11233
- Ebert, T. A. (2010). Demographic patterns of the purple sea urchin *Strongylocentrotus purpuratus* along a latitudinal gradient, 1985–1987. *Mar. Ecol. Prog. Ser.* 406, 105–120. doi: 10.3354/meps08547
- Ely, K. A., Bischoff, L. A., and Weiss, V. L. (2018). Wnt signaling in thyroid homeostasis and carcinogenesis. *Genes* 9:204. doi: 10.3390/genes9040204
- ENCODE Project Consortium Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi: 10.1038/nature05874
- Evans, T., Rosenthal, E. T., Youngblom, J., Distel, D., and Hunt, T. (1983). Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell* 33, 389–396. doi: 10.1016/0092-8674(83)90420-8
- Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annu. Rev. Biochem.* 79, 351–379. doi: 10.1146/annurev-biochem-060308-103103
- Fehon, R. G., Kooh, P. J., Rebay, I., Regan, C. L., Xu, T., Muskavitch, M. A., et al. (1990). Molecular interactions between the protein products of the neurogenic loci notch and delta, two EGF-homologous genes in Drosophila. *Cell* 61, 523–534. doi: 10.1016/0092-8674(90)90534-1
- Fernandez-Guerra, A., Aze, A., Morales, J., Mulner-Lorillon, O., Cosson, B., Cormier, P., et al. (2006). The genomic repertoire for cell cycle control and DNA metabolism in *S. purpuratus*. *Dev. Biol.* 300, 238–251. doi: 10.1016/j.ydbio.2006.09.012
- Finnegan, E. J., and Matzke, M. A. (2003). The small RNA world. *J. Cell Sci.* 116, 4689–4693. doi: 10.1242/jcs.00838
- Flanagan, D. J., Pheesse, T. J., Barker, N., Schwab, R. H., Amin, N., Malaterre, J., et al. (2015). Frizzled7 functions as a Wnt receptor in intestinal epithelial Lgr5(+) stem cells. *Stem Cell Reports* 4, 759–767. doi: 10.1016/j.stemcr.2015.03.003
- Flanagan, D. J., Vincan, E., and Pheesse, T. J. (2017). Winding back Wnt signalling: potential therapeutic targets for treating gastric cancers. *Br. J. Pharmacol.* 174, 4666–4683. doi: 10.1111/bph.13890
- Fu, P. F., Zheng, X., Fan, X., and Lin, A. F. (2019). Role of cytoplasmic lncRNAs in regulating cancer signaling pathways. *J. Zhejiang Univ. Sci. B* 20, 1–8. doi: 10.1631/jzus.B1800254
- Garneau, N. L., Wilusz, J., and Wilusz, C. J. (2007). The highways and byways of mRNA decay. *Nat. Rev. Mol. Cell Biol.* 8, 113–126. doi: 10.1038/nrm2104
- Gildor, T., Malik, A., Sher, N., Avraham, L., and De-Leon, S. B. -T. (2016). Quantitative developmental transcriptomes of the Mediterranean sea urchin *Paracentrotus lividus*. *Mar. Genomics* 25, 89–94. doi: 10.1016/j.margen.2015.11.013
- Giudice, G. (1973). *Developmental biology of the sea urchin embryo*. New York: Academic Press.
- Giudice, G. (1962). Restitution of whole larvae from disaggregated cells of sea urchin embryos. *Dev. Biol.* 5, 402–411. doi: 10.1016/0012-1606(62)90021-0
- Goldstone, J. V., Hamdoun, A., Cole, B. J., Howard-Ashby, M., Nebert, D. W., Scally, M., et al. (2006). The chemical defenseome: environmental sensing and response genes in the *Strongylocentrotus purpuratus* genome. *Dev. Biol.* 300, 366–384. doi: 10.1016/j.ydbio.2006.08.066
- Gordon, W. R., Arnett, K. L., and Blacklow, S. C. (2008). The molecular logic of notch signaling--a structural and biochemical perspective. *J. Cell Sci.* 121, 3109–3119. doi: 10.1242/jcs.035683
- Gordon, R. E., Zhang, L., Peri, S., Kuo, Y. M., Du, F., Egleston, B. L., et al. (2018). Statins synergize with hedgehog pathway inhibitors for treatment of medulloblastoma. *Clin. Cancer Res.* 24, 1375–1388. doi: 10.1158/1078-0432.CCR-17-2923

- Grewal, S. I., and Elgin, S. C. (2007). Transcription and RNA interference in the formation of heterochromatin. *Nature* 447, 399–406. doi: 10.1038/nature05914
- Gutzmer, R., and Solomon, J. A. (2019). Hedgehog pathway inhibition for the treatment of basal cell carcinoma. *Target. Oncol.* 14, 253–267. doi: 10.1007/s11523-019-00648-2
- Hadijargyrou, M., and Delihis, N. (2013). The intertwining of transposable elements and non-coding RNAs. *Int. J. Mol. Sci.* 14, 13307–13328. doi: 10.3390/ijms140713307
- Hamdoun, A. M., Cherr, G. N., Roepke, T. A., and Epel, D. (2004). Activation of multidrug efflux transporter activity at fertilization in sea urchin embryos (*Strongylocentrotus purpuratus*). *Dev. Biol.* 276, 452–462. doi: 10.1016/j.ydbio.2004.09.013
- Hanahan, D., and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* 100, 57–70. doi: 10.1016/S0092-8674(00)81683-9
- Hedges, S. B. (2002). The origin and evolution of model organisms. *Nat. Rev. Genet.* 3, 838–849. doi: 10.1038/nrg929
- Henrique, D., and Schweisguth, F. (2019). Mechanisms of notch signaling: a simple logic deployed in time and space. *Development* 146:dev172148. doi: 10.1242/dev.172148
- Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., and Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep.* 11, 1110–1122. doi: 10.1016/j.celrep.2015.04.023
- Hibino, T., Loza-Coll, M., Messier, C., Majeske, A. J., Cohen, A. H., Terwilliger, D. P., et al. (2006). The immune gene repertoire encoded in the purple sea urchin genome. *Dev. Biol.* 300, 349–365. doi: 10.1016/j.ydbio.2006.08.065
- Hinegardner, R. T. (1971). An improved fluorometric assay for DNA. *Anal. Biochem.* 39, 197–201. doi: 10.1016/0003-2697(71)90476-3
- Holm, S. (2003). The ethical case against stem cell research. *Camb. Q. Healthc. Ethics* 12, 372–383. doi: 10.1017/S0963180103124061
- Horstadius, S. (1939). The mechanics of sea urchin development, studied by operative methods. *Biol. Rev.* 14, 132–179. doi: 10.1111/j.1469-185X.1939.tb00929.x
- Howard-Ashby, M., Materna, S. C., Brown, C. T., Chen, L., Cameron, R. A., and Davidson, E. H. (2006). Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. *Dev. Biol.* 300, 90–107. doi: 10.1016/j.ydbio.2006.08.033
- Huarte, M. (2015). The emerging role of lncRNAs in cancer. *Nat. Med.* 21, 1253–1261. doi: 10.1038/nm.3981
- Israel, J. W., Martik, M. L., Byrne, M., Raff, E. C., Raff, R. A., McClay, D. R., et al. (2016). Comparative developmental transcriptomics reveals rewiring of a highly conserved gene regulatory network during a major life history switch in the sea urchin genus *heliocidarid*. *PLoS Biol.* 14:e1002391. doi: 10.1371/journal.pbio.1002391
- Janies, D. A., Witter, Z., Linchangco, G. V., Foltz, D. W., Miller, A. K., Kerr, A. M., et al. (2016). EchinoDB, an application for comparative transcriptomics of deeply-sampled clades of echinoderms. *BMC Bioinformatics* 17:48. doi: 10.1186/s12859-016-0883-2
- Karhadkar, S. S., Bova, G. S., Abdallah, N., Dhara, S., Gardner, D., Maitra, A., et al. (2004). Hedgehog signalling in prostate regeneration, neoplasia and metastasis. *Nature* 431, 707–712. doi: 10.1038/nature02962
- Kazanian, H. H. Jr., and Moran, J. V. (2017). Mobile DNA in health and disease. *N. Engl. J. Med.* 377, 361–370. doi: 10.1056/NEJMra1510092
- Kober, K. M., and Bernardi, G. (2013). Phylogenomics of stronglycentrotid sea urchins. *BMC Evol. Biol.* 13:88. doi: 10.1186/1471-2148-13-88
- Konishi, J., Yi, F., Chen, X., Vo, H., Carbone, D. P., and Dang, T. P. (2010). Notch3 cooperates with the EGFR pathway to modulate apoptosis through the induction of bim. *Oncogene* 29, 589–596. doi: 10.1038/nc.2009.366
- Kontomanolis, E. N., Kalagasidou, S., Pouliliou, S., Anthoulaki, X., Georgiou, N., Papamanolis, V., et al. (2018). The notch pathway in breast cancer progression. *ScientificWorldJournal* 2018:2415489. doi: 10.1155/2018/2415489
- Kumburegama, S., and Wikramanayake, A. H. (2008). Wnt signaling in the early sea urchin embryo. *Methods Mol. Biol.* 469, 187–199. doi: 10.1007/978-1-60327-469-2_14
- Kuo, C. -J., Hansen, M., and Troemel, E. (2018). Autophagy and innate immunity: insights from invertebrate model organisms. *Autophagy* 14, 233–242. doi: 10.1080/15548627.2017.1389824
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi: 10.1038/35057062
- Langton, P. F., Kakugawa, S., and Vincent, J. P. (2016). Making, exporting, and modulating Wnts. *Trends Cell Biol.* 26, 756–765. doi: 10.1016/j.tcb.2016.05.011
- Laubichler, M. D., and Davidson, E. H. (2008). Boveri's long experiment: sea urchin merogones and the establishment of the role of nuclear chromosomes in development. *Dev. Biol.* 314, 1–11. doi: 10.1016/j.ydbio.2007.11.024
- Laumer, C. E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R. C., Sørensen, M. V., et al. (2015). Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol.* 25, 2000–2006. doi: 10.1016/j.cub.2015.06.068
- Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The landscape of long noncoding RNA classification. *Trends Genet.* 31, 239–251. doi: 10.1016/j.tig.2015.03.007
- Lebedev, E. E., Ostromyshenskii, D. I., Solovyeva, A. I., Turenkod, A. S., Drozdov, A. L., Podgornaya, O. I., et al. (2019). The transposons of the sea urchin *strongylocentrotus intermedius* agassiz, 1863: in silico versus in vitro. *Russ. J. Mar. Biol.* 45, 418–424. doi: 10.1134/S1063074019060051
- Lee, I., Ajay, S. S., Yook, J. I., Kim, H. S., Hong, S. H., Kim, N. H., et al. (2009). New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. *Genome Res.* 19, 1175–1183. doi: 10.1101/gr.089367.108
- Lefort, K., Mandinova, A., Ostano, P., Kolev, V., Calpini, V., Kolfschoten, I., et al. (2007). Notch1 is a p53 target gene involved in human keratinocyte tumor suppression through negative regulation of ROCK1/2 and MRCKalpha kinases. *Genes Dev.* 21, 562–577. doi: 10.1101/gad.1484707
- Lehmann, W., Mossman, D., Kleemann, J., Mock, K., Meisinger, C., Brummer, T., et al. (2016). ZEB1 turns into a transcriptional activator by interacting with YAP1 in aggressive cancer types. *Nat. Commun.* 7:10498. doi: 10.1038/ncomms10498
- Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* 120, 15–20. doi: 10.1016/j.cell.2004.12.035
- Li, W., and Kang, Y. (2014). A new lnc in metastasis: long noncoding RNA mediates the prometastatic functions of TGF- β . *Cancer Cell* 25, 557–559. doi: 10.1016/j.ccr.2014.04.014
- Li, J., and Liu, C. (2019). Coding or noncoding, the converging concepts of RNAs. *Front. Genet.* 10:496. doi: 10.3389/fgene.2019.00496
- Lim, J., and Thiery, J. P. (2012). Epithelial-mesenchymal transitions: insights from development. *Development* 139, 3471–3486. doi: 10.1242/dev.071209
- Liu, S., Dontu, G., Mantle, I. D., Patel, S., Ahn, N. S., Jackson, K. W., et al. (2006). Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells. *Cancer Res.* 66, 6063–6071. doi: 10.1158/0008-5472.CAN-06-0054
- Lloyd-Lewis, B., Mourikis, P., and Fre, S. (2019). Notch signalling: sensor and instructor of the microenvironment to coordinate cell fate and organ morphogenesis. *Curr. Opin. Cell Biol.* 61, 16–23. doi: 10.1016/j.cub.2019.06.003
- Lo, B., and Parham, L. (2009). Ethical issues in stem cell research. *Endocr. Rev.* 30, 204–213. doi: 10.1210/er.2008-0031
- Long, Y., Wang, X., Youmans, D. T., and Cech, T. R. (2017). How do lncRNAs regulate transcription? *Sci. Adv.* 3:eaa02110. doi: 10.1126/sciadv.aao2110
- Lowry, W. E., Richter, L., Yachechko, R., Pyle, A. D., Tchieu, J., Sridharan, R., et al. (2008). Generation of human induced pluripotent stem cells from dermal fibroblasts. *Proc. Natl. Acad. Sci. U. S. A.* 105, 2883–2888. doi: 10.1073/pnas.0711983105
- Lu, W., Liu, C. C., Thottassery, J. V., Bu, G., and Li, Y. (2010). Mesd is a universal inhibitor of Wnt coreceptors LRP5 and LRP6 and blocks Wnt/beta-catenin signaling in cancer cells. *Biochemistry* 49, 4635–4643. doi: 10.1021/bi1001486
- Ma, J., Xia, J., Miele, L., Sarkar, F. H., and Wang, Z. (2013). Notch signaling pathway in pancreatic cancer progression. *Pancreat. Disord. Ther.* 3:1000114.
- Marques, I. J., Lupi, E., and Mercader, N. (2019). Model systems for regeneration: zebrafish. *Development* 146:dev167692. doi: 10.1242/dev.167692
- Martik, M. L., Lyons, D. C., and McClay, D. R. (2016). Developmental gene regulatory networks in sea urchins and what we can learn from them. *F1000Res.* 5:203. doi: 10.12688/f1000research.7381.1
- McClay, D. R. (2011). Evolutionary crossroads in developmental biology: sea urchins. *Development* 138, 2639–2648. doi: 10.1242/dev.048967
- McClay, D. R., and Fink, R. D. (1982). Sea urchin hyalin: appearance and function in development. *Dev. Biol.* 92, 285–293. doi: 10.1016/0012-1606(82)90175-0

- Mzoughi, S., Zhang, J., Hequet, D., Teo, S. X., Fang, H., Xing, Q. R., et al. (2017). PRDM15 safeguards naive pluripotency by transcriptionally regulating WNT and MAPK-ERK signaling. *Nat. Genet.* 49, 1354–1363. doi: 10.1038/ng.3922
- Nakajima, Y., Kaneko, H., Murray, G., and Burke, R. D. (2004). Divergent patterns of neural development in larval echinoids and asteroids. *Evol. Dev.* 6, 95–104. doi: 10.1111/j.1525-142x.2004.04011.x
- Nguyen, V., Hough, R., Bernaud, S., and Peng, C. (2019). Wnt/ β -catenin signalling in ovarian cancer: insights into its hyperactivation and function in tumorigenesis. *J. Ovarian Res.* 12:122. doi: 10.1186/s13048-019-0596-z
- Nielsen, J. (2019). Yeast systems biology: model organism and cell factory. *Biotechnol. J.* 14:e1800421. doi: 10.1002/biot.201800421
- Nieto, M. A., Huang, R. Y., Jackson, R. A., and Thiery, J. P. (2016). EMT: 2016. *Cell* 166, 21–45. doi: 10.1016/j.cell.2016.06.028
- Nowell, C. S., and Radtke, F. (2017). Notch as a tumour suppressor. *Nat. Rev. Cancer* 17, 145–159. doi: 10.1038/nrc.2016.145
- Nusse, R., and Clevers, H. (2017). Wnt/ β -catenin signaling, disease, and emerging therapeutic modalities. *Cell* 169, 985–999. doi: 10.1016/j.cell.2017.05.016
- Oliveri, P., Tu, Q., and Davidson, E. H. (2008). Global regulatory logic for specification of an embryonic cell lineage. *Proc. Natl. Acad. Sci. U. S. A.* 105, 5955–5962. doi: 10.1073/pnas.0711220105
- Ong, M. S., Cai, W., Yuan, Y., Leong, H. C., Tan, T. Z., Mohammad, A., et al. (2017). 'Lnc'-ing Wnt in female reproductive cancers: therapeutic potential of long non-coding RNAs in Wnt signalling. *Br. J. Pharmacol.* 174, 4684–4700. doi: 10.1111/bph.13958
- Paço, A., Freitas, R., and Vieira-da-Silva, A. (2019). Conversion of DNA sequences: from a transposable element to a tandem repeat or to a gene. *Gene* 10:1014. doi: 10.3390/genes10121014
- Parr, C., Watkins, G., and Jiang, W. G. (2004). The possible correlation of Notch-1 and Notch-2 with clinical outcome and tumour clinicopathological parameters in human breast cancer. *Int. J. Mol. Med.* 14, 779–786. doi: 10.3892/ijmm.14.5.779
- Patel, N. S., Li, J. L., Generali, D., Poulson, R., Cranston, D. W., and Harris, A. L. (2005). Up-regulation of delta-like 4 ligand in human tumor vasculature and the role of basal expression in endothelial cell function. *Cancer Res.* 65, 8690–8697. doi: 10.1158/0008-5472.CAN-05-1208
- Peng, W. X., Koirala, P., and Mo, Y. Y. (2017). LncRNA-mediated regulation of cell signaling in cancer. *Oncogene* 36, 5661–5667. doi: 10.1038/onc.2017.184
- Pérez-Portela, R., Turon, X., and Riesgo, A. (2016). Characterization of the transcriptome and gene expression of four different tissues in the ecologically relevant sea urchin *Arbacia lixula* using RNA-seq. *Mol. Ecol. Resour.* 16, 794–808. doi: 10.1111/1755-0998.12500
- Peter, I. S., and Davidson, E. H. (2010). The endoderm gene regulatory network in sea urchin embryos up to mid-blastula stage. *Dev. Biol.* 340, 188–199. doi: 10.1016/j.ydbio.2009.10.037
- Peter, I. S., and Davidson, E. H. (2017). Assessing regulatory information in developmental gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* 114, 5862–5869. doi: 10.1073/pnas.1610616114
- Peter, I. S., Faure, E., and Davidson, E. H. (2012). Predictive computation of genomic logic processing functions in embryonic development. *Proc. Natl. Acad. Sci. U. S. A.* 109, 16434–16442. doi: 10.1073/pnas.1207852109
- Peterson, K. J., Dietrich, M. R., and McPeck, M. A. (2009). MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays* 31, 736–747. doi: 10.1002/bies.200900033
- Pires-daSilva, A., and Sommer, R. J. (2003). The evolution of signaling pathways in animal development. *Nat. Rev. Genet.* 4, 39–49. doi: 10.1038/nrg977
- Poli, V., Fagnocchi, L., and Zippo, A. (2018). Tumorigenic cell reprogramming and cancer plasticity: interplay between signaling, microenvironment, and epigenetics. *Stem Cells Int.* 2018:4598195. doi: 10.1155/2018/4598195
- Rafiq, K., Shashikant, T., McManus, C. J., and Ettensohn, C. A. (2014). Genome-wide analysis of the skeletal gene regulatory network of sea urchins. *Development* 141, 950–961. doi: 10.1242/dev.105585
- Rast, J. P., Smith, L. C., Loza-Coll, M., Hibino, T., and Litman, G. W. (2006). Genomic insights into the immune system of the sea urchin. *Science* 314, 952–956. doi: 10.1126/science.1134301
- Reya, T., Morrison, S. J., Clarke, M. F., and Weissman, I. L. (2001). Stem cells, cancer, and cancer stem cells. *Nature* 414, 105–111. doi: 10.1038/35102167
- Robert, N., Hammami, F., Lhomond, G., Dru, P., Lepage, T., Schubert, M., et al. (2019). A wnt2 ortholog in the sea urchin *Paracentrotus lividus*. *Genesis* 57:e23331. doi: 10.1002/dvg.23331
- Robert, N., Lhomond, G., Schubert, M., and Croce, J. C. (2014). A comprehensive survey of wnt and frizzled expression in the sea urchin *Paracentrotus lividus*. *Genesis* 52, 235–250. doi: 10.1002/dvg.22754
- Romancino, D. P., Anello, L., Lavanco, A., Buffa, V., Di Bernardo, M., and Bongiovanni, A. (2017). A sea urchin in vivo model to evaluate epithelial-mesenchymal transition. *Dev. Growth Differ.* 59, 141–151. doi: 10.1111/dgd.12353
- Rzepnikowska, W., Flis, K., Muñoz-Braceras, S., Menezes, R., Escalante, R., and Zoladek, T. (2017). Yeast and other lower eukaryotic organisms for studies of Vps13 proteins in health and disease. *Traffic* 18, 711–719. doi: 10.1111/tra.12523
- Sari, I. N., Phi, L., Jun, N., Wijaya, Y. T., Lee, S., and Kwon, H. Y. (2018). Hedgehog signaling in cancer: a prospective therapeutic target for eradicating cancer stem cells. *Cells* 7:208. doi: 10.3390/cells7110208
- Sea Urchin Genome Sequencing Consortium, Sodergren, E., Weinstock, G. M., Davidson, E. H., Cameron, R. A., Gibbs, R. A., et al. (2006). The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* 314, 941–952. doi: 10.1126/science.1133609
- Segditsas, S., and Tomlinson, I. (2006). Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* 25, 7531–7537. doi: 10.1038/sj.onc.1210059
- Shahabi, S., Kumaran, V., Castillo, J., Cong, Z., Nandagopal, G., Mullen, D. J., et al. (2019). LINC00261 is an epigenetically regulated tumor suppressor essential for activation of the DNA damage response. *Cancer Res.* 79, 3050–3062. doi: 10.1158/0008-5472.CAN-18-2034
- Shen, P., Pichler, M., Chen, M., Calin, G. A., and Ling, H. (2017). To Wnt or lose: the missing non-coding linc in colorectal cancer. *Int. J. Mol. Sci.* 18:2003. doi: 10.3390/ijms18092003
- Sheng, T., Li, C., Zhang, X., Chi, S., He, N., Chen, K., et al. (2004). Activation of the hedgehog pathway in advanced prostate cancer. *Mol. Cancer* 3:29. doi: 10.1186/1476-4598-3-29
- Sherwood, D. R., and McClay, D. R. (1999). LvNotch signaling mediates secondary mesenchyme specification in the sea urchin embryo. *Development* 126, 1703–1713.
- Sherwood, D. R., and McClay, D. R. (2001). LvNotch signaling plays a dual role in regulating the position of the ectoderm-endoderm boundary in the sea urchin embryo. *Development* 128, 2221–2232.
- Skoda, A. M., Simovic, D., Karin, V., Kardum, V., Vranic, S., and Serman, L. (2018). The role of the hedgehog signaling pathway in cancer: a comprehensive review. *Bosn. J. Basic Med. Sci.* 18, 8–20. doi: 10.17305/bjbm.2018.2756
- Song, J. L., Stoekius, M., Maaskola, J., Friedländer, M., Stepicheva, N., Juliano, C., et al. (2012). Select microRNAs are essential for early development in the sea urchin. *Dev. Biol.* 362, 104–113. doi: 10.1016/j.ydbio.2011.11.015
- Stecca, B., Mas, C., Clement, V., Zbinden, M., Correa, R., Piguet, V., et al. (2007). Melanomas require HEDGEHOG-GLI signaling regulated by interactions between GLI1 and the RAS-MEK/AKT pathways. *Proc. Natl. Acad. Sci. U. S. A.* 104, 5895–5900. doi: 10.1073/pnas.0700776104
- Steitz, J. A., and Vasudevan, S. (2009). miRNPs: versatile regulators of gene expression in vertebrate cells. *Biochem. Soc. Trans.* 37, 931–935. doi: 10.1042/BST0370931
- Stepicheva, N. A., and Song, J. L. (2015). microRNA-31 modulates skeletal patterning in the sea urchin embryo. *Development* 142, 3769–3780. doi: 10.1242/dev.127969
- Stimpson, W. (1857). On the Crustacea and Echinodermata of the Pacific shores of North America. *Boston J. Nat. Hist.* 6, 444–532.
- Sun, T., Jiao, L., Wang, Y., Yu, Y., and Ming, L. (2018). SIRT1 induces epithelial-mesenchymal transition by promoting autophagic degradation of E-cadherin in melanoma cells. *Cell Death Dis.* 9:136. doi: 10.1038/s41419-017-0167-4
- Sweet, H. C., Gehring, M., and Ettensohn, C. A. (2002). LvDelta is a mesoderm-inducing signal in the sea urchin embryo and can endow blastomeres with organizer-like properties. *Development* 129, 1945–1955.
- Tang, J., Xu, Z., Huang, L., Luo, H., and Zhu, X. (2019). Transcriptional regulation in model organisms: recent progress and clinical implications. *Open Biol.* 9:190183. doi: 10.1098/rsob.190183
- Telford, M. J., Budd, G. E., and Philippe, H. (2015). Phylogenomic insights into animal evolution. *Curr. Biol.* 25, R876–R887. doi: 10.1016/j.cub.2015.07.060

- Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M. A., del Campo, J., et al. (2015). Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Curr. Biol.* 25, 2404–2410. doi: 10.1016/j.cub.2015.07.053
- Trimarchi, T., Bilal, E., Ntziachristos, P., Fabbri, G., Dalla-Favera, R., Tsigiris, A., et al. (2014). Genome-wide mapping and characterization of notch-regulated long noncoding RNAs in acute leukemia. *Cell* 158, 593–606. doi: 10.1016/j.cell.2014.05.049
- Tu, Q., Cameron, R. A., and Davidson, E. H. (2014). Quantitative developmental transcriptomes of the sea urchin *Strongylocentrotus purpuratus*. *Dev. Biol.* 385, 160–167. doi: 10.1016/j.ydbio.2013.11.019
- Tulchinsky, E., Demidov, O., Kriajevska, M., Barlev, N. A., and Imyaninov, E. (2019). EMT: a mechanism for escape from EGFR-targeted therapy in lung cancer. *Biochim. Biophys. Acta Rev. Cancer* 1871, 29–39. doi: 10.1016/j.bbcan.2018.10.003
- Venkatesh, V., Nataraj, R., Thangaraj, G. S., Karthikeyan, M., Gnanasekaran, A., Kagineili, S. B., et al. (2018). Targeting notch signalling pathway of cancer stem cells. *Stem Cell Investig.* 5:5. doi: 10.21037/sci.2018.02.02
- Viatour, P., Ehmer, U., Saddic, L. A., Dorrell, C., Andersen, J. B., Lin, C., et al. (2011). Notch signaling inhibits hepatocellular carcinoma following inactivation of the RB pathway. *J. Exp. Med.* 208, 1963–1976. doi: 10.1084/jem.20110198
- Vinson, K. E., George, D. C., Fender, A. W., Bertrand, F. E., and Sigounas, G. (2016). The notch pathway in colorectal cancer. *Int. J. Cancer* 138, 1835–1842. doi: 10.1002/ijc.29800
- Wang, Z., and Kunze, R. (2015). “Transposons in eukaryotes: structures, mechanisms and applications” in *eLS*. Chichester: John Wiley & Sons, Ltd.
- Wang, Q., Zhou, Y., Rychahou, P., Harris, J. W., Zaytseva, Y. Y., Liu, J., et al. (2018). Deptor is a novel target of Wnt/ β -catenin/c-Myc and contributes to colorectal cancer cell growth. *Cancer Res.* 78, 3163–3175. doi: 10.1158/0008-5472.CAN-17-3107
- Warner, J. F., Miranda, E. L., and McClay, D. R. (2016). Contribution of hedgehog signaling to the establishment of left-right asymmetry in the sea urchin. *Dev. Biol.* 411, 314–324. doi: 10.1016/j.ydbio.2016.02.008
- Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., et al. (2009). The deep evolution of metazoan microRNAs. *Evol. Dev.* 11, 50–68. doi: 10.1111/j.1525-142X.2008.00302.x
- Whitaker, M. (2006). Calcium microdomains and cell cycle control. *Cell Calcium* 40, 585–592. doi: 10.1016/j.ceca.2006.08.018
- Wu, C. Y., Tsai, Y. P., Wu, M. Z., Teng, S. C., and Wu, K. J. (2012). Epigenetic reprogramming and post-transcriptional regulation during the epithelial-mesenchymal transition. *Trends Genet.* 28, 454–463. doi: 10.1016/j.tig.2012.05.005
- Yuan, J. H., Yang, F., Wang, F., Ma, J. Z., Guo, Y. J., Tao, Q. F., et al. (2014). A long noncoding RNA activated by TGF- β promotes the invasion-metastasis cascade in hepatocellular carcinoma. *Cancer Cell* 25, 666–681. doi: 10.1016/j.ccr.2014.03.010
- Zaidel-Bar, R. (2009). Evolution of complexity in the integrin adhesome. *J. Cell Biol.* 186, 17–21. doi: 10.1083/jcb.200811067
- Zhou, W., Fu, X. Q., Zhang, L. L., Zhang, J., Huang, X., Lu, X. H., et al. (2013). The AKT1/NF-kappaB/Notch1/PTEN axis has an important role in chemoresistance of gastric cancer cells. *Cell Death Dis.* 4:e847. doi: 10.1038/cddis.2013.375
- Zhou, J. X., Jia, L. W., Liu, W. M., Miao, C. L., Liu, S., Cao, Y. J., et al. (2006). Role of sonic hedgehog in maintaining a pool of proliferating stem cells in the human fetal epidermis. *Hum. Reprod.* 21, 1698–1704. doi: 10.1093/humrep/del086

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Adonin, Drozdov and Barlev. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Gene Expression Changes and Associated Pathways Involved in the Progression of Prostate Cancer Advanced Stages

Elena A. Pudova¹, George S. Krasnov¹, Anastasiya A. Kobelyatskaya¹, Maria V. Savvateeva¹, Maria S. Fedorova¹, Vladislav S. Pavlov¹, Kirill M. Nyushko², Andrey D. Kaprin², Boris Y. Alekseev², Dmitry Y. Trofimov³, Gennady T. Sukhikh³, Anastasiya V. Snezhkina¹ and Anna V. Kudryavtseva^{1*}

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Stanislav Rybtsov,
University of Edinburgh,
United Kingdom
Angelika Timofeeva,
National Medical Research Center of
Obstetrics, Gynecology and
Perinatology named after
Academician V.I. Kulakova, Russia
Vladimir Kalinin,
TMA-Science, Ireland

*Correspondence:

Anna V. Kudryavtseva
rhizamoeba@mail.ru

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 01 October 2020

Accepted: 24 December 2020

Published: 21 January 2021

Citation:

Pudova EA, Krasnov GS,
Kobelyatskaya AA, Savvateeva MV,
Fedorova MS, Pavlov VS,
Nyushko KM, Kaprin AD,
Alekseev BY, Trofimov DY,
Sukhikh GT, Snezhkina AV and
Kudryavtseva AV (2021)
Gene Expression Changes and
Associated Pathways Involved in the
Progression of Prostate Cancer
Advanced Stages.
Front. Genet. 11:613162.
doi: 10.3389/fgene.2020.613162

¹Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, ²National Medical Research Radiological Center, Ministry of Health of the Russian Federation, Moscow, Russia, ³National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I. Kulakov, Ministry of Health of the Russian Federation, Moscow, Russia

Prostate cancer (PC) is one of the most common cancers among men worldwide, and advanced PCs, such as locally advanced PC (LAPC) and castration-resistant PC (CRPC), present the greatest challenges in clinical management. Current indicators have limited capacity to predict the disease course; therefore, better prognostic markers are greatly needed. In this study, we performed a bioinformatic analysis of The Cancer Genome Atlas (TCGA) datasets, including RNA-Seq data from the prostate adenocarcinoma (PRAD; $n = 55$) and West Coast Dream Team – metastatic CRPC (WCDT-MCRPC; $n = 84$) projects, to evaluate the transcriptome changes associated with progression-free survival (PFS) for LAPC and CRPC, respectively. We identified the genes whose expression was positively/negatively correlated with PFS. In LAPC, the genes with the most significant negative correlations were *ZC2HC1A*, *SQLE*, and *KIF11*, and the genes with the most significant positive correlations were *SOD3*, *LRRC26*, *MIR22HG*, *MEG3*, and *MIR29B2CHG*. In CRPC, the most significant positive correlations were found for *BET1*, *CTAGE5*, *IFNGR1*, and *GIMAP6*, and the most significant negative correlations were found for *CLPB*, *PRPF19*, *ZNF610*, *MPST*, and *LINC02001*. In addition, we performed a gene network interaction analysis using STRINGdb, which revealed a significant relationship between genes predominantly involved in the cell cycle and characterized by upregulated expression in early recurrence. Based on the results, we propose several genes that can be used as potential prognostic markers.

Keywords: prostate cancer, gene expression, PFS, TCGA, PRAD, WCDT-MCRPC

INTRODUCTION

Prostate cancer (PC) is the second most common cancer in men worldwide (Rawla, 2019). Advanced PC presents as locally advanced and castration-resistant tumors. Locally advanced PC (LAPC) is characterized by the spread of the tumor beyond the prostate capsule and is more aggressive than localized PC. Androgen deprivation therapy aimed at reducing the level of circulating testosterone is

often used in LAPC treatment (Yap et al., 2016). However, despite rapid patient responses to this therapy, after 18–36 months, the disease frequently progresses to castration-resistant PC (CRPC; Yap et al., 2016). Metastatic CRPC is a prognostically unfavorable disease that requires regular systematic examination and monitoring and significantly impairs quality of life.

During therapy, in some patients, PC has an aggressive course, leading to metastasis, while in others, the disease has an indolent course, with a low tendency for progression. For prognostic assessment of the disease course, various clinical parameters are used, such as the level of prostate-specific antigen (PSA) and/or Gleason score; however, these parameters are not sufficiently informative (Keyes et al., 2013). To solve this problem, new molecular markers are needed that are highly associated with disease progression, which, when used in combination with existing clinical parameters, have a high predictive value. Currently, a promising area in the search for potential markers is the analysis of the most significant and consistent changes in the tumor transcriptome (Larkin et al., 2012; Cao et al., 2019; Pudova et al., 2019; Song et al., 2019).

An important criterion in the study of cancer data is progression-free survival (PFS). This parameter is the time from a random assignment in a clinical trial to disease progression or death from any cause (Forsythe et al., 2018). The study of PFS-specific molecular events will help to identify the major changes associated with the onset of disease progression.

This study aimed to analyze the transcriptome profiles of LAPC and CRPC based on the RNA-Seq data from the prostate adenocarcinoma (PRAD) and West Coast Dream Team – metastatic CRPC (WCDT-MCRPC) projects in The Cancer Genome Atlas (TCGA), respectively. We analyzed the changes in gene expression and molecular pathways related to PFS. The results of this study improve our understanding of the mechanisms underlying PC progression and identify molecules with potential as prognostic markers.

MATERIALS AND METHODS

Data Collection

The study included RNA-Seq data for LAPC samples from the PRAD project ($n = 55$) and CRPC samples from the WCDT-MCRPC project ($n = 84$), which were donated to the TCGA consortium. We focused on the study of Caucasian patients (Caucasians were identified as “white” in the databases). The trimmed mean of M-values (TMM) method was used to normalize the RNA-Seq data of each dataset. Ethical approval was not available for the study as our data were revealed from the public database.

Relative Gene Expression and Downstream Analysis

RNA-Seq data from the TCGA project (read counts per gene evaluated by HTSeq) were downloaded from the repository of GDC Data Portal¹ and then analyzed in the statistical

environment R.² Data normalization and gene expression analysis with generalized linear models was performed using edgeR package (Robinson et al., 2010). The obtained results were considered statistically significant when the p -value of the quasi-likelihood F-test (QLF test) of <0.05 . We did not use Benjamini-Hochberg (BH) p -value adjustment here because very few genes pass the threshold after it is applied.

In order to identify the genes whose expression change is most strongly associated with PFS, we evaluated Spearman's rank correlation coefficient between the normalized gene expression level and PFS. Here, the obtained results were also considered statistically significant when both Spearman's $p < 0.05$ and QLF test $p < 0.05$. For this analysis, we pre-filtered genes by the following parameters: read counts per million (CPM) > 8 and gene expression fold change is two- or (LogFC) > 1 .

Next, we performed KEGG (Kyoto Encyclopedia of Genes and Genomes database) pathway enrichment analysis using the clusterProfile package (Yu et al., 2012). Additionally, we performed interaction network reconstruction and Gene Ontology (GO) pathway enrichment analysis using the STRINGdb (Szklarczyk et al., 2017). For enrichment analyses, BH adjustment to calculate the false discovery rate (FDR) was applied. The obtained results of this analysis were considered statistically significant when the FDR-value of <0.05 . When constructing the networks, only data on direct protein-protein interactions were used (other associations such as co-expression, co-occurrence, gene fusions, and neighborhood, which are set by default, were excluded).

RESULTS

Using PRAD and WCDT-MCRPC datasets, we analyzed changes of the relative gene expression associated with PFS (the number of days to recurrence) and found 889 genes for the LAPC and 1,889 genes for CRPC with the $p < 0.05$ according to the QLF test (Supplementary Tables S1 and S2).

Pathways Enrichment Analysis

First, we focused on the pathways enrichment analysis (KEGG database) to identify the major pathways, the expression of whose members may be positively or negatively associated with PFS (Figure 1). For this analysis, we selected top-80 genes, increased expression of which was positively associated with PFS (further – “upregulated genes”) and top-80 genes with negatively associated expression (further – “downregulated genes”). In LAPC, the most statistically significant enrichment with upregulated genes was noted for “Transcriptional misregulation in cancer” pathway (hsa05202). The analysis of downregulated genes showed a significant enrichment for the “Cell cycle” pathway (hsa04110; Table 1).

For CRPC, a pathway enrichment analysis (top-250 up- and downregulated genes) revealed two pathways, “Complement and coagulation cascades” (hsa04610) and “Drug metabolism-cytochrome

¹<https://portal.gdc.cancer.gov/>

²<https://www.r-project.org/>

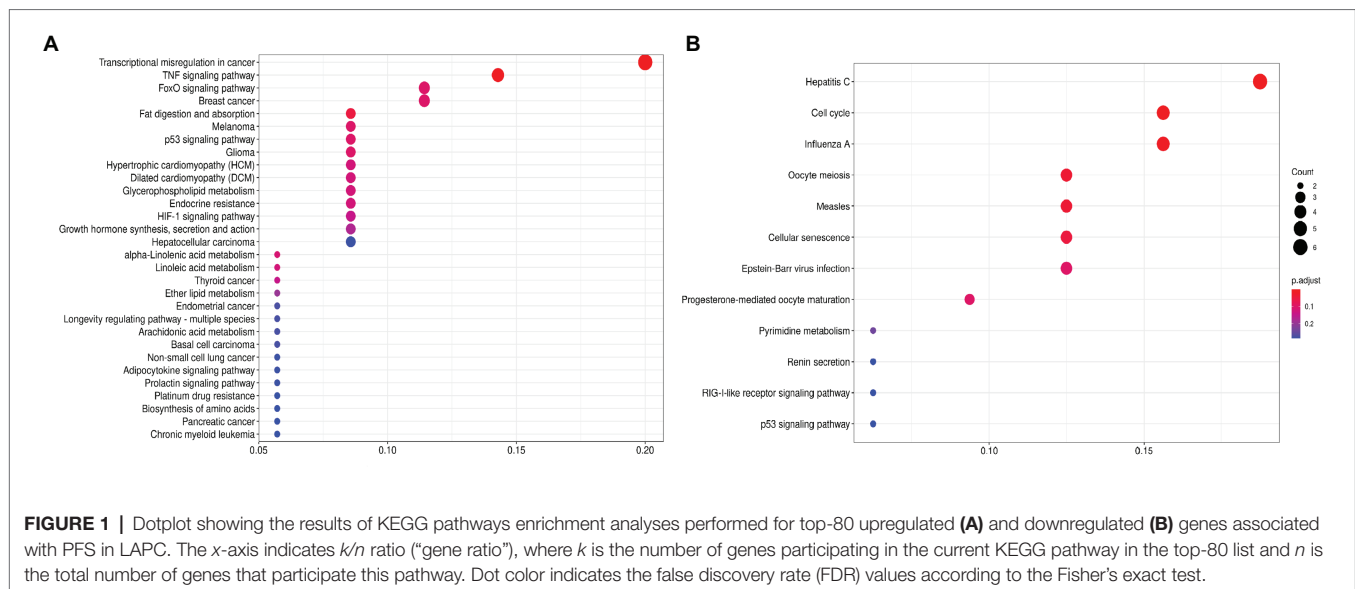


TABLE 1 | Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways that are significantly enriched with genes associated with progression-free survival (PFS) in locally advanced prostate cancer (LAPC).

KEGG ID	Pathway name	Enrichment p -value	Enrichment FDR	Genes
Enriched with upregulated genes				
hsa05202	Transcriptional misregulation in cancer	1,53E-05	1,98E-03	<i>CEBPB, RARA, GADD45B, CDKN1A, NR4A3, IGF1, SUPT3H</i>
hsa04668	TNF signaling pathway	1,12E-04	7,29E-03	<i>BCL3, CEBPB, JUNB, SOCS3, IRF1</i>
hsa04975	Fat digestion and absorption	8,31E-04	3,60E-02	<i>PLA2G5, CD36, PLA2G2A</i>
Enriched with downregulated genes				
hsa05160	Hepatitis C	2,81E-05	2,84E-03	<i>RSAD2, IFIT1, MX1, DDX58, OAS3, EIF2AK2</i>
hsa04110	Cell cycle	1,17E-04	5,89E-03	<i>E2F5, CCNB1, RAD21, CCNB2, BUB1</i>
hsa05164	Influenza A	5,08E-04	1,71E-02	<i>RSAD2, MX1, DDX58, OAS3, EIF2AK2</i>
hsa04114	Oocyte meiosis	1,57E-03	3,96E-02	<i>CALML3, CCNB1, CCNB2, BUB1</i>
hsa05162	Measles	2,07E-03	4,17E-02	<i>MX1, DDX58, OAS3, EIF2AK2</i>

P-450” (hsa00982), which were enriched with genes that have expression positively associated with PFS. When considering pathways negatively associated with PFS, we have seen an overrepresentation of genes participating in such cancer-associated pathways as the “TGF-beta signaling pathway” (hsa04350), “Hippo signaling pathway” (hsa04390), and others (Figure 2; Table 2).

Correlation Analysis Between Gene Expression and PFS

Further, a correlation analysis between the gene expression and PFS was performed in LAPC and CRPC using Spearman’s rank correlation calculation.

For the LAPC, 34 genes were found to be correlated with PFS ($p < 0.05$); an equal number of genes had positive (17/34) and negative (17/34) correlation with PFS (Figure 3).

For the CRPC, 118 genes were found, including 81 genes with positive correlation and 37 genes with negative correlation with PFS (Figure 4).

Interaction Network Analysis Using STRINGdb

To analyze the LAPC, we examined 190 genes, expression of which is negatively, and 115 genes – positively associated with PFS.

These genes passed a threshold of $p < 0.05$ (both QLF test and Spearman’s rank correlation analysis), irrespectively of CPM and $\text{abs}(\text{LogFC})$ values. A statistically significant result was obtained for genes with a negative correlation. For these genes, the network has a strong enrichment of protein-protein interactions (PPI) with $p < 1.0\text{e-}16$ (Figure 5).

The genes shown in Figure 5 encode for proteins participating in such processes as the “Cell cycle” (GO: 0007049; enrichment FDR = $2.24\text{e-}15$), “Organelle organization” (GO: 0006996; FDR = $8.53\text{e-}06$), “Response to stress” (GO: 0006950; FDR = 0.0490) terms. We have also noted the cluster with the highest number of interactions, including *RAD21*, *SGOL1*, *DYNC1LI1*, *CKAP5*, *CENPI*, *PPP2R5C*, *SKA2*, *SPC25*, *MAD2L1*, *CENPL*, *BUB1*, *ERCC6L*, *SGOL2*, and *CDCA8* genes. Most of them participate cell cycle process.

Next, we performed a protein-protein interaction network analysis using 196 genes with negative expression correlation with PFS for CRPC. The analysis (PPI enrichment $p = 0.000112$) revealed involved enrichment in genes participating the “Cell cycle” (GO: 0007049; enrichment FDR = 0.0207), “Regulation of gene expression” (GO: 0010468; FDR = 0.0215), and “Metabolic process” (GO: 0008152; FDR = 0.0345) terms (Figure 6).

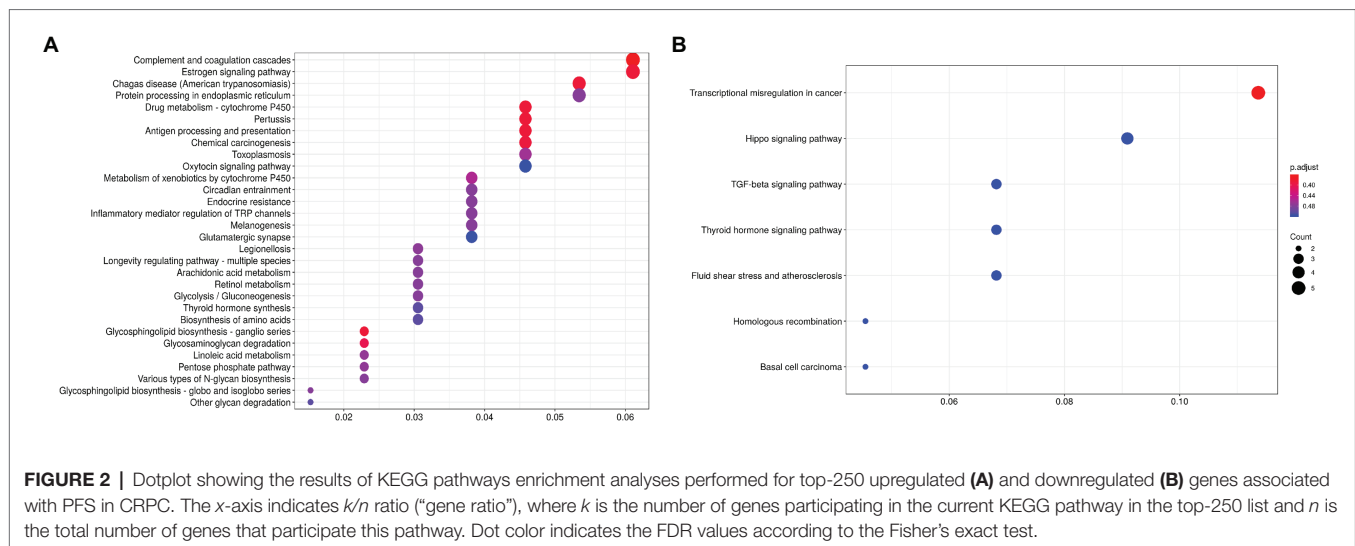


TABLE 2 | KEGG pathways that are significantly enriched with genes associated with PFS in castration-resistant prostate cancer (CRPC).

KEGG ID	Pathway name	Enrichment <i>p</i> -value	Enrichment FDR	Genes
Enriched with upregulated genes				
hsa04610	Complement and coagulation cascades	8,14E-05	1,88E-02	<i>CPB2, SERPIND1, CFI, VTN, KNG1, SERPING1, SERPINA1, C3</i>
hsa00982	Drug metabolism – cytochrome P450	1,45E-04	1,88E-02	<i>ADH6, ADH1B, CYP3A4, CYP2E1, CYP2C8, GSTP1</i>
Enriched with downregulated genes				
hsa05202	Transcriptional misregulation in cancer	3,28E-03	3,65E-01	<i>HIST1H3C, TAF15, ASPSCR1, NCOR1, EYA1</i>
hsa04390	Hippo signaling pathway	1,04E-02	5,17E-01	<i>BMP4, TGFB2, BMP5, NKD1</i>
hsa04350	TGF-beta signaling pathway	1,45E-02	5,17E-01	<i>BMP4, TGFB2, BMP5</i>

The interaction network formed with genes positively correlated with PFS did not pass the PPI enrichment *p*-value threshold.

DISCUSSION

In the present study, we examined RNA-Seq data from the two most advanced stages of PC, LAPC and CRPC, with an aim to reveal the major changes in the transcriptome associated with disease progression within each analyzed category. The pathological manifestation of LAPC is invasion of the prostatic capsule as well as invasion beyond it, whereas CRPC, which is the next progressive stage, is characterized by the presence of distant metastases.

Using KEGG pathway enrichment analysis, we found several pathways, activation of which is possibly associated with PFS period, either positively or negatively. In LAPC, the most prominent associations were the increased expression of the participants of “TNF signaling pathway” and “Transcriptional misregulation in cancer.” Transcriptional misregulation involves an extensive network of processes that play dual roles in cancers, and probably in LAPC.

In CRPC, we identified putative activation of “Complement and coagulation cascades” and “Drug metabolism-cytochrome

P-450” KEGG pathways to be associated with increased PFS time. According to various studies, the expression levels of complement system genes vary in different cancers. In PC, upregulated expression of complement system genes in both the classical and alternative pathways is associated with a good prognosis and long-term patient survival (Roumenina et al., 2019). Based on our results, putative activation of the complement pathway increased the relapse time in CRPC. Therefore, in patients with activation of the complement pathway genes, a longer response to therapy is expected, which may lead to the elimination of tumor-induced immunosuppression and increased antitumor immunity. Among the identified pathways enriched with genes negatively associated with PFS in CRPC, we noticed several cancer-associated pathways such as the “Hippo signaling pathway” and “TGF-beta signaling pathway” as well as “Transcriptional misregulation in cancer.”

Next, we performed PPI network and GO enrichment analysis using the STRINGdb. We examined the interaction networks for sets of genes that were positively and negatively correlated with PFS. In both datasets, statistically significant results were obtained only for genes with negative correlations. In LAPC, an extensive network of protein-protein interactions was identified for downregulated genes, which suggests close biological relationships among the genes and proteins under study.

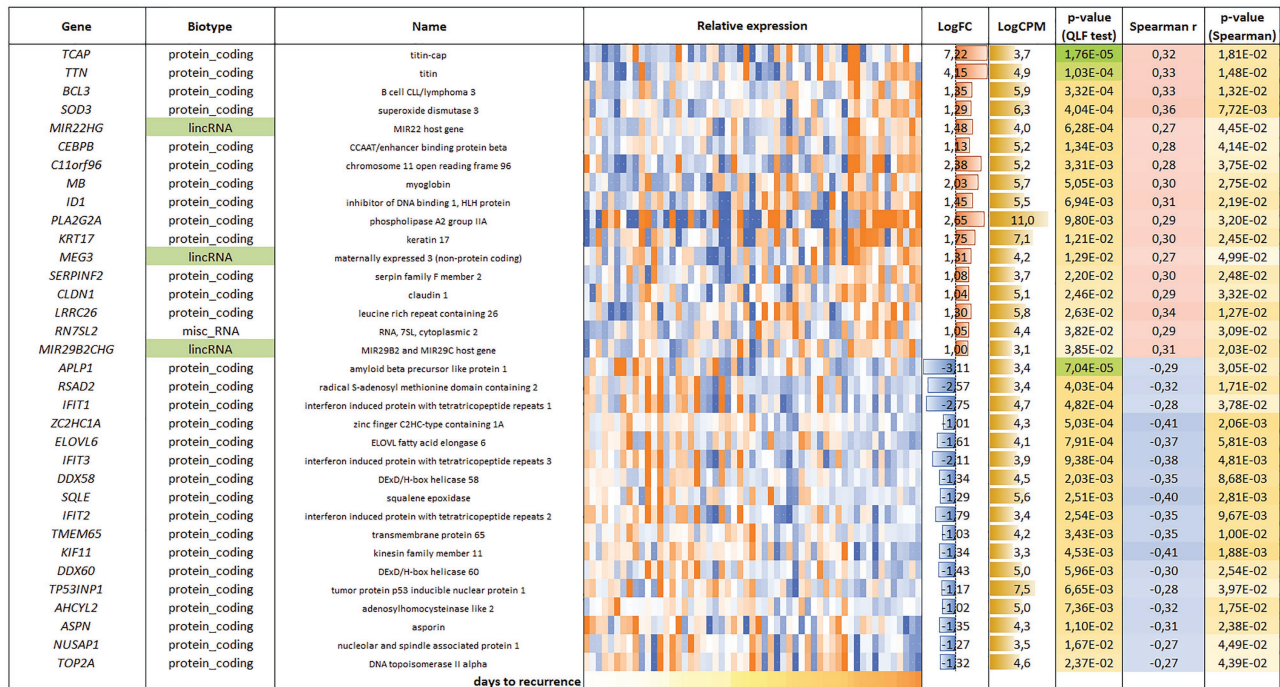


FIGURE 3 | List and heatmap demonstrating log relative expression level of genes with significant Spearman's rank correlation coefficient relative to PFS for LAPC. Cell colors (blue-white-orange gradient) correspond to the binary logarithm of the ratio of the expression level in a current sample to the average level across all the samples (per each gene). Blue – expression level is below the average, orange – above the average. LogFC – binary logarithm of the ratio of GLM-approximated expression values between samples with maximal and minimal PFS; LogCPM – binary logarithm of read counts per million (CPM); *p*-value (QLF test) – *p*-value according to quasi-likelihood F-test (edgeR); Spearman *r* – Spearman's rank correlation coefficient between gene expression level and PFS; *p*-value (Spearman) – *p*-value according to Spearman's rank correlation test.

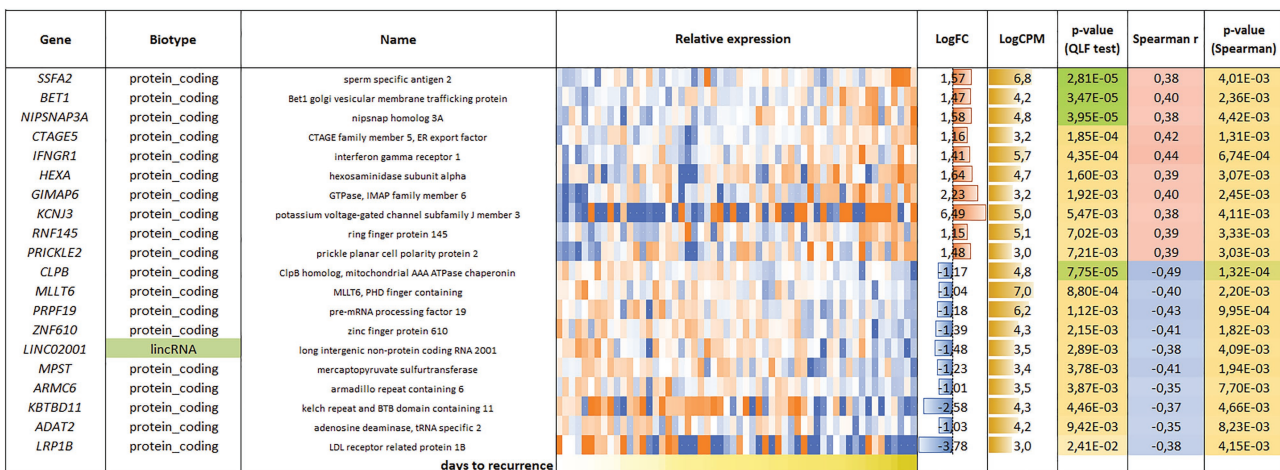


FIGURE 4 | List and heatmap demonstrating log relative expression level of genes with significant Spearman's rank correlation coefficient relative to PFS for CRPC. For comments, see the Figure 3 legend.

We observed significant enrichment with participants of several biological processes such as “Cell cycle,” “Organelle organization,” and “Response to stress.” In CRPC, we observed

strong enrichment with genes involved “Cell cycle,” “Regulation of gene expression,” and “Metabolic process.” This suggests that the activation of the cell cycle and an increase in the

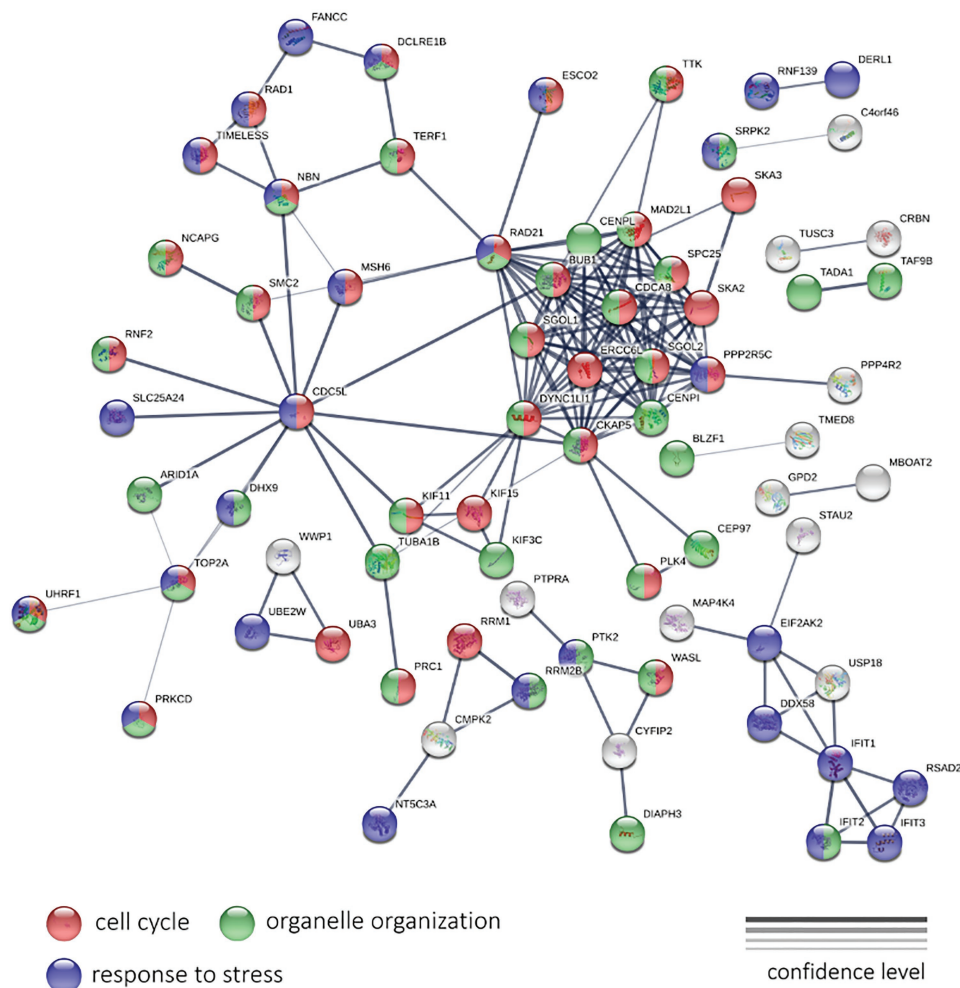


FIGURE 5 | The protein-protein interaction network of 190 genes with a significant negative expression correlation with PFS in LAPC. Circle colors indicate proteins participating in the cell cycle regulation (red), organelle organization (green), and response to stress (dark blue). If a gene does not belong to these categories, it is marked with gray. The thickness of lines indicates the reliability of evidence of the interaction between two proteins (direct or indirect experimental confirmation, predictions from homologues, etc.).

expression of cell cycle-related genes may be behind the formation of more aggressive tumor phenotype and shortened PFS period, for both LAPC and CRPC. This is to be expected. Needless to say once again, the cell cycle, in particular cellular mitosis, and cancer are closely related, as cancer cells undergo abnormal, uncontrolled mitosis, which supports tumor growth and metastasis, two processes that are integral to disease progression (Weaver and Cleveland, 2005).

In our study, we also focused on identifying potential prognostic markers among the genes that have the strongest correlations with PFS. In LAPC, *KIF11* (Spearman's rank correlation coefficient, $r_s = 0.41$), *ZC2HC1A* ($r_s = -0.41$), and *SQLE* ($r_s = -0.40$) had the greatest negative correlations with PFS, while *SOD3* ($r_s = 0.36$) and *LRRC26* ($r_s = 0.34$) had the highest positive correlations. We reviewed the literature regarding the role of these genes in cancer, especially PC.

The kinesin family member 11 (*KIF11*) gene encodes mitotic kinesin, which plays a central role in mitosis (Wojcik et al., 2013). In PC, assessment of tumor cell differentiation, which is recorded as the Gleason score, is an important prognostic parameter. A value of 8 or higher corresponds to a poorly differentiated tumor and is associated with an unfavorable prognosis (Gordetsky and Epstein, 2016). A value of 8 or higher corresponds to a poorly differentiated tumor and is associated with an unfavorable prognosis. It was shown that *KIF11* gene expression was higher in PC tumor samples with a Gleason score of 8 (poorly differentiated tumors) than in tumor samples with a Gleason score of 7 (moderately differentiated tumors; Piao et al., 2017). In our study, we found that *KIF11* gene expression was negatively correlated with an increase in the number of days before recurrence, which confirms the relationship of this gene with unfavorable prognosis in PC.

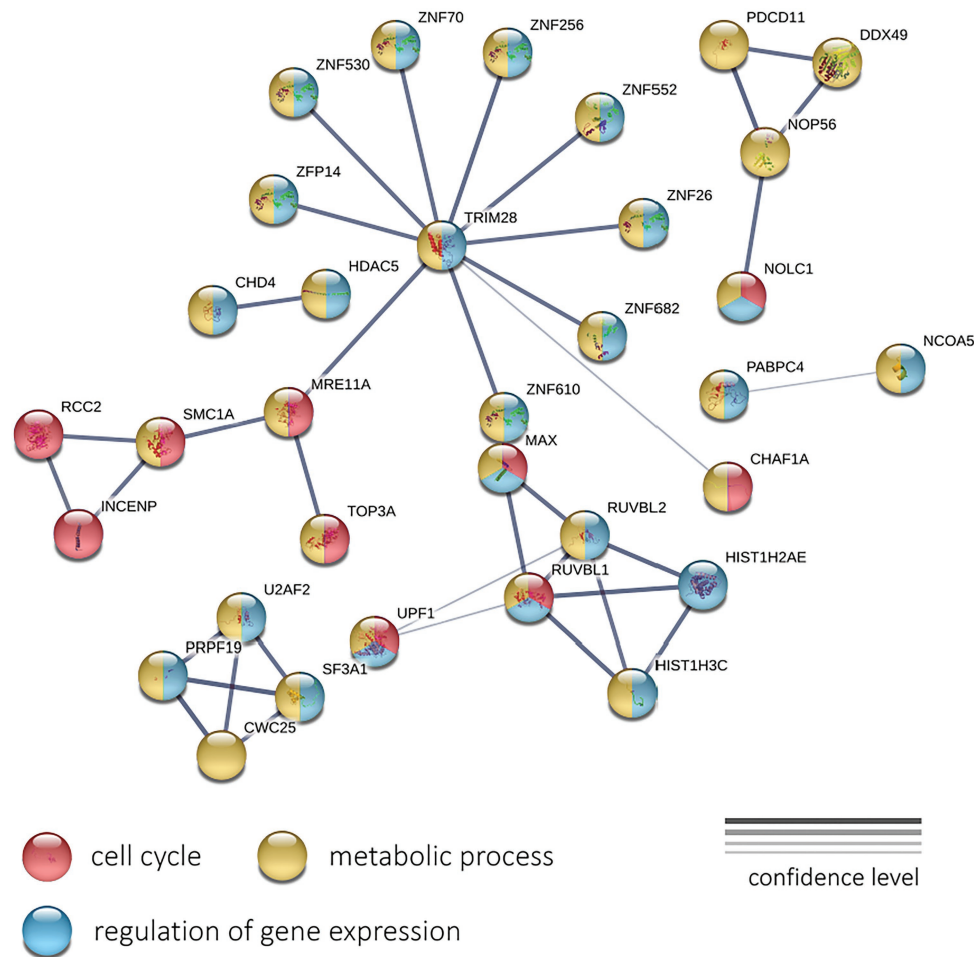


FIGURE 6 | The protein-protein interaction network of 196 genes with a significant negative expression correlation with PFS in CRPC. Circle colors indicate proteins participating in the cell cycle regulation (red), metabolic process (yellow), and regulation of gene expression (light blue). The thickness of lines indicates the reliability of evidence of the interaction between two proteins (direct or indirect experimental confirmation, predictions from homologues, etc.).

There are few data regarding zinc finger C2HC-type containing 1A (*ZC2HC1A*). However, Zhu et al. (2019) showed that overexpression of the *ZC2HC1A* gene was associated with unfavorable prognosis in hepatocellular carcinoma. In our study, *ZC2HC1A* expression was increased in PC samples with early onset recurrence.

The *SQLE* gene encodes squalene epoxidase, which is involved in cholesterol synthesis. Cholesterol is an essential component of cell membranes and a precursor for the synthesis of androgens (Pelton et al., 2012). According to previous data, the *SQLE* gene is expressed in aggressive PC, and its expression is correlated with the Gleason score. It has been suggested that the progression of PC depends on the *de novo* synthesis of cholesterol catalyzed by *SQLE* (Stopsack et al., 2017). In our study, we observed higher *SQLE* gene expression in samples with early onset relapse.

Prostate cancer can also progress due to oxidative stress, which produces reactive oxygen species (ROS; Shiota et al., 2011). Cells are protected against ROS by antioxidant enzymes, such

as superoxide dismutase (SOD), which functions as a first line antioxidant enzyme. The *SOD3* gene encodes an extracellular isoform of the enzyme, and data suggest that *SOD3* acts as a tumor suppressor in PC (Faraci and Didion, 2004). Thus, the effect of *SOD3* gene expression on cell proliferation, migration, and invasion of PC-3 cells was assessed, and the results showed that overexpression of *SOD3* inhibits these processes (Kim et al., 2014). Our data also demonstrate that *SOD3* gene expression was increased relative to time to relapse and was lower in tumor samples from patients with an unfavorable prognosis.

Studies on the function of *LRRC26* gene in LNCaP cell culture have shown that its overexpression leads to suppression of the NF- κ B pathway, which is involved in cancer progression and metastasis (Liu et al., 2012). Thus, decreased expression of the *LRRC26* gene is associated with an unfavorable prognosis in PC, which is consistent with our results.

Correlation analysis of gene expression with PFS in CRPC revealed that the following genes had the strongest Spearman

correlation coefficients: *CLPB* ($r_s = -0.49$), *PRPF19* ($r_s = -0.43$), *MPST* ($r_s = 0.41$), *IFNGR1* ($r_s = 0.44$), *CTAGE5* ($r_s = 0.42$), *GIMAP6* ($r_s = 0.40$), and *BET1* ($r_s = 0.40$).

The caseinolytic mitochondrial matrix peptidase chaperone subunit B (*CLPB*) is an ATPase associated with a variety of cellular processes. Currently, there are no data on the role of this gene in cancer. Our study showed that an increase in expression was associated with an unfavorable prognosis, as evidenced by the strong correlation and high statistical significance.

The pre-mRNA processing factor 19 (*PRPF19*) gene encodes the hPrp19 protein, which is involved in many physiological processes, such as the ubiquitin-proteasome system, DNA damage response, proliferation, and apoptosis (Yin et al., 2012). The hPrp19 protein has been reported to play a potential pro-oncogenic role due to its proliferation-promoting activity. hPrp19 is also required for the expression of p21, which has an intense cell cycle arrest-promoting effect (Chen et al., 2011). The mechanism of hPrp19 in cancer is still unclear; however, its involvement in DNA repair is expected to be correlated with tumor progression. According to our results, in CRPC, increased expression of the *PRPF19* gene is associated with early relapse.

The mercaptopyruvate sulfurtransferase (*MPST*) gene encodes an enzyme involved in the catalysis of endogenous hydrogen sulfide from L-cysteine. Various studies have shown that endogenous hydrogen sulfide can regulate the occurrence and development of tumors and can participate in cancer progression by stimulating angiogenesis and cell growth in colon and ovarian cancer (Bhattacharyya et al., 2013; Szabo et al., 2013; Hellmich and Szabo, 2015). In CRPC, we observed an increase in *MPST* expression at relapse, which suggests the involvement of this gene in PC progression.

Interferon-gamma receptor 1 (*IFNGR1*) encodes a subunit of the IFN- γ receptor that acts in IFN- γ pathways and regulates the immune response. Reduced expression of this gene was observed in MYC-dependent metastatic PC. Experiments that activate *IFNGR1* gene expression demonstrated strong activation of tumor-suppression signaling and sustained apoptosis (Wee et al., 2014). We observed similar results in CRPC, which showed decreased expression of the *IFNGR1* gene in tumor samples with early relapse.

The CTAGE family member 5 (*CTAGE5*) gene product (an ER export factor) is involved in the transport of collagen VII in the endoplasmic reticulum. In PC, the *CTAGE5* gene is involved in tumor-specific splicing (Ren et al., 2012). Our study showed a decrease in the expression of *CTAGE5* in samples with early relapse.

The GTPase, IMAP family member 6 (*GIMAP6*) gene encodes a protein belonging to the GIMAP family of proteins, which are mainly expressed in immune system cells and contribute to the development of thymocytes, apoptosis of peripheral lymphocytes, and T-helper differentiation (Filen et al., 2009). Dysregulation of *GIMAP6* expression is observed in non-small cell lung cancer, where the *GIMAP6* gene expression is decreased in tumor samples (Shiao et al., 2008). Decreased expression of the *GIMAP6* gene has also been

noted in hepatocellular carcinoma (Huang et al., 2016). In CRPC, we also observed that decreased expression of *GIMAP6* was associated with the onset of disease relapse.

Bet1 Golgi vesicular membrane trafficking protein (*BET1*) is a membrane protein associated with the Golgi complex that is involved in vesicular transport. There are few published studies on the function of this gene in cancer. There is a report that *BET1* was identified as part of a gene expression signature associated with a favorable prognosis in glioblastoma (Cai et al., 2020). We showed that decreased *BET1* expression was associated with early relapse in CRPC. In addition to the identified protein-coding genes, in our correlation analysis, we also found long non-coding RNAs (lncRNAs) with altered expression patterns. lncRNAs are involved in the regulation of various biological processes, including many cancer-associated pathways (Spizzo et al., 2012; Flynn and Chang, 2014; Quinn and Chang, 2016). Aberrant lncRNAs expression has been observed in a variety of cancers, and many studies have shown a link between these molecules and cancer development and progression. In LAPC, we found changed expression of the following lncRNAs that were associated with PFS: *MIR22HG* ($r_s = 0.27$), *MEG3* ($r_s = 0.27$), and *MIR29B2CHG* ($r_s = 0.31$).

Regarding the role of MIR22 host gene (*MIR22HG*) in PC, decreased *MIR22HG* expression has been shown to be significantly associated with a higher Gleason score and shorter PFS time, highlighting its prognostic potential (Shen et al., 2019). In our analysis, we also observed an association between decreased *MIR22HG* expression and early onset of disease recurrence.

Studies on maternally expressed 3 (*MEG3*) have shown that this lncRNA inhibits the proliferation and metastasis of gastric cancer through p53 signaling (Wei and Wang, 2017). In PC, *MEG3* is characterized by a downregulated expression, and increased expression has an inhibitory effect on tumor growth (Wu et al., 2019). Therefore, our results are consistent with the published data.

There are currently no data on the role of MIR29B2 and MIR29C host gene (*MIR29B2CHG*) in cancer. Therefore, we report here, for the first time, the involvement of *MIR29B2CHG* in the progression of LAPC.

In CRPC, we identified an association between an increase in the expression of *LINC02001* (small nucleolar RNA host gene 30 [*SNHG30*]; $r_s = -0.38$) and early onset of disease relapse. No studies on the involvement of *LINC02001* in cancer have been reported.

Summing up, in this study, genes and pathways associated with PFS in advanced-stage PC (LAPC and CRPC) were identified. Our results are consistent with the previous studies that reported the participation of *KIF11*, *SQLE*, *SOD3*, *LRRC26*, *IFNGR1*, *MIR22HG*, and *MEG3* in the carcinogenesis and progression of PC. The possible association with the progression of PC was first shown for genes *ZC2HC1A*, *CLPB*, *PRPF19*, *MPST*, *GIMAP6*, *BET1*, *MIR29B2CHG*, and *LINC02001*. All listed genes showed strong correlations with PFS and thus could be considered as potential prognostic markers.

DATA AVAILABILITY STATEMENT

The study included RNA-Seq data for LAPC samples from the PRAD project ($n = 55$) and CRPC samples from the WCDT-MCRPC project ($n = 84$), which were donated to the TCGA consortium.

AUTHOR CONTRIBUTIONS

EAP, AVK, KMN, ADK, BYA and GTS conceived and designed the work. GSK, AAK, and VSP carried out bioinformatics analysis. MVS, MSF, and DYT carried out statistical analysis. EAP, AVS, and GSK wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the Russian Science Foundation (grant 18-75-10127) and RFBR (grant 17-29-06085).

REFERENCES

- Bhattacharyya, S., Saha, S., Giri, K., Lanza, I. R., Nair, K. S., Jennings, N. B., et al. (2013). Cystathionine beta-synthase (CBS) contributes to advanced ovarian cancer progression and drug resistance. *PLoS One* 8:e79167. doi: 10.1371/journal.pone.0079167
- Cai, H. Q., Liu, A. S., Zhang, M. J., Liu, H. J., Meng, X. L., Qian, H. P., et al. (2020). Identifying predictive gene expression and signature related to temozolomide sensitivity of glioblastomas. *Front. Oncol.* 10:669. doi: 10.3389/fonc.2020.00669
- Cao, Z. X., Xiao, G. A., Zhang, W., Ji, J., Ye, C., Liu, D., et al. (2019). Comprehensive investigation of alternative splicing and development of a prognostic risk score for prostate cancer based on six-gene signatures. *J. Cancer* 10, 5585–5596. doi: 10.7150/jca.31725
- Chen, Y., Zhang, L., and Jones, K. A. (2011). SKIP counteracts p53-mediated apoptosis via selective regulation of p21Cip1 mRNA splicing. *Genes Dev.* 25, 701–716. doi: 10.1101/gad.2002611
- Faraci, F. M., and Didion, S. P. (2004). Vascular protection: superoxide dismutase isoforms in the vessel wall. *Arterioscler. Thromb. Vasc. Biol.* 24, 1367–1373. doi: 10.1161/01.ATV.0000133604.20182.cf
- Filen, J. J., Filen, S., Moulder, R., Tuomela, S., Ahlfors, H., West, A., et al. (2009). Quantitative proteomics reveals GIMAP family proteins 1 and 4 to be differentially regulated during human T helper cell differentiation. *Mol. Cell. Proteomics* 8, 32–44. doi: 10.1074/mcp.M800139-MCP200
- Flynn, R. A., and Chang, H. Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell* 14, 752–761. doi: 10.1016/j.stem.2014.05.014
- Forsythe, A., Chandiwana, D., Barth, J., Thabane, M., Baek, J., and Tremblay, G. (2018). Progression-free survival/time to progression as a potential surrogate for overall survival in HR+, HER2– metastatic breast cancer. *Breast Cancer* 10, 69–78. doi: 10.2147/BCTT.S162841
- Gordetsky, J., and Epstein, J. (2016). Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn. Pathol.* 11:25. doi: 10.1186/s13000-016-0478-2
- Hellmich, M. R., and Szabo, C. (2015). Hydrogen sulfide and cancer. *Handb. Exp. Pharmacol.* 230, 233–241. doi: 10.1007/978-3-319-18144-8_12
- Huang, Z., Zhang, W., Gao, C., Ji, B., Chi, X., Zheng, W., et al. (2016). Dysregulation of GTPase IMAF family members in hepatocellular cancer. *Mol. Med. Rep.* 14, 4119–4123. doi: 10.3892/mmr.2016.5764
- Keyes, M., Crook, J., Morton, G., Vigneault, E., Usmani, N., and Morris, W. J. (2013). Treatment options for localized prostate cancer. *Can. Fam. Physician* 59, 1269–1274.

ACKNOWLEDGMENTS

We thank the National Medical Research Center for Obstetrics, Gynecology and Perinatology named after Academician V.I. Kulakov and the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, EIMB RAS for providing the computing power and techniques for the data analysis. This work was performed using the equipment of EIMB RAS “Genome” center (http://www.eimb.ru/ru1/ckp/ccu_genome_c.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.613162/full#supplementary-material>

SUPPLEMENTARY TABLE S1 | List statistically significant genes according to quasi-likelihood F-test for the LAPC.

SUPPLEMENTARY TABLE S2 | List statistically significant genes according to quasi-likelihood F-test for the CRPC.

- Kim, J., Mizokami, A., Shin, M., Izumi, K., Konaka, H., Kadono, Y., et al. (2014). SOD3 acts as a tumor suppressor in PC-3 prostate cancer cells via hydrogen peroxide accumulation. *Anticancer Res.* 34, 2821–2831.
- Larkin, S. E., Holmes, S., Cree, I. A., Walker, T., Basketter, V., Bickers, B., et al. (2012). Identification of markers of prostate cancer progression using candidate gene expression. *Br. J. Cancer* 106, 157–165. doi: 10.1038/bjc.2011.490
- Liu, X. F., Xiang, L., Zhang, Y., Becker, K. G., Bera, T. K., and Pastan, I. (2012). CAPC negatively regulates NF-kappaB activation and suppresses tumor growth and metastasis. *Oncogene* 31, 1673–1682. doi: 10.1038/onc.2011.355
- Pelton, K., Freeman, M. R., and Solomon, K. R. (2012). Cholesterol and prostate cancer. *Curr. Opin. Pharmacol.* 12, 751–759. doi: 10.1016/j.coph.2012.07.006
- Piao, X. -M., Byun, Y. J., Jeong, P., Ha, Y. -S., Yoo, E. S., Yun, S. J., et al. (2017). Kinesin family member 11 mRNA expression predicts prostate cancer aggressiveness. *Clin. Genitourin. Cancer* 15, 450–454. doi: 10.1016/j.clgc.2016.10.005
- Pudova, E. A., Lukyanova, E. N., Nyushko, K. M., Mikhaylenko, D. S., Zaretsky, A. R., Snezhkina, A. V., et al. (2019). Differentially expressed genes associated with prognosis in locally advanced lymph node-negative prostate cancer. *Front. Genet.* 10:730. doi: 10.3389/fgene.2019.00730
- Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet.* 17, 47–62. doi: 10.1038/nrg.2015.10
- Rawla, P. (2019). Epidemiology of prostate cancer. *World J. Oncol.* 10, 63–89. doi: 10.14740/wjon1191
- Ren, S., Peng, Z., Mao, J. H., Yu, Y., Yin, C., Gao, X., et al. (2012). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Res.* 22, 806–821. doi: 10.1038/cr.2012.30
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Roumenina, L. T., Daugan, M. V., Petitprez, F., Sautes-Fridman, C., and Fridman, W. H. (2019). Context-dependent roles of complement in cancer. *Nat. Rev. Cancer* 19, 698–715. doi: 10.1038/s41568-019-0210-0
- Shen, H., Weng, X. D., Yang, D., Wang, L., and Liu, X. H. (2019). Long noncoding RNA MIR22HG is down-regulated in prostate cancer. *Math. Biosci. Eng.* 17, 1776–1786. doi: 10.3934/mbe.2020093
- Shiao, Y. M., Chang, Y. H., Liu, Y. M., Li, J. C., Su, J. S., Liu, K. J., et al. (2008). Dysregulation of GIMAP genes in non-small cell lung cancer. *Lung Cancer* 62, 287–294. doi: 10.1016/j.lungcan.2008.03.021

- Shiota, M., Yokomizo, A., and Naito, S. (2011). Oxidative stress and androgen receptor signaling in the development and progression of castration-resistant prostate cancer. *Free Radic. Biol. Med.* 51, 1320–1328. doi: 10.1016/j.freeradbiomed.2011.07.011
- Song, Z. Y., Chao, F., Zhuo, Z., Ma, Z., Li, W., and Chen, G. (2019). Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis. *Aging* 11, 4736–4756. doi: 10.18632/aging.102087
- Spizzo, R., Almeida, M. I., Colombatti, A., and Calin, G. A. (2012). Long non-coding RNAs and cancer: a new frontier of translational research? *Oncogene* 31, 4577–4587. doi: 10.1038/ncr.2011.621
- Stopsack, K. H., Gerke, T. A., Andrén, O., Andersson, S. -O., Giovannucci, E. L., Mucci, L. A., et al. (2017). Cholesterol uptake and regulation in high-grade and lethal prostate cancers. *Carcinogenesis* 38, 806–811. doi: 10.1093/carcin/bgx058
- Szabo, C., Coletta, C., Chao, C., Modis, K., Szczesny, B., Papapetropoulos, A., et al. (2013). Tumor-derived hydrogen sulfide, produced by cystathionine-beta-synthase, stimulates bioenergetics, cell proliferation, and angiogenesis in colon cancer. *Proc. Natl. Acad. Sci. U. S. A.* 110, 12474–12479. doi: 10.1073/pnas.1306241110
- Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937
- Weaver, B. A., and Cleveland, D. W. (2005). Decoding the links between mitosis, cancer, and chemotherapy: the mitotic checkpoint, adaptation, and cell death. *Cancer Cell* 8, 7–12. doi: 10.1016/j.ccr.2005.06.011
- Wee, Z. N., Li, Z., Lee, P. L., Lee, S. T., Lim, Y. P., and Yu, Q. (2014). EZH2-mediated inactivation of IFN-gamma-JAK-STAT1 signaling is an effective therapeutic target in MYC-driven prostate cancer. *Cell Rep.* 8, 204–216. doi: 10.1016/j.celrep.2014.05.045
- Wei, G. H., and Wang, X. (2017). lncRNA MEG3 inhibit proliferation and metastasis of gastric cancer via p53 signaling pathway. *Eur. Rev. Med. Pharmacol. Sci.* 21, 3850–3856.
- Wojcik, E. J., Buckley, R. S., Richard, J., Liu, L., Huckaba, T. M., and Kim, S. (2013). Kinesin-5: cross-bridging mechanism to targeted clinical therapy. *Gene* 531, 133–149. doi: 10.1016/j.gene.2013.08.004
- Wu, M., Huang, Y., Chen, T., Wang, W., Yang, S., Ye, Z., et al. (2019). lncRNA MEG3 inhibits the progression of prostate cancer by modulating miR-9-5p/QKI-5 axis. *J. Cell. Mol. Med.* 23, 29–38. doi: 10.1111/jcmm.13658
- Yap, T. A., Smith, A. D., Ferraldeschi, R., Al-Lazikani, B., Workman, P., and De Bono, J. S. (2016). Drug discovery in advanced prostate cancer: translating biology into therapy. *Nat. Rev. Drug Discov.* 15, 699–718. doi: 10.1038/nrd.2016.120
- Yin, J., Zhu, J. M., and Shen, X. Z. (2012). New insights into pre-mRNA processing factor 19: a multi-faceted protein in humans. *Biol. Cell.* 104, 695–705. doi: 10.1111/boc.201200011
- Yu, G., Wang, L. G., Han, Y., and He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 16, 284–287. doi: 10.1089/omi.2011.0118
- Zhu, G. Q., Yang, Y., Chen, E. B., Wang, B., Xiao, K., Shi, S. M., et al. (2019). Development and validation of a new tumor-based gene signature predicting prognosis of HBV/HCV-included resected hepatocellular carcinoma patients. *J. Transl. Med.* 17:203. doi: 10.1186/s12967-019-1946-8

Conflict of Interest: The reviewer AT declared a shared affiliation with two of the authors, GS and DT, to the handling editor at time of review.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Pudova, Krasnov, Kobelyatskaya, Savvateeva, Fedorova, Pavlov, Nyushko, Kaprin, Alekseev, Trofimov, Sukhikh, Snezhkina and Kudryavtseva. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Predicting Genome Architecture: Challenges and Solutions

Polina Belokopytova^{1,2} and Veniamin Fishman^{1,2*}

¹ Natural Sciences Department, Novosibirsk State University, Novosibirsk, Russia, ² Institute of Cytology and Genetics Siberian Branch of Russian Academy of Sciences (SB RAS), Novosibirsk, Russia

Genome architecture plays a pivotal role in gene regulation. The use of high-throughput methods for chromatin profiling and 3-D interaction mapping provide rich experimental data sets describing genome organization and dynamics. These data challenge development of new models and algorithms connecting genome architecture with epigenetic marks. In this review, we describe how chromatin architecture could be reconstructed from epigenetic data using biophysical or statistical approaches. We discuss the applicability and limitations of these methods for understanding the mechanisms of chromatin organization. We also highlight the emergence of new predictive approaches for scoring effects of structural variations in human cells.

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Davide Marenduzzo,
University of Edinburgh,
United Kingdom
Vladimir B. Teif,
University of Essex, United Kingdom
Ron Schwesinger,
University of Oxford, United Kingdom

*Correspondence:

Veniamin Fishman
minja-f@ya.ru;
minja-f@yandex.ru

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 14 October 2020

Accepted: 15 December 2020

Published: 22 January 2021

Citation:

Belokopytova P and Fishman V
(2021) Predicting Genome
Architecture: Challenges
and Solutions.
Front. Genet. 11:617202.
doi: 10.3389/fgene.2020.617202

Keywords: Hi-C, modeling, polymer physics, machine learning, predicting approaches

STUDYING GENOME ARCHITECTURE: METHODS AND MECHANISMS

The human genome has a three-dimensional structure, which folds in the nucleus, producing specific chromatin interactions. These chromatin interactions can be experimentally assessed by modern microscopy methods (reviewed in Boettiger and Murphy, 2020) or sequencing approaches, such as genome-wide modifications of chromatin conformation capture (Hi-C) (Lieberman-Aiden et al., 2009; Rao et al., 2014), split-pool recognition of interactions by tag extension (Quinodoz et al., 2018), and genome architecture mapping (Beagrie et al., 2017). These methods are covered by comprehensive reviews (Kempfer and Pombo, 2020) and comparative studies (Fiorillo et al., 2020). Here, we focus mainly on the Hi-C technique and its results because this method was most widely applied in various genomic studies during the last decade, allowing the accumulation of a huge amount of experimental data. Both methodological aspects of the Hi-C technique (Fiorillo et al., 2020) and biological principles revealed by applying this method to study genome architecture (Szabo et al., 2019) are discussed in detail in several recent reviews. We refer readers to **Box 1**, where we briefly discuss the main concepts of this field for the sake of completeness.

WHY MODELING 3-D GENOME FOLDING?

The models and algorithms predicting genome architecture can be used in different ways. First, we can apply modeling to get new insights or test our hypotheses of molecular mechanisms underlying 3-D genome folding. Polymer modeling is used more often for this purpose, but convolutional neural networks, such as, for example, Akita (Fudenberg et al., 2020) and DeepC (Schwesinger et al., 2020), also enable identifying the main chromosome features contributing to genome architecture. Such approaches give remarkable results. During the last few years, we gained a significant amount of data describing the main features of 3-D genome folding and understanding the molecular mechanisms underlying these data, including loop extrusion and

BOX 1 | Start of Box 1 Hi-C Technology Uncovers Principles of Genome Organization

Hi-C includes crosslinking and digestion of chromatin, followed by proximity ligation and sequencing of ligation products (Lieberman-Aiden et al., 2009; Rao et al., 2014). During the proximity ligation step, only those genomic regions that spatially co-localize have a chance to be ligated. Thus, counting ligation products by next-generation sequencing allows deciphering the spatial proximity of loci. Although several single-cell Hi-C methods are published (Flyamer et al., 2017), the technique is most often applied to large cell populations, and ligation event frequency (also referred to as *interaction* or *contact* frequency throughout this review) should be interpreted as the average frequency of loci co-localization among the studied cell population. This snapshot of averaged chromatin contacts in a population, typically represented by a matrix of pairwise interaction frequencies, is known as a Hi-C map.

Using Hi-C and other methods, several important principles of genome architecture were recently discovered. At the largest scales, chromosomes occupy distinct territories, showing only limited intermingling (Tavares-Cadete et al., 2020) and characterized by an exponential decay of contact frequencies with the genomic distance between loci (Lieberman-Aiden et al., 2009). Within the territories, one can distinguish compartments that correspond to different chromatin types (Lieberman-Aiden et al., 2009). Mechanisms underlying compartment formation are actively debated, and there is a growing body of theoretical and experimental pieces of evidence suggesting the essential role of liquid-liquid phase separation in these processes (Kantidze and Razin, 2020; Razin and Gavrillov, 2020; Razin and Ulianov, 2020). At a finer scale, specific loci may preferentially interact with each other, forming topologically associated domains (TADs) (Dixon et al., 2012), stripes (Vian et al., 2018), cliques (Petrovic et al., 2019), and loops (Rao et al., 2014). Although the terminology is not well established in this field (de Wit, 2020), the current mechanisms underlying the formation of these structures fall into two categories.

First is a recently proposed loop extrusion mechanism (Sanborn et al., 2015; Fudenberg et al., 2016). It is considered that ring-shaped cohesin and condensin proteins bind chromatin and form and continuously extend loops in an ATP-dependent manner. Extrusion stops encountering another extrusion complex or, in the case of cohesins, when reaching CTCF protein bound to DNA in a specific orientation. This results in increased interaction frequency between loci bound by cohesin, displayed on Hi-C maps as loops (two-point interactions) (Rao et al., 2014) or stripes (one-to-many-points interactions) (Vian et al., 2018). The chromatin interaction patterns arising from loop extrusion mechanisms could be qualitatively described by the landscape of CTCF binding and also depend on the loading and processivity of cohesin (Fudenberg et al., 2016). Moreover, loop extrusion results in increased proximity of all loci located between convergently oriented CTCF sites, which is captured by the formation of looping domains (Rao et al., 2014).

The second mechanism responsible for the formation of loops and cliques is mediated by the formation of regulatory protein complexes, for example, polycomb complexes (Eagen et al., 2017), and certain transcription factors (Petrovic et al., 2019). This mechanism is at least partially independent of cohesin-mediated extrusion because the subset of loops remains stable upon degradation of the cohesin complex (Rao et al., 2017).

It is important to note that profiles of chromatin interactions captured by the Hi-C experiment are formed by the joint action of different mechanisms. For example, the formation of TADs, which represent self-interacting regions in the genome, is affected both by loop extrusion and compartmentalization processes (Szabo et al., 2019; de Wit, 2020), which is consistent with both convergent CTCF sites and chromatin state transition enrichment at TAD boundaries (Dixon et al., 2012; Rao et al., 2014; Huang et al., 2015).

phase separation, which was largely facilitated by biophysical modeling and statistical analysis of chromatin properties. This field of research is well described in reviews (Imakaev et al., 2015; Lin et al., 2019). However, known mechanisms do not explain all 3-D chromatin features, which limits hypothesis-driven models and further research is required to explain them.

Second, 3-D genome models can be used to predict functional consequences caused by changes in 3-D genome folding. It is shown that alterations of chromatin topology accompanying genomic variations, especially large structural variations, can cause changes of gene expression (Franke et al., 2016; Rodríguez-Carballo et al., 2017; Kraft et al., 2019). One can find examples of such gene expression changes and their underlying mechanisms in the last part of this review. In these cases, modeling of 3-D genome architecture is essential for accurate prediction of the consequences of the genomic mutations.

Last, one can use modeling for predicting the 3-D genome architecture of new data. It is possible to predict chromatin interactions for different cell types lacking experimental Hi-C data (Belokopytova et al., 2020). Machine learning methods often gain applicability in this way.

WHICH 3-D GENOME STRUCTURES CAN BE PREDICTED, AND WHY THEY ARE RELEVANT?

Chromosome-capturing methods, such as Hi-C, allow deciphering the main features of chromatin folding. Since the first Hi-C experiments, chromatin structures as compartments,

TADs, and loops were revealed (see **Box 1** for details of mechanisms underlying these structures). In the following, we describe the main Hi-C map features and algorithms used to predict them. Also, it may be helpful for readers new to the field to use the table of algorithms (**Table 1**) containing algorithms for predicting different 3-D genome features.

Promoter-Enhancer Interactions

Interactions between promoters and enhancers are essential for expression regulation. Pioneering attempts to find such regulatory connections rely on either the correlation of epigenetic marks of promoters and enhancers across different cell types or evolutionary conservation of promoter-enhancer proximity in the linear DNA molecule (Spicuglia and Vanhille, 2012; Andersson and Sandelin, 2020). With the advent of genome-wide 3C-methods, we gain the ability to measure spatial proximity between genomic segments. The question about the exact role of spatial contacts between regulatory elements in the control of gene expression is still under active debate; however, much research defines “interacting” enhancers and promoters as pairs of loci belonging to the anchors of one Hi-C loop. Although we argue that using this loop-based definition of interacting promoters and enhancers might be confusing (see **Box 1** and limitations section below for additional discussion), several algorithms are designed to predict enhancer-promoter pairs located within the anchors of one loop (Whalen et al., 2016).

Loops

Instead of predicting whether promoters and enhancers overlap loop anchors, some algorithms, such as Lollipop (Kai et al., 2018),

TABLE 1 | Tools for modeling and predicting chromatin interactions.

Tool name	Input features	Target features	Method/algorithm
See review by Xu et al. (2018)	Histone marks, TFs binding, DHS	Promoter–enhancer interactions	See review by Xu et al. (2018)
MacPherson et al. (2018) model	HP1, H3K9me3	Compartments	Polymer modeling
MichroM + MEGABASE (Di Pierro et al.)	Histone marks, TFs binding	Compartments	NN classifier + polymer modeling
Huang et al. (2015) model	Histone marks	TADs	BART
3Disease Browser (Li et al., 2016)	Enhancers and TAD boundaries	Rearranged TADs	Linear model
Lollipop (Kai et al., 2018)	Chip-seq data, CTCF directionality	Loops	ML ensemble classifier (random forest)
3DEpiloop (Al Bkhetan and Plewczynski, 2018)	Histone marks, TFs binding	Loops	ML ensemble classifier (random forest)
CTCF-MP (Zhang et al., 2018)	CTCF binding, DHS, nucleotide sequence	Loops	ML ensemble classifier/NN (Boosted trees/word2vec)
EpiTensor (Zhu et al., 2016)	Histone marks, TFs binding	Loops	Tensor modeling + PCA
DeepMILO (Trieu et al., 2020)	Sequence of loop anchors	Rearranged loops	CNN and RNN
3D-GNOME (Sadowski et al., 2019)	CTCF ChIA-PET	Rearranged loops	linear models
3DPredictor (Belokopytova et al., 2020)	CTCF, RNA-seq	Whole hi-c map	ML ensemble regression (gradient boosting)
Hi-C Reg (Zhang et al., 2019)	Histone marks, TFs binding, DHS	Whole hi-c map	ML ensemble regression (random forest)
Akita (Fudenberg et al., 2020)	Sequence	Whole hi-c map	CNN
DeepC (Schwessinger et al., 2020)	Sequence	Whole hi-c map	CNN
Yifeng Qi and Bin Zhang model (Qi and Zhang, 2019)	CTCF binding, Chromatin states	Whole hi-c map	Polymer modeling
HiP-HoP (Buckle et al., 2018)	CTCF and cohesin binding, Histone marks or DHS	Whole hi-c map	Polymer modeling
Rowley et al. (2017) model	GRO-seq + CTCF binding	Whole hi-c map	Explicit algebraic model
PRISMR (Bianco et al., 2018)	Wild-type Hi-C data	Whole hi-c map in mutated cells	Polymer modeling

DHS, DNase I hypersensitivity sites; TFs, transcription factors; TADs, topologically associated domains; ML, machine learning; NN, neural network; CNN, convolutional neural network; RNN, recurrent neural network; BART, Bayesian additive regression trees; PCA, principle component analysis.

3DEpiloop (Al Bkhetan and Plewczynski, 2018), and EpiTensor (Zhu et al., 2016), are designed to directly infer all loop positions using epigenetic data. In mammals, most of the looping interactions are formed due to the cohesin-mediated loop extrusion process (see **Box 1** for details). Thus, some algorithms, such as CTCF-MP (Zhang et al., 2018) or Lollipop (Kai et al., 2018), are focused exclusively on the prediction of CTCF-mediated interactions or separately access quality of prediction for CTCF-mediated and all other loops as in the DeepMILO algorithm (Trieu et al., 2020).

TADs

TADs have the shape of triangles on Hi-C maps, which indicates an increase of chromatin interaction frequency within TADs and insulation at TAD borders. These structures are largely dependent on the extrusion process and also influenced by other mechanisms (see references provided in **Box 1** for discussion of the TAD definition and current views on mechanisms explaining TAD formation). TADs are also relevant for promoter–enhancer interactions as the majority of the functional interactions occur within the same TAD. It is known that TAD boundaries are enriched by CTCF binding sites (usually in convergent orientation) and different epigenetic marks (Dixon et al., 2012). Based on these observations, Huang et al. (2015) use ChIP-seq data for different proteins in a computational model predicting TAD boundaries and chromatin interaction hubs.

Compartments

Chromatin compartments are the main features of distant contacts revealed by chromosome conformation capture. Hi-C maps show that interactions occur more often within each compartment rather than across compartments (Lieberman-Aiden et al., 2009). The presence of compartments results in a checkerboard-like (or “plaid-like”) pattern of contacts on Hi-C maps. It is shown that compartments reflect the clustering of different types of chromatin (see **Box 1** for details). Seminal work proposed binary division of the genome into eu- and heterochromatin, which correspond to A- and B-compartments. Subsequent research extends this view, suggesting that multiple chromatin states exist, each described by a unique profile of spatial interactions (Rao et al., 2014). In accord with this, several models are proposed, allowing the prediction of compartmental interactions based on epigenetic data (Di Pierro et al., 2017; MacPherson et al., 2018). Most of these algorithms utilize physical modeling to infer spatial chromatin interactions. Machine learning methods are often used as a part of the algorithm to attribute genomic loci to a certain compartment based on its epigenetic signatures.

Hi-C Maps

Predictions of all aforementioned features require similar epigenetic information. Thus, it should be possible to develop an algorithm predicting all topological structures simultaneously. Because it is widely assumed that biologically

relevant interactions do not occur at a distance above several megabases, most of the algorithms limit their prediction to these distances, which reduces computational time and resources. For instance, machine learning algorithms, such as 3Dpredictor (Belokopytova et al., 2020), HiC-Reg (Zhang et al., 2019), Akita (Fudenberg et al., 2020), and DeepC (Schwessinger et al., 2020), predict all interactions within an $\sim 1\text{--}3$ Mb window. In addition, some polymer modeling approaches, such as Hip-Hop (Buckle et al., 2018) and PRISMR (Bianco et al., 2018), could be used to predict the whole Hi-C heat map.

From Contact Frequencies to 3-D Models

Hi-C and other 3C-based methods provide a snapshot of pairwise interactions between loci. Although we call this “3-D” information, it cannot be trivially transformed into 3-D structures. An approach known as restraint-based (RB) modeling interprets the 3C-based data as a set of spatial restraints to build a 3-D model of the chromatin fiber by satisfying the input restraints. The chromatin fiber is represented as a polymer of consecutive monomers, and several computational optimization strategies can be employed to find 3-D models of chromatin (Dekker et al., 2013; Serra et al., 2015). The challenge of predicting 3-D genomic structures from high-resolution chromosome conformation capture data was recently taken by several groups, and we refer the reader to the recent review by Kimberly MacKay and Anthony Kusalik describing problems and solutions in this field (MacKay and Kusalik, 2020) and to the articles collected in the recently published book *Modeling the 3D Conformation of Genomes* (Tiana and Luca, 2019).

HOW DO THE MODELING ALGORITHMS WORK? PROBLEMS AND LIMITATIONS

All models and algorithms that are currently used to infer chromatin contacts from epigenetic data could be divided into two categories. First are the models derived from the physical simulation of chromatin behavior, i.e., polymer modeling. The second includes statistical algorithms searching for interdependencies between genetic and epigenetic properties and patterns of 3-D contacts. Here, we described the principles and limitations of both approaches.

POLYMER MODELING

The physics of chromatin has been the subject of intense research over many decades. Seminal studies by de Gennes and Witten (1980) provide basic rules describing polymer behavior under different conditions. Importantly, these studies show that, when a polymer is large (i.e., its size increases the size of individual monomers significantly), its physical properties do not depend on the monomer's chemical structure. Instead, the behavior of a polymer depends on several physical parameters, such as monomer concentration, solvent quality, and temperature. For different combinations of these parameters, the polymer would

exist in one of the well-described equilibrium states, such as the random coil, the swollen coil, the equilibrium globular state, and others (Fudenberg and Mirny, 2012). Thus, knowing the key parameters and using the laws of polymer physics would allow the description (and prediction) of chromatin behavior within the nucleus. These ideas gave rise to the first physical models of chromatin architecture.

Development and validation of physical models during recent decades are linked to the development of experimental techniques measuring genome architecture (**Figure 1**). The presence of chromosome territories as well as measures of mean distances between defined loci by FISH disagree with basic swollen coil or random coil polymer properties (Hahnfeldt et al., 1993). There were multiple attempts to improve these disagreements, of which the fractal globule (Mirny, 2011) is currently the most accepted. This model, originally proposed by Grosberg et al. (1988) suggests that chromatin exists in a highly unknotted fractal-like non-equilibrium state, and the predictions obtained using this model fit well with the experimentally measured scaling of Hi-C contacts (Lieberman-Aiden et al., 2009).

Although the fractal globule recapitulates the experimentally observed scaling of chromatin contacts better than the equilibrium globule state, it is still far from a complete description of chromatin folding in a real cell. Not to mention all disagreements (see Grosberg, 2016, for a detailed review), the fractal globule represents a pictorial description of the chromatin structures and does not include locus-specific features. Thus, to build a more comprehensive description of chromatin conformation and dynamics in a real cell, active (energy-consuming) locus-specific mechanisms should be introduced into the system.

One such mechanism, which maintains the structure of chromatin, is a loop extrusion process (see **Box 1** for details on this mechanism). This process was recently introduced into physical models of chromatin by Fudenberg et al. (2016) and Sanborn et al. (2015), and later experimentally validated by Ganji et al. (2018), Davidson et al. (2019), and Kim et al. (2019). A recent preprint from Banigan et al. (2020) shows another impressive application of polymer modeling in which it helps to investigate if a one- or two-sided loop extrusion model works in the cell and to identify a class of one-sided extrusion models that can reproduce *in vivo* experiments. The models of loop extrusion show good agreement with the experimental Hi-C data. Importantly, loop extrusion models use epigenetic information about CTCF binding to account for CTCF-mediated extrusion barriers. This allows making the model locus-specific; moreover, modifying CTCF anchors *in silico* results in different chromatin packaging as revealed by the models (Sanborn et al., 2015). Thus, such physical models allow predicting chromatin packaging and its perturbations knowing CTCF-binding sites.

Another class of locus-specific models is designed to study and predict the packaging of different chromatin types. Distinct types of chromatin differentially interact with themselves and surrounding proteins. This can be imagined as a polymer composed of several distinct units or blocks. Such polymers are called block copolymers, and their behavior could be modeled

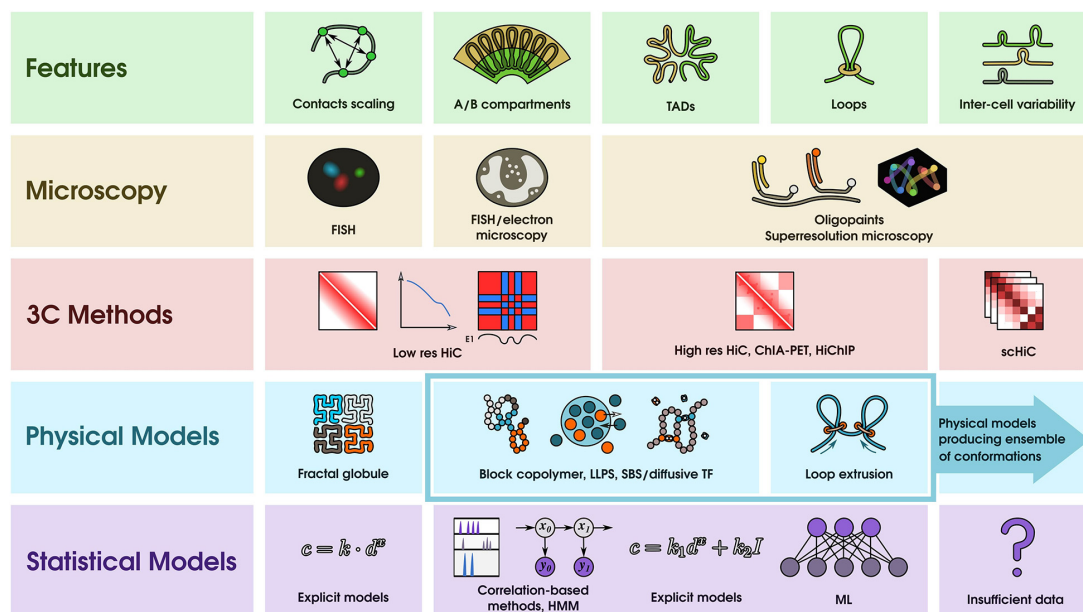


FIGURE 1 | Modeling and predicting the main features of the 3-D-genome organization using physical and statistical approaches. The features row contains a schematic representation of the main features of 3-D-genome organization: scaling of contacts with genomic distance, compartments, TADs, loops, and inter-cell variability of genome architecture (from left to right). The 3C methods row shows that contact scaling and compartments could be found using low-resolution Hi-C data, whereas identification of loops and dissection of TAD structure requires high resolution. Studying inter-cell variability is challenging and could be done using single-cell Hi-C approaches (scHi-C). Microscopy methods shown in the second row include conventional 3D-FISH (fluorescent *in situ* hybridization) to measure spatial distances; electron microscopy, which is helpful to visualize segregation of eu- and heterochromatin; and modern super-resolution microscopy methods, which, in combination with oligopaints, allow dissection of the internal structure of TADs in individual cells (Boettiger and Murphy, 2020; Szabo et al., 2020). Physical description of chromatin interactions (fourth row) includes generic models such as the fractal globule as well as locus-specific models. For the latter, researchers employ block-copolymer models and models with chromatin binders, such as strings and binders switch (SBS) and diffusive transcription factor (TF) models and concepts of liquid–liquid phase separation (LLPS). All these physical models allow studying the dynamics and inter-cell variability of 3-D structures, providing ensembles of possible chromatin conformations (this is schematically shown in the last cell of the physical models row). Statistical methods (the last row) could utilize interconnections between epigenetic data and chromatin organization using different approaches. This includes approaches in which explicitly defined algebraic expressions contain free parameters, which could be fit from the data, hidden Markov models (HMM), and various machine learning (ML) algorithms. TADs, loops, and compartments were predicted using these methods. However, for single-cell data, these approaches are not applicable, mainly due to the large amount of data required for the implementation of these algorithms.

knowing the interaction potential between blocks (Bates and Fredrickson, 1990). Several attempts have been made to apply this logic for modeling chromatin interactions in *Drosophila* and *Human* (Jost et al., 2014; Di Pierro et al., 2016; Ulianov et al., 2016). These models predict that specific preferences of interactions between similar blocks of chromatin result in spatial segregation of distinct chromatin domains in the process of liquid–liquid phase separation (Nuebler et al., 2018).

Block copolymer models rely on the epigenetic information about histone modifications and/or architectural factor binding to assign DNA segments to specific chromatin types. Once developed, these models could be used to predict chromatin architecture if epigenetic data is available. Indeed, several studies show that such prediction recapitulates Hi-C data very well (Di Pierro et al., 2017), especially when accounting for the loop extrusion process (Nuebler et al., 2018; Qi and Zhang, 2019).

To further extend block copolymer models, one should consider the physical nature of interactions between blocks. In a nucleus, these interactions are mediated by specific factors, such as polycomb-group proteins (Plys et al., 2019; Eeftens et al., 2020), BRD-domain containing proteins (Gibson et al., 2019),

HP1 (Larson et al., 2017; Sanulli et al., 2019), mediator and RNA polymerase II (Cho et al., 2018), or interactions between DNA and nuclear lamina proteins (Chiang et al., 2019; Ulianov et al., 2019). The above-described block copolymer models account for these interactions implicitly by setting specific interaction potentials between different block types. Other models explicitly introduce binder proteins that mediate interactions in the system.

There are multiple ligand-binding theories applied to model DNA–protein interactions in chromatin, reviewed, for example, in Teif and Rippe, 2010. Among recent models that aim to explain genome-wide interaction profiles revealed by 3C-based methods, several consider specific chromatin binders, such as HP1 (Teif et al., 2015; MacPherson et al., 2018), lamina proteins (Chiang et al., 2019; Ulianov et al., 2019), or generic active and inactive complexes (Brackley et al., 2016b), whereas others describe binders, such as abstract molecules with defined physical properties but unknown biological nature (Nicodemi and Prisco, 2009; Barbieri et al., 2012; Brackley et al., 2013, 2017; Chiariello et al., 2016). Mechanistically, chromatin clustering may be reproduced by these models either due to the affinity of binders or because of multivalent interactions between binders

and chromatin, which results in bridging-induced attraction (Brackley et al., 2013, 2017; Johnson et al., 2015). In addition to compartmentalization, these mechanisms could explain TAD and loops formation (Brackley et al., 2016b). For more details on these and other physical models, we refer the reader to a recently published extensive review (Brackley et al., 2020) and a collection of articles provided with the book (Tiana and Luca, 2019).

Here, it is pertinent to note that the binder positions are inferred from epigenetic data even in those models that use “abstract” binders. This allows predicting chromatin folding in normal and mutated genomes, knowing epigenetic data with high accuracy (Scialdone et al., 2011; Bianco et al., 2012, 2018; Brackley et al., 2016a,b; Barbieri et al., 2017; Chiariello et al., 2017; Kragesteen et al., 2018). For example, the Hip-Hop model (Buckle et al., 2018) infers binder positions based on H3K27 acetylation data and/or chromatin accessibility, and the authors show that this epigenetic information is sufficient for prediction of chromatin interactions. In the PRISMR model (Bianco et al., 2018), Hi-C data obtained from wild-type cells are used to define the number of binder types and their affinities, and this information can be further used to model chromatin conformation after a deletion or duplication event occurs.

The examples mentioned above show that physical modeling could be a powerful tool for both validation of proposed molecular mechanisms underlying chromatin architecture and predicting spatial interactions based on epigenetic data. In the following, we discuss some limitations that should be addressed to allow a comprehensive description of genome organization by physical modeling.

Limitations of Physical Models

Physical Modeling Is Hypothesis-Driven

As was mentioned above, physical models rely on an explicitly defined set of rules to describe polymer behavior. However, we are still far from a complete understanding of all biophysical processes involved in chromatin organization. Thus, it is clear that none of the currently developed models can accurately explain all details of genome architecture and dynamics.

For example, PRISMR and Hip-Hop models introduce specific binders whose positions and affinity could be inferred from experimental Hi-C or ChIP-seq data. The problem is not only that we do not know the correspondence between the model’s abstract binders and real proteins. The major concern is that these abstract binders might not be given the same physical properties as real proteins. Biochemical dissection of regulatory complexes, such as PRC1 or Mediator, show the complexity of their structural organization and regulation, which is not described by current models. This limits modeling approaches to qualitative predictions of trends rather than quantitative comparison with contact maps.

Inferring Key Physical Parameters Might Be Challenging

There are many biophysical parameters that are currently unknown but essential for modeling. This includes affinity constants and concentrations of chromatin binders, the position of boundaries, and processivity of loop extruders and other

factors. One solution to this problem is extracting the missing parameters from available ChIP-seq data. For example, in the MEGABASE + MiChroM model developed by Di Pierro and colleagues, chromatin states are first inferred from epigenetic data using a machine learning approach and then used in a block copolymer model optimized to fit Hi-C data (Di Pierro et al., 2017). However, in many cases, available ChIP-seq data is only indirectly connected to the affinity and concentration of the key architectural factors, and the dependence between ChIP-seq signals and biophysical properties of chromatin may vary in different cell types. Thus, the model developed using one cell type might not be well transferable to another.

There are also models that fit their parameters directly using Hi-C data. This is, for example, the PRISMR model (Bianco et al., 2018), which defines binder types and positions based on Hi-C maps. The transferability of this model to other cell types or loci without knowing corresponding experimental Hi-C data could be problematic.

There are also several technical parameters of simulation that could influence the results, including the finite volume effect, polymer conformation used for model initialization, equilibration time, sampling size, etc. We refer those readers interested in this subject to a recent review describing potential pitfalls and methods developed to overcome these limitations (Gartner and Jayaraman, 2019).

Physical Modeling Is Computationally Intensive and Often Requires Coarse-Graining

Using a polymer modeling approach is computationally intensive. Technically, the vast majority of the physical models describe chromatin as a string with beads. Ideally, each bead should represent a single nucleosome as histone octamers are monomers of chromatin organization. However, this leads to a huge number of beads required to simulate chromosome-scaled loci. The behavior of beads is typically simulated using LAMMPS software, which is computationally intensive for such a large number of objects. Great computational resources are needed for every modeling attempt, and these are not always accessible. Although it is possible to model only a particular chromosomal region, whole chromosome or whole genome modeling is computationally too expensive.

One solution could be to decrease the resolution and use more coarse-grained models, with which several atoms or molecules are grouped and represented by a single simple object. However, this comes at a cost of the inability to resolve fine patterns of interactions. There are multiple levels of chromatin coarse-graining, starting from atomic resolution and up to hundreds of thousands of base pairs, each suitable for the specific problem of interest (see **Table 1** in the recent review published by Brackley et al., 2020). The choice of coarse-graining should be considered carefully in order to find a balance between the detail of the model and computational cost.

To sum up, physical modeling is essential for validating hypotheses about mechanisms driving chromatin organization. When using epigenetic data to infer properties of chromatin monomers, it is easy to repurpose a physical model from hypothesis validation to prediction of locus-specific chromatin

organization. However, there are several limitations of these predictions, and we next describe another class of approaches based on machine learning techniques that have the potential to overcome some of the aforementioned limitations.

STATISTICAL APPROACH

It is known that different epigenetic marks and transcriptional factors correlate with various regulatory elements, chromatin states, and other genomic features. For example, histone modification H3K9me3 correlates well with constitutive heterochromatin, which correlates with the B compartment (Strom et al., 2017), TAD boundaries are enriched by CTCF protein (Dixon et al., 2012; Rao et al., 2014), and open chromatin regions are enriched by specific histone modification. Thus, one can simply use regression to predict 3-D genome features based on epigenetics data. For example, correlation-based methods are used for the prediction of enhancer–promoter interactions using histone modifications, CAGE, ChIP-seq, and other chromatin features as input (Xu et al., 2020).

Although linear models could explain 3-D organization to some extent, it is clear that certain dependencies between genetic features and chromatin interactions are not linear. The most prominent example of such non-linearity is the scaling of the average chromatin contact frequency with genomic distance, which could be well described as a power law. This dependence, $P(s) \sim s^{-x}$, has only one free parameter x , which could be easily obtained by fitting experimental data. Of course, it is not enough to account for distance dependence to obtain accurate estimations of contact frequencies. One should also describe locus-specific insulation, compartmentalization, and other features of genome organization. This description should be done in the form of algebraic expressions with some free parameters that could be fit from the data. This was utilized recently by Rowley et al. (2017), who proposed an algebraic expression combining linear and exponential terms to predict genomic contacts based on GRO-seq transcription data, CTCF binding, and genomic distance. As a result, Rowley et al. simulate Hi-C maps including main 3-D structures, such as TADs and loops with high accuracy.

However, there might be multiple non-linear dependencies between histone modifications, transcription factor binding, and chromatin interactions, which cannot be defined analytically as an algebraic expression, such as a power law. These dependencies could be found by sophisticated machine learning algorithms, such as logistic regression, gradient boosting, random forest regression, neural networks, and others (Eraslan et al., 2019; **Figure 2**).

Machine learning algorithms operate with a numerical representation of input information (*features*): nucleotide sequence; genomic distance or epigenetic marks; and experimentally measured target feature values, such as contact frequency between loci, positions of loop anchors, etc. The main result of machine learning training is a function that transforms input features into predictions of target values. The similarity between predictions and experimental data is measured using a

user-defined loss function. During a training step, the portion of available data called the training subsample is used to optimize the transforming function so that the loss function is minimal; this is how the algorithm finds interdependencies between features and target values. These interdependencies might represent general biological mechanisms or be subsampling artifacts specific to the training subsample. Moreover, the function transforming the input features into predictions of target values typically has numerous adjustable parameters. This could allow fitting the detail and noise in the training data to the extent that it negatively impacts the performance of the model on held-out data. In this case, the developed algorithm is of no use even if prediction accuracy is high as it cannot generalize over unseen samples. This problem is well known in the machine learning field under the name of “overfitting.” To verify that any increase in accuracy over the training subset is generalizable, an evaluation of the algorithm using a portion of unseen data (*validation subset*) should be done. It is essential that the validation subset does not contain samples presented in the training subset. However, during the design of training and validation subsets, one should note that genomic objects that are not equivalent from a mathematical point of view might share a large amount of biological information. For example, nested chromatin loops might share a large portion of epigenetic information encoded by the window spanning loop anchors although the anchors themselves do not overlap and formally represent different pairs of genomic regions. Such indirect overlapping results in the sharing of information between training and validation data sets, leading to the overestimation of prediction accuracy (Belokopytova et al., 2020). To overcome this problem, one can use different chromosomes for training and validation data sets.

It is considered that machine learning-based algorithms can find complex non-linear patterns when fitting the model. Machine learning is used for binary classifiers for regression-based models, enabling the prediction of structures ranging from two-point interactions to whole Hi-C maps. Several algorithms employing these methods for promoter–enhancer interaction prediction were recently developed, including TargetFinder (Whalen et al., 2016), DeepTACT (Li et al., 2019), 3DPredictor (Belokopytova et al., 2020), and HiC-Reg (Zhang et al., 2019). We refer the reader to the informative review of Xu et al. (2020) describing different algorithms for the prediction of enhancer–promoter interactions. Other spatial chromatin structures, such as loops (Zhu et al., 2016; Al Bkhetan and Plewczynski, 2018; Kai et al., 2018; Zhang et al., 2018; Trieu et al., 2020) and contact probabilities (Zhang et al., 2019; Belokopytova et al., 2020; Fudenberg et al., 2020; Schwessinger et al., 2020) also can be predicted by machine learning-based algorithms (see the section above). Furthermore, a machine learning-based approach enables revealing biological features underlying 3-D genome folding, which improves our understanding of biological mechanisms. For example, extracting matrix positional weights from layers of convolution neural networks helps to find the main features, in particular, sequences giving the main contribution to the prediction and consequently to the 3-D chromatin structure. Another example is the analysis of feature importance in a

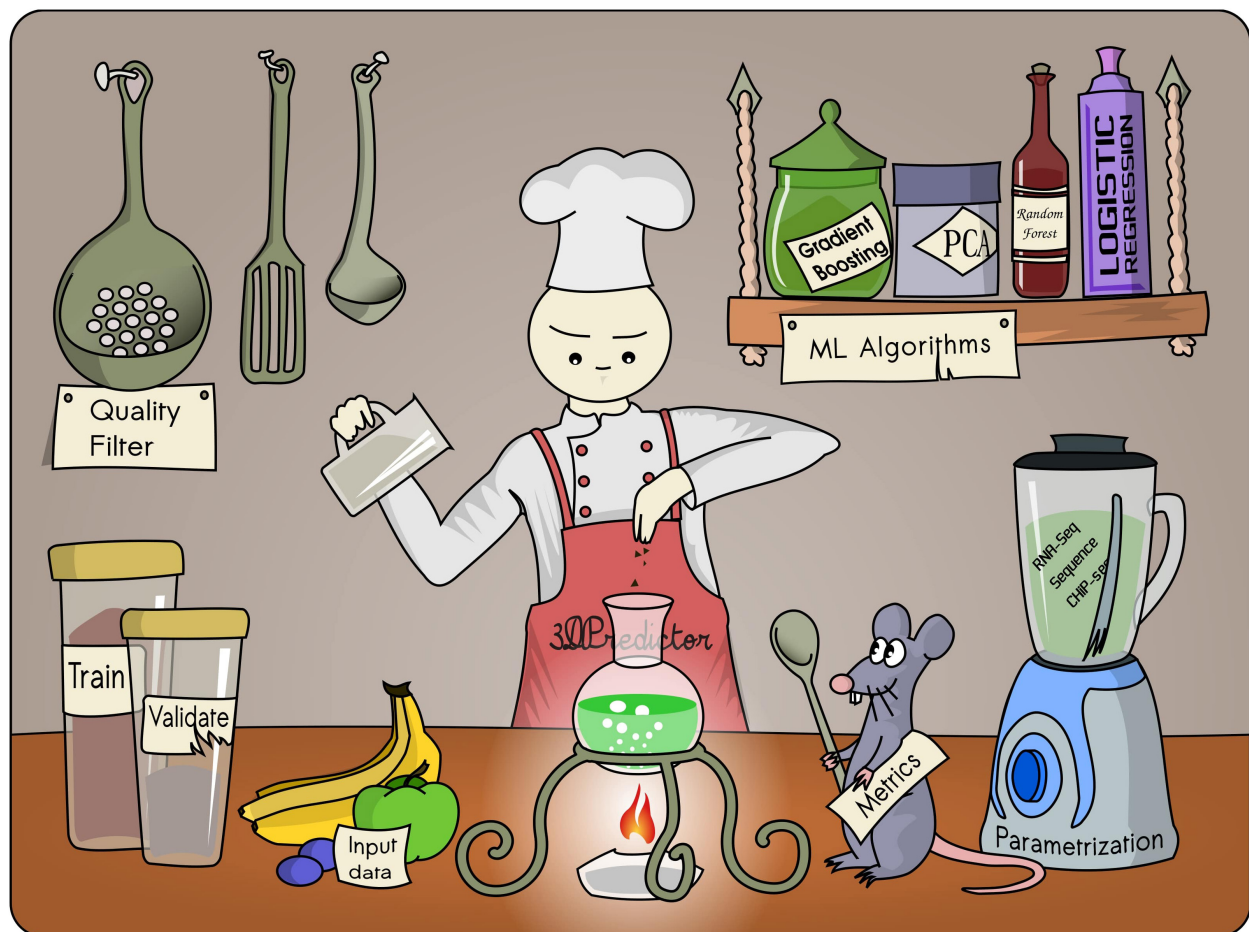


FIGURE 2 | In designing a machine learning-based algorithm, one should carefully choose the main “ingredients” required for good prediction.

gradient-boosting algorithm that gives the ranked list of features that helps to find the best feature. Anyway, analysis of features and algorithm parameters can inspire thoughts of biological mechanisms underlying the studying process.

Challenges and Limitations

Defining Target Features and Their Properties

The development of a predictive algorithm should start from a clear statement of biological features one wants to predict. Clear definitions of the features are important for the selection of positive and negative samples as well as for the choice of the machine learning algorithm.

Let us consider the goal of the prediction of interacting promoter–enhancer pairs. How would one define positive cases, i.e., interacting pairs? Now, it is clear that the majority of loops (see **Box 1** for details of mechanisms underlying these structures) observed on Hi-C maps are due to the synergetic activity of cohesin and CTCF proteins. These complexes form loops that might facilitate interactions of promoters and enhancers located within the looping region by reducing the spatial distance between them but do not necessarily directly

mediate contacts between these regulatory elements. In accord with this, direct functional tests based on targeted enhancer deletions or CRISPR-interference approaches (Gasperini et al., 2019) indicate that the vast majority of interacting enhancer–promoter pairs do not overlap with loop anchors although they are often located within a reasonable distance from them (Belokopytova et al., 2020). Thus, functionally interacting enhancer–promoter pairs might show only a slight increase in contact frequency. It is worth noting that the NG Capture-C approach (Davies et al., 2015) provides more sensitive and robust quantitation and enables detecting more significant interactions than Hi-C; however, typical Hi-C data are more widespread and available. At the same time, the majority of algorithms predicting 3-D genome structures are classifiers, so they solve the question of whether the promoter and enhancer interact, answering yes or no. We argue that quantitative measurement and prediction of spatial enhancer–promoter interactions are more informative than qualitative attribution to the loop anchors, and regression-based methods are more suited for such predictions.

Another example of varying feature definition is loop prediction. In this case, authors often use loops called by

specific algorithms as positive samples. A large proportion of loop calls varies between algorithms and visually assessed loops (Belokopytova et al., 2020; Salameh et al., 2020). Methods for loop detection, such as for TAD detection, are constantly improving. For example, the last published method Peakachu for loop calling can detect more loops than previous algorithms (Salameh et al., 2020). The same applies to TAD calling: Zufferey et al. (2018) compared 22 different TAD caller algorithms and found that TAD sizes and numbers vary significantly among callers and data resolutions.

To sum up, it is very important to consider the nature and biological properties of target features and carefully design positive and negative samples if using classifiers for prediction.

Predicting Single-Cell Data

The statistical approach is well applicable for 3-D genome structure prediction and investigation, but it uses population data. It allows getting a prediction that is actually a mean value for a cell population, which does not provide information about the 3-D genome organization of a single cell and differences of spatial contacts between distinct cells. Conversely, physical modeling always produces ensembles of single-cell chromatin configurations. Nevertheless, it does not mean that this prediction matches a real biological cell exactly even if its average matches population Hi-C data. However, recently Conte et al. (2020) show the consistent agreement between the predicted structures and independent single-cell super-resolution microscopy data, which provides evidence that, at least in the studied loci, polymer physics approaches accurately capture single-cell chromatin conformation. This issue is under active debate, however.

Understanding Mechanisms Underlying Prediction

Another limitation is that one cannot extract a simple algebraic formula transforming features into target feature values from a trained machine learning model. Therefore, the statistical dependencies found by machine learning algorithms are difficult to interpret in biological terms. Nevertheless, it is possible to evaluate the feature's contribution to prediction. We have already discussed several approaches for estimation of feature importance above; in addition, modifying features *in silico* and accessing how the modifications impact prediction could provide insights about the role of biological features used for prediction (Fudenberg et al., 2020).

Choosing Data Parameterization Function

To train a machine learning model, input data should be represented in a specific format, typically as a numeric vector of fixed length. The process of conversion of the input data into the desired format is called parameterization, and choosing the parameterization function might not be trivial. For example, ChIP-seq data is often used for the prediction of spatial chromatin contacts. There are several ways to submit these data to the algorithm: as a sum of ChIP-seq signals in the interval between two genome loci of interest, the total number of peaks in this region, the signal value of the nearest ChIP-seq peaks, or the *p*-values of peaks, etc. In our

experience, differences in parameterization could significantly affect prediction accuracy. Thus, the most challenging part is to choose the best way of parameterization to achieve the best performance of the algorithm.

Input Data Quality

Another important issue is the quality of the training data. Some machine learning algorithms are sensitive to outliers presented in the data. In this case, data smoothing should be performed before training the model. For example, for Hi-C and RNA-seq data, it is often useful to log-transform values.

Recently, high-resolution Hi-C maps were published (Hsieh et al., 2015, 2020; Krietenstein et al., 2020). They reveal chromatin structures in more detail and thereby improve predictions. Moreover, we noticed that the prediction of higher resolution heat maps is more accurate than the prediction of the same heat map but with a lower resolution (Belokopytova et al., 2020). This aspect is explained by features used for prediction. We gain lots of information from ChIP-seq data, in which the protein-binding event is attributed to a small locus (usually less than 200 base pairs). In this case, using an ultra-high resolution of Hi-C maps provides a better correspondence between protein-binding sites and interacting loci, allowing the model to learn effects mediated by specific proteins in a more direct way.

Overfitting

Another problem of machine learning approaches is overfitting. In this case, the model performs well on the training data set but does not perform well on a holdout sample, actually not capturing real complex patterns underlying the 3-D genome structure. Non-overlapping subsets for training and validation help to detect overfitting. There are two main ways to minimize overfitting: training the network on more examples and changing the complexity of the network. However, in the case of biological data, it is not always possible to have enough training samples. To increase the number of samples, it may be necessary to combine data from multiple sources. This leads to the next challenge: to normalize data from different sources that require rigorous data preprocessing (Xu and Jackson, 2019).

WHAT DO WE CONSIDER A GOOD PREDICTION?

Any data type has its data specificities, and this is also true for the Hi-C maps discussed below. It should be remembered that, usually for 3-D chromatin architecture, prediction binary classifiers or regression-based methods are used. There are some common metrics to access the binary classifier's performance, such as *f1*-score, AUC, and others. These metrics do not have any special characteristics related to genomic data.

The performance estimation of regression-based methods is more specific for Hi-C maps. How can we understand that one heat map is similar to another? Actually, a Hi-C map is a matrix of numbers, so we can apply any metrics for matrices comparison.

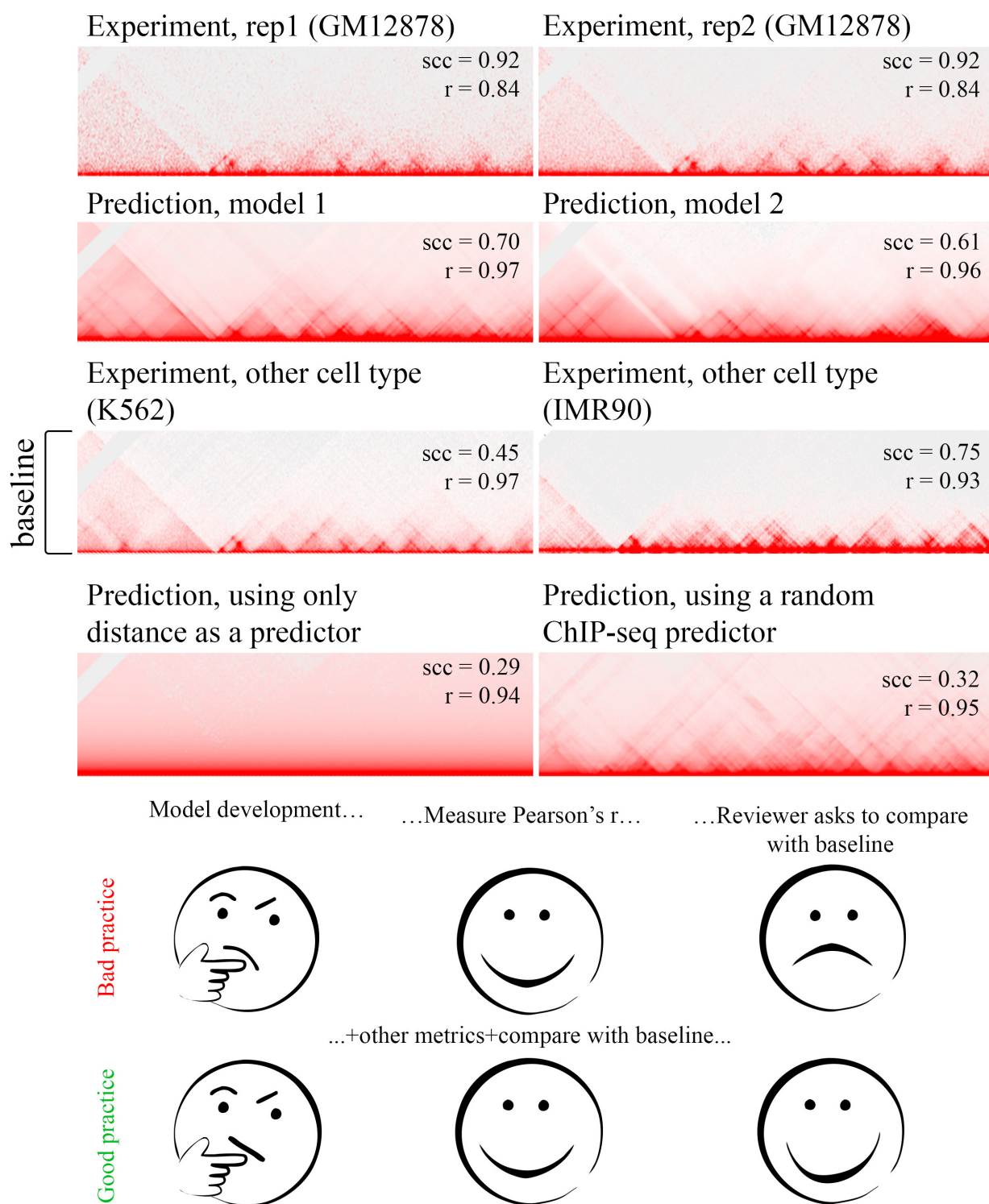


FIGURE 3 | The choice of the baseline plays a key role in assessing the prediction accuracy. Experimental data are from Rao et al. (2014); predictions generated using 3DPredictor (Belokopytova et al., 2020) supplemented with following data: genomic distance, CTCF, and RNA-seq (model 1); (Qi and Zhang, 2019) (model 2).

The basic metric is Pearson's correlation. Let us consider, for instance, a Pearson's correlation equal to 0.8: Does this correspond to a good or bad prediction? Intuitively, it seems

that a Pearson's correlation equal to 0.8 indicates accurate prediction. However, using absolute values is not a good idea. As we discussed above, contact probability shows prominent

dependence from distance, and even very simple prediction algorithms efficiently capture this dependence. Even when the distance between loci is not directly provided, it could be inferred from many epigenetic features. For example, cumulative ChIP-seq signals scale with the length of the genomic region, allowing prediction of contact probability. As we show in **Figure 3**, using randomly shuffled ChIP-seq signals, which have no biological meaning, allows the generation of predictions highly correlating with experimental data. Also, the whole-map correlation coefficient does not reflect the prediction of specific topological structures, such as TADs, loops, or compartments.

There are several workarounds allowing the comparison of Hi-C maps using correlation coefficients. First, one can compare the correlation between predicted and experimental data with the correlation between experimental replicates. Ideally, the prediction should be as similar to the experimental data as replicates among themselves. However, replicates are not always available; in addition, Tao Yang et al. show that Pearson's correlation between unrelated samples sometimes is equal to differences between replicates (**Figure 3** in Yang et al., 2017).

Another baseline could be obtained by scoring differences of Hi-C maps between distinct cell types. Chromatin organization is moderately conserved between different cell types (Dixon et al., 2012; Battulin et al., 2015) and even between different species (Fishman et al., 2019; Nuriddinov and Fishman, 2019), thus predicting cell type-specific features might be more challenging than an overall 3-D organization. For a high-quality algorithm, one would expect the difference between prediction and experimental data on the target cell type to be less than between different cell types. Besides this, one should carefully select data sets for comparison, accounting for their noise level. The lower noise level in the experimental data on target cell type results in higher measures of prediction accuracy, whereas a high noise level in a cell type used for baseline results in low baseline metrics, thus overestimating predictive power.

To overcome the limitations of standard correlations as measurements of Hi-C map similarity, Tao Yang et al. propose a framework that minimizes the effect of noise and biases by smoothing the Hi-C matrix, and then it addresses the distance-dependence effect by stratifying Hi-C data according to their genomic distance (Yang et al., 2017). This SCC metric distinguishes subtle differences between closely related cell lines, biological replicates, and pseudoreplicates, which was shown in the paper (**Figure 3** in Yang et al., 2017).

Besides Pearson's correlation and SCC standard metrics for comparison of matrices, such as MAE, MRE and others can be used for algorithm performance estimation. Similar to Pearson's correlation, understanding the values of these metrics requires a comparison with the baseline. Overall, we recommend using several metrics and several baselines for the optimal assessment of prediction accuracy (**Figure 3**).

Nevertheless, it is useful to visualize the predicted Hi-C map for empirical assessment to be confident that the chosen metric correctly reflects the differences between heat maps. Another way is to estimate the prediction of 3-D chromatin structures, such as TADs and loops. For some statistics, one can call loops or

insulator boundaries at experimental and predicting maps and then compare and overlap detected structures.

The selection of metrics for prediction accuracy estimation is an important issue for every algorithm. It should correctly reflect differences of 3-D chromatin features.

PREDICTION OF FUNCTIONAL CONSEQUENCES OF REARRANGEMENTS

Some rearrangements have been known to change the 3-D chromatin structure, causing diseases. Several works show the importance of chromatin folding in the gene regulation process (Franke et al., 2016; Rodríguez-Carballo et al., 2017; Kraft et al., 2019). Inversions, duplications, and other rearrangements can lead to TAD disruption, changing of promoter–enhancer interactions, and the emergence of new interactions between regulatory elements and genes. These insights are significant for medical genetics because the interpretation of chromosomal rearrangements in non-coding regions remains a big challenge. Zepeda-Mendoza et al. (2018) suggest detailed instructions on how to run a computational pipeline that identifies relevant candidates of non-coding balanced and apparently balanced chromosomal abnormality position effects. This pipeline includes analysis of TADs and the possibility of changing enhancer–promoter interactions due to rearrangement. Hence, the analysis of chromosomal rearrangement consequences in the context of the 3-D genome structure becomes a routine assay. The recently published machine learning algorithm TADA (Hertzberg et al., 2020) can prioritize large chromosomal alterations, such as copy number variants (CNVs) based on their pathogenicity.

Besides the prediction of the overall rearrangement effect, it is possible to predict changes in 3-D genome structures as TADs and loops. The 3D-GNOME algorithm (Sadowski et al., 2019; Wlasnowolski et al., 2020) generates chromatin 3-D structures using a Monte Carlo approach based on chromatin conformation capture (3C) data. It uses high-quality CTCF or RNA polymerase II ChIA-PET data as a reference chromatin interaction pattern. For rearrangement prediction, it applies a series of simple rules to recover chromatin interaction patterns. The 3D-GNOME algorithm can visualize alterations emerging in genomic structures after the introduction of SVs¹. Another approach is to predict changes in chromatin loops by a machine learning-based DeepMilo algorithm (Trieu et al., 2020). The algorithm can extract features directly from DNA sequences of loop anchors not using information about the presence and orientation of CTCF motifs. It allows predicting true Hi-C loops not having a CTCF signal at their anchors. DeepMILO can predict effects even of small mutations, and authors identified insulator loops predicted to change in multiple cancer patients and genes affected by these loops.

The aforementioned algorithms predict the perturbation of specific chromatin structures, such as loops and TADs. Other

¹<https://3dgenome.cent.uw.edu.pl/>

tools are capable of predicting a complete Hi-C map of the mutated locus. Algorithms such as Akita (Fudenberg et al., 2020), DeepC (Schwessinger et al., 2020), 3DPredictor (Belokopytova et al., 2020), PRISMR (Bianco et al., 2018), and others can predict alterations of 3-D chromatin architecture induced by structural variants.

An area of increasing interest and active research is the effect of small INDELs and single base pair variants on chromatin architecture. It is known that even single nucleotide replacement can lead to changes in 3-D genome structure, for example, by modifying CTCF binding sites (Schmiedel et al., 2016; Sun et al., 2020). A separate mission of predictive algorithms is to foresee the consequences of such mutations. Some algorithms, such as DeepMILO (Trieu et al., 2020), Akita (Fudenberg et al., 2020), and DeepC (Schwessinger et al., 2020) use a nucleotide sequence as the main feature for prediction. These algorithms are very powerful in predicting changes induced by small mutations because the mutations directly affect input features. On the other hand, training these algorithms requires knowledge of 3-D chromatin organization in wild-type cells of the same type because a nucleotide sequence does not provide cell type-specific epigenetic information.

Other algorithms do not use nucleotide sequences for prediction directly. In this case, it is important to model changes in input features caused by SNP or small INDEL. For instance, in the case of polymer modeling, it needs to change binder position or to remove the part of the polymer corresponding to the mutated DNA. All the same is about statistical approaches not using nucleotides as features for the prediction.

CONCLUSION

The mechanisms that underlie genome organization are intensively studied. Multiple groups developed computational algorithms to explain mechanisms underlying genome architecture and predict chromatin folding in normal and mutated cells. However, there is still no approach that is able to completely describe the whole complexity of the

nuclear organization. Physical models are limited by incomplete knowledge of mechanisms and relevant system parameters, such as interaction affinities and concentrations. Statistical methods do not allow understanding of the exact mechanisms underlying captured dependencies. And for both methods, it is not clear whether developed algorithms trained and validated using several cell types could be broadly and efficiently transferred to other cell types and conditions.

The latter question could be answered using the rapidly growing number of high-resolution Hi-C data sets. There are multiple published experimental data studying 3-D genome structure in normal and rearranged genomes. Such experiments provide detailed Hi-C maps of mutated regions that can be used as validation data for predictive algorithms. We believe that benchmarking and comparing existing predictive algorithms using these data sets would help to describe their power and limitations and to develop new, comprehensive approaches for the prediction of chromatin organization and dynamics in the future.

AUTHOR CONTRIBUTIONS

Both authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication

FUNDING

This work was supported by the RSF grant #19-74-00102. Computations have shown in **Figure 3** were performed using nodes of the Novosibirsk State University high-throughput computation cluster [supported by the Ministry of Education and Science of Russian Federation, grant #2019-0546 (FSUS-2020-0040)].

ACKNOWLEDGMENTS

We thank Emil Valeev and Olga Gladkih who helped us with designing illustrations.

REFERENCES

- Al Bkhetan, Z., and Plewczynski, D. (2018). Three-dimensional epigenome statistical model: genome-wide chromatin looping prediction. *Sci. Rep.* 8:5217. doi: 10.1038/s41598-018-23276-23278
- Andersson, R., and Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* 21, 71–87. doi: 10.1038/s41576-019-0173-178
- Banigan, E., van den Berg, A., Brandão, H., Marko, J., and Mirny, L. (2020). Chromosome organization by one-sided and two-sided loop extrusion. *Elife* 9:815340. doi: 10.1101/815340
- Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L. M., Dostie, J., Pombo, A., et al. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16173–16178. doi: 10.1073/pnas.1204799109
- Barbieri, M., Xie, S. Q., Torlai Triglia, E., Chiariello, A. M., Bianco, S., De Santiago, I., et al. (2017). Active and poised promoter states drive folding of the extended HoxB locus in mouse embryonic stem cells. *Nat. Struct. Mol. Biol.* 24, 515–524. doi: 10.1038/nsmb.3402
- Bates, F. S., and Fredrickson, G. H. (1990). Block copolymer thermodynamics: theory and experiment. *Annu. Rev. Phys. Chem.* 41, 525–557. doi: 10.1146/annurev.pc.41.100190.002521
- Battulin, N., Fishman, V. S., Mazur, A. M., Pomaznoy, M., Khabarova, A. A., Afonnikov, D. A., et al. (2015). Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol.* 16:77. doi: 10.1186/s13059-015-0642-640
- Beagrie, R. A., Scialdone, A., Schueler, M., Kraemer, D. C. A., Chotalia, M., Xie, S. Q., et al. (2017). Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* 543, 519–524. doi: 10.1038/nature21411

- Belokopytova, P. S., Nuriddinov, M. A., Mozheiko, E. A., Fishman, D., and Fishman, V. (2020). Quantitative prediction of enhancer–promoter interactions. *Genome Res.* 30, 72–84. doi: 10.1101/gr.249367.119
- Bianco, S., Lupiáñez, D. G., Chiariello, A. M., Annunziatella, C., Kraft, K., Schöpflin, R., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat. Genet.* 50, 662–667. doi: 10.1038/s41588-018-0098-98
- Bianco, V., Scialdone, A., and Nicodemi, M. (2012). Colocalization of multiple DNA loci: a physical mechanism. *Biophys. J.* 103, 2223–2232. doi: 10.1016/j.bpj.2012.08.056
- Boettiger, A., and Murphy, S. (2020). Advances in chromatin imaging at kilobase-scale resolution. *Trends Genet.* 36, 273–287. doi: 10.1016/j.tig.2019.12.010
- Brackley, C. A., Marenduzzo, D., and Gilbert, N. (2020). Mechanistic modeling of chromatin folding to understand function. *Nat. Methods* 17, 767–775. doi: 10.1038/s41592-020-0852-856
- Brackley, C. A., Brown, J. M., Waithe, D., Babbs, C., Davies, J., Hughes, J. R., et al. (2016a). Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol.* 17:59. doi: 10.1186/s13059-016-0909-900
- Brackley, C. A., Johnson, J., Kelly, S., Cook, P. R., and Marenduzzo, D. (2016b). Simulated binding of transcription factors to active and inactive regions folds human chromosomes into loops, rosettes and topological domains. *Nucleic Acids Res.* 44, 3503–3512. doi: 10.1093/nar/gkw135
- Brackley, C. A., Liebchen, B., Michieletto, D., Mouvet, F., Cook, P. R., and Marenduzzo, D. (2017). Ephemeral protein binding to DNA shapes stable nuclear bodies and chromatin domains. *Biophys. J.* 112, 1085–1093. doi: 10.1016/j.bpj.2017.01.025
- Brackley, C. A., Taylor, S., Papantonis, A., Cook, P. R., and Marenduzzo, D. (2013). Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc. Natl. Acad. Sci. U S A.* 110:E3605–E3611. doi: 10.1073/pnas.1302950110
- Buckle, A., Brackley, C. A., Boyle, S., Marenduzzo, D., and Gilbert, N. (2018). Polymer simulations of heteromorphic chromatin predict the 3D folding of complex genomic Loci. *Mol. Cell* 72, 786–797.e11. doi: 10.1016/j.molcel.2018.09.016
- Chiang, M., Michieletto, D., Brackley, C. A., Rattanaivirotkul, N., Mohammed, H., Marenduzzo, D., et al. (2019). Polymer modeling predicts chromosome reorganization in senescence. *Cell Rep.* 28, 3212–3223.e6. doi: 10.1016/j.celrep.2019.08.045
- Chiariello, A. M., Annunziatella, C., Bianco, S., Esposito, A., and Nicodemi, M. (2016). Polymer physics of chromosome large-scale 3D organisation. *Sci. Rep.* 6:29775. doi: 10.1038/srep29775
- Chiariello, A. M., Esposito, A., Annunziatella, C., Bianco, S., Fiorillo, L., Prisco, A., et al. (2017). A polymer physics investigation of the architecture of the murine orthologue of the 7q11.23 human locus. *Front. Neurosci.* 11:559. doi: 10.3389/fnins.2017.00559
- Cho, W. K., Spille, J. H., Hecht, M., Lee, C., Li, C., Grube, V., et al. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* 361, 412–415. doi: 10.1126/science.aar4199
- Conte, M., Fiorillo, L., Bianco, S., Chiariello, A. M., Esposito, A., and Nicodemi, M. (2020). Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. *Nat. Commun.* 11:3289. doi: 10.1038/s41467-020-17141-17144
- Davidson, I. F., Bauer, B., Goetz, D., Tang, W., Wutz, G., and Peters, J. M. (2019). DNA loop extrusion by human cohesin. *Science* 366, 1338–1345. doi: 10.1126/science.aaz3418
- Davies, J. O. J., Telenius, J. M., McGowan, S. J., Roberts, N. A., Taylor, S., Higgs, D. R., et al. (2015). Multiplexed analysis of chromosome conformation at vastly improved sensitivity. *Nat. Methods* 13, 74–80. doi: 10.1038/nmeth.3664
- de Gennes, P. G., and Witten, T. A. (1980). Scaling concepts in polymer physics. *Phys. Today* 33, 51–54. doi: 10.1063/1.2914118
- de Wit, E. (2020). TADs as the caller calls them. *J. Mol. Biol.* 432, 638–642. doi: 10.1016/j.jmb.2019.09.026
- Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390–403. doi: 10.1038/nrg3454
- Di Pierro, M., Cheng, R. R., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2017). De novo prediction of human chromosome structures: epigenetic marking patterns encode genome architecture. *Proc. Natl. Acad. Sci. U S A.* 114, 12126–12131. doi: 10.1073/pnas.1714980114
- Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., and Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. U S A.* 113, 12168–12173. doi: 10.1073/pnas.1613607113
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380. doi: 10.1038/nature11082
- Eagen, K. P., Aiden, E. L., and Kornberg, R. D. (2017). Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proc. Natl. Acad. Sci. U S A.* 114, 8764–8769. doi: 10.1073/pnas.1701291114
- Eeftens, J. M., Kapoor, M., and Brangwynne, C. P. (2020). Epigenetic memory as a time integral over prior history of Polycomb phase separation. *bioRxiv* [preprint]. doi: 10.1101/2020.08.19.254706
- Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* 20, 389–403. doi: 10.1038/s41576-019-0122-126
- Fiorillo, L., Musella, F., Kempfer, R., Chiariello, A., Bianco, S., Kukalev, A., et al. (2020). Comparison of the Hi-C, GAM and SPRITE methods by use of polymer models of chromatin. *bioRxiv* [preprint]. doi: 10.1101/2020.04.24.059915
- Fishman, V., Battulin, N., Nuriddinov, M., Maslova, A., Zlotina, A., Strunov, A., et al. (2019). 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. *Nucleic Acids Res.* 47, 648–665. doi: 10.1093/nar/gkyl1103
- Flyamer, I. M., Gassler, J., Imakaev, M., Brandão, H. B., Ulianov, S. V., Abdennur, N., et al. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* 544, 110–114. doi: 10.1038/nature21711
- Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* 538, 265–269. doi: 10.1038/nature19800
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell Rep.* 15, 2038–2049. doi: 10.1016/j.celrep.2016.04.085
- Fudenberg, G., Kelley, D. R., and Pollard, K. S. (2020). Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods* 17, 1111–1117. doi: 10.1038/s41592-020-0958-x
- Fudenberg, G., and Mirny, L. A. (2012). Higher-order chromatin structure: bridging physics and biology. *Curr. Opin. Genet. Dev.* 22, 115–124. doi: 10.1016/j.jgde.2012.01.006
- Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., et al. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science* 360, 102–105. doi: 10.1126/science.aar7831
- Gartner, T. E., and Jayaraman, A. (2019). Modeling and simulations of polymers: a roadmap. *Macromolecules* 52, 755–786. doi: 10.1021/acs.macromol.8b01836
- Gasparini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., et al. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* 176, 377–390.e19. doi: 10.1016/j.cell.2018.11.029
- Gibson, B. A., Doolittle, L. K., Schneider, M. W. G., Jensen, L. E., Gamarra, N., Henry, L., et al. (2019). Organization of chromatin by intrinsic and regulated phase separation. *Cell* 179, 470–484.e21. doi: 10.1016/j.cell.2019.08.037
- Grosberg, A. Y. (2016). Extruding loops to make loopy globules? *Biophys. J.* 110, 2133–2135. doi: 10.1016/j.bpj.2016.04.008
- Grosberg, A. Y., Nechaev, S. K., and Shakhnovich, E. I. (1988). The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys.* 49, 2095–2100. doi: 10.1051/jphys:0198800490120209500
- Hahnfeldt, P., Hearst, J. E., Brenner, D. J., Sachs, R. K., and Hlatky, L. R. (1993). Polymer models for interphase chromosomes. *Proc. Natl. Acad. Sci. U S A.* 90, 7854–7858. doi: 10.1073/pnas.90.16.7854
- Hertzberg, J., Mundlos, S., Vingron, M., and Gallone, G. (2020). TADA – a Machine learning tool for functional annotation based prioritisation of putative pathogenic CNVs. *bioRxiv* [preprint]. doi: 10.1101/2020.06.30.180711
- Hsieh, T. H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Rando, O. J., Tjian, R., et al. (2020). Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Mol. Cell* 78, 539–553.e8. doi: 10.1016/j.molcel.2020.03.002

- Hsieh, T. H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O. J. (2015). Mapping nucleosome resolution chromosome folding in yeast by Micro-C. *Cell* 162, 108–119. doi: 10.1016/j.cell.2015.05.048
- Huang, J., Marco, E., Pinello, L., and Yuan, G. C. (2015). Predicting chromatin organization using histone marks. *Genome Biol.* 16:162. doi: 10.1186/s13059-015-0740-z
- Imakaev, M. V., Fudenberg, G., and Mirny, L. A. (2015). Modeling chromosomes: beyond pretty pictures. *FEBS Lett.* 589, 3031–3036. doi: 10.1016/j.febslet.2015.09.004
- Johnson, J., Brackley, C. A., Cook, P. R., and Marenduzzo, D. (2015). A simple model for DNA bridging proteins and bacterial or human genomes: bridging-induced attraction and genome compaction. *J. Phys. Condens. Matter* 27:064119. doi: 10.1088/0953-8984/27/6/064119
- Jost, D., Carrivain, P., Cavalli, G., and Vaillant, C. (2014). Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.* 42, 9553–9561. doi: 10.1093/nar/gku698
- Kai, Y., Andricovich, J., Zeng, Z., Zhu, J., Tzatsos, A., and Peng, W. (2018). Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. *Nat. Commun.* 9:4221. doi: 10.1038/s41467-018-06664-6666
- Kantidze, O. L., and Razin, S. V. (2020). Weak interactions in higher-order chromatin organization. *Nucleic Acids Res.* 48, 4614–4626. doi: 10.1093/nar/gkaa261
- Kempfer, R., and Pombo, A. (2020). Methods for mapping 3D chromosome architecture. *Nat. Rev. Genet.* 21, 207–226. doi: 10.1038/s41576-019-0195-192
- Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J., and Yu, H. (2019). Human cohesin compacts DNA by loop extrusion. *Science* 366, 1345–1349. doi: 10.1126/science.aaz4475
- Kraft, K., Magg, A., Heinrich, V., Riemenschneider, C., Schöpflin, R., Markowski, J., et al. (2019). Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.* 21, 305–310. doi: 10.1038/s41556-019-0273-x
- Kragsteven, B. K., Spielmann, M., Paliou, C., Heinrich, V., Schöpflin, R., Esposito, A., et al. (2018). Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis. *Nat. Genet.* 50, 1463–1473. doi: 10.1038/s41588-018-0221-x
- Krientein, N., Abraham, S., Venev, S. V., Abdennur, N., Gibcus, J., Hsieh, T. H. S., et al. (2020). Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* 78, 554–565.e7. doi: 10.1016/j.molcel.2020.03.003
- Larson, A. G., Elnatan, D., Keenen, M. M., Trnka, M. J., Johnston, J. B., Burlingame, A. L., et al. (2017). Liquid droplet formation by HP1 α suggests a role for phase separation in heterochromatin. *Nature* 547, 236–240. doi: 10.1038/nature22822
- Li, R., Liu, Y., Li, T., and Li, C. (2016). 3Disease browser: a web server for integrating 3D genome and disease-associated chromosome rearrangement data. *Nat. Publ. Gr.* 6:34651. doi: 10.1038/srep34651
- Li, W., Wong, W. H., and Jiang, R. (2019). DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res.* 47:e60. doi: 10.1093/nar/gkz167
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293. doi: 10.1126/science.1181369
- Lin, D., Bonora, G., Yardimci, G. G., and Noble, W. S. (2019). Computational methods for analyzing and modeling genome structure and organization. *wiley interdiscip. Rev. Syst. Biol. Med.* 11:e1435. doi: 10.1002/wsbm.1435
- MacKay, K., and Kusalik, A. (2020). Computational methods for predicting 3D genomic organization from high-resolution chromosome conformation capture data. *Brief. Funct. Genom.* 19, 292–308. doi: 10.1093/bfpg/ela004
- MacPherson, Q., Beltran, B., and Spakowitz, A. J. (2018). Bottom-up modeling of chromatin segregation due to epigenetic modifications. *Proc. Natl. Acad. Sci. U S A.* 115, 12739–12744. doi: 10.1073/pnas.1812268115
- Mirny, L. A. (2011). The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.* 19, 37–51. doi: 10.1007/s10577-010-9177-9170
- Nicodemi, M., and Prisco, A. (2009). Thermodynamic pathways to genome spatial organization in the cell nucleus. *Biophys. J.* 96, 2168–2177. doi: 10.1016/j.bpj.2008.12.3919
- Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., and Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. U S A.* 115, E6697–E6706. doi: 10.1073/pnas.1717730115
- Nuriddinov, M., and Fishman, V. (2019). C-InterSector—a computational tool for interspecies comparison of genome architecture. *Bioinformatics* 35, 4912–4921. doi: 10.1093/bioinformatics/btz415
- Petrovic, J., Zhou, Y., Fasolino, M., Goldman, N., Schwartz, G. W., Mumbach, M. R., et al. (2019). Oncogenic notch promotes long-range regulatory interactions within hyperconnected 3D cliques. *Mol. Cell* 73, 1174–1190.e12. doi: 10.1016/j.molcel.2019.01.006
- Plys, A. J., Davis, C. P., Kim, J., Rizki, G., Keenen, M. M., Marr, S. K., et al. (2019). Phase separation of polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. *Genes Dev.* 33, 799–813. doi: 10.1101/gad.326488.119
- Qi, Y., and Zhang, B. (2019). Predicting three-dimensional genome organization with chromatin states. *PLoS Comput. Biol.* 15:e1007024. doi: 10.1371/journal.pcbi.1007024
- Quinodoz, S. A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J. M., Detmar, E., et al. (2018). Higher-Order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* 174, 744–757.e24. doi: 10.1016/j.cell.2018.05.024
- Rao, S. S. P., Huang, S. C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K. R., et al. (2017). Cohesin loss eliminates all loop domains. *Cell* 171, 305–320.e24. doi: 10.1016/j.cell.2017.09.026
- Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159, 1665–1680. doi: 10.1016/j.cell.2014.11.021
- Razin, S. V., and Gavrilov, A. A. (2020). The role of liquid–liquid phase separation in the compartmentalization of cell nucleus and spatial genome organization. *Biochemistry* 85, 643–650. doi: 10.1134/S0006297920060012
- Razin, S. V., and Ulianov, S. V. (2020). Divide and rule: phase separation in eukaryotic genome functioning. *Cells* 9:2480. doi: 10.3390/cells9112480
- Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P. J., Beccari, L., El-Idrissi, I., et al. (2017). The HoxD cluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes Dev.* 31, 2264–2281. doi: 10.1101/gad.307769.117
- Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., et al. (2017). Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* 67, 837–852.e7. doi: 10.1016/j.molcel.2017.07.022
- Sadowski, M., Kraft, A., Szalaj, P., Wlasnowolski, M., Tang, Z., Ruan, Y., et al. (2019). Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. *Genome Biol.* 20:148. doi: 10.1186/s13059-019-1728-x
- Salameh, T. J., Wang, X., Song, F., Zhang, B., Wright, S. M., Khunsriraksakul, C., et al. (2020). A supervised learning framework for chromatin loop detection in genome-wide contact maps. *Nat. Commun.* 11:3428. doi: 10.1038/s41467-020-17239-17239
- Sanborn, A. L., Rao, S. S. P., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U S A.* 112, E6456–E6465. doi: 10.1073/pnas.1518552112
- Sanulli, S., Trnka, M. J., Dharmarajan, V., Tibble, R. W., Pascal, B. D., Burlingame, A. L., et al. (2019). HP1 reshapes nucleosome core to promote phase separation of heterochromatin. *Nature* 575, 390–394. doi: 10.1038/s41586-019-1669-1662
- Schmiedel, B. J., Seumois, G., Samaniego-Castruita, D., Cayford, J., Schulten, V., Chavez, L., et al. (2016). 17q21 asthma-risk variants switch CTCF binding and regulate IL-2 production by T cells. *Nat. Commun.* 7:13426. doi: 10.1038/ncomms13426
- Schwessinger, R., Gosden, M., Downes, D., Brown, R. C., Oudelaar, A. M., Telenius, J., et al. (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods* 17, 1118–1124. doi: 10.1038/s41592-020-0960-963
- Scialdone, A., Cataudella, I., Barbieri, M., Prisco, A., and Nicodemi, M. (2011). Conformation regulation of the X chromosome inactivation center: a model. *PLoS Comput. Biol.* 7:1002229. doi: 10.1371/journal.pcbi.1002229
- Serra, F., Di Stefano, M., Spill, Y. G., Cuartero, Y., Goodstadt, M., Baù, D., et al. (2015). Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett.* 589, 2987–2995. doi: 10.1016/j.febslet.2015.05.012
- Spicuglia, S., and Vanhille, L. (2012). Chromatin signatures of active enhancers. *Nucleus* 3, 126–131. doi: 10.4161/nucl.19232
- Strom, A. R., Emelyanov, A. V., Mir, M., Fyodorov, D. V., Darzacq, X., and Karpen, G. H. (2017). Phase separation drives heterochromatin domain formation. *Nature* 547, 241–245. doi: 10.1038/nature22989

- Sun, Y., Dong, L., Zhang, Y., Lin, D., Xu, W., Ke, C., et al. (2020). 3D genome architecture coordinates trans and cis regulation of differentially expressed ear and tassel genes in maize. *Genome Biol.* 21:143. doi: 10.1186/s13059-020-02063-2067
- Szabo, Q., Bantignies, F., and Cavalli, G. (2019). Principles of genome folding into topologically associating domains. *Sci. Adv.* 5:eaaw1668. doi: 10.1126/sciadv.aaw1668
- Szabo, Q., Donjon, A., Jerković, I., Papadopoulos, G. L., Cheutin, T., Bonev, B., et al. (2020). Regulation of single-cell genome organization into TADs and chromatin nanodomains. *Nat. Genet.* 52, 1151–1157. doi: 10.1038/s41588-020-00716-718
- Tavares-Cadete, F., Norouzi, D., Dekker, B., Liu, Y., and Dekker, J. (2020). Multi-contact 3C reveals that the human genome during interphase is largely not entangled. *Nat. Struct. Mol. Biol.* 27, 1105–1114. doi: 10.1038/s41594-020-0506-5
- Teif, V. B., Kepper, N., Yserentant, K., Wedemann, G., and Rippe, K. (2015). Affinity, stoichiometry and cooperativity of heterochromatin protein 1 (HP1) binding to nucleosomal arrays. *J. Phys. Condens. Matter.* 27:064110. doi: 10.1088/0953-8984/27/6/064110
- Teif, V. B., and Rippe, K. (2010). Statistical-mechanical lattice models for protein-DNA binding in chromatin. *J. Phys. Condens. Matter.* 22:414105. doi: 10.1088/0953-8984/22/41/414105
- Tiana, G., and Luca, G. (2019). “Modeling the 3D conformation of genomes,” in *Raton: Taylor & Francis, 2018. | Series: Series in Computational Biophysics*, eds G. Tiana and L. Giorgetti Boca (Boca Raton, FL: CRC Press). doi: 10.1201/9781315144009.
- Trieu, T., Martinez-Fundichely, A., and Khurana, E. (2020). DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biol.* 21:79. doi: 10.1186/s13059-020-01987-1984
- Ulianov, S. V., Doronin, S. A., Khrameeva, E. E., Kos, P. I., Luzhin, A. V., Starikov, S. S., et al. (2019). Nuclear lamina integrity is required for proper spatial organization of chromatin in *Drosophila*. *Nat. Commun.* 10:1176. doi: 10.1038/s41467-019-09185-y
- Ulianov, S. V., Khrameeva, E. E., Gavrilov, A. A., Flyamer, I. M., Kos, P., Mikhaleva, E. A., et al. (2016). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. *Genome Res.* 26, 70–84. doi: 10.1101/gr.196006.115
- Vian, L., Pękowska, A., Rao, S. S. P., Kieffer-Kwon, K. R., Jung, S., Baranello, L., et al. (2018). The energetics and physiological impact of cohesin extrusion. *Cell* 173, 1165–1178.e20. doi: 10.1016/j.cell.2018.03.072
- Whalen, S., Truty, R. M., and Pollard, K. S. (2016). Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48, 488–496. doi: 10.1038/ng.3539
- Wlasnowolski, M., Sadowski, M., Czarnota, T., Jodkowska, K., Szalaj, P., Tang, Z., et al. (2020). 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. *Nucleic Acids Res.* 48, W170–W176. doi: 10.1093/nar/gkaa388
- Xu, C., and Jackson, S. A. (2019). Machine learning and complex biological data. *Genome Biol.* 20:76. doi: 10.1186/s13059-019-1689-1680
- Xu, H., Zhang, S., Yi, X., Plewczynski, D., and Li, M. J. (2020). Exploring 3D chromatin contacts in gene regulation: the evolution of approaches for the identification of functional enhancer-promoter interaction. *Comput. Struct. Biotechnol. J.* 18, 558–570. doi: 10.1016/j.csbj.2020.02.013
- Xu, T., Zheng, X., Li, B., Jin, P., Qin, Z., and Wu, H. (2018). A comprehensive review of computational prediction of genome-wide features. *Brief. Bioinform.* doi: 10.1093/bib/bby110 Online ahead of print .
- Yang, T., Zhang, F., Yardimci, G. G., Song, F., Hardison, R. C., Noble, W. S., et al. (2017). HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* 27, 1939–1949. doi: 10.1101/gr.220640.117
- Zepeda-Mendoza, C. J., Menon, S., and Morton, C. C. (2018). Computational prediction of position effects of human chromosome rearrangements. *Curr. Protoc. Hum. Genet.* 97:e57. doi: 10.1002/cphg.57
- Zhang, R., Wang, Y., Yang, Y., Zhang, Y., and Ma, J. (2018). Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* 34, i133–i141. doi: 10.1093/bioinformatics/bty248
- Zhang, S., Chasman, D., Knaack, S., and Roy, S. (2019). In silico prediction of high-resolution Hi-C interaction matrices. *Nat. Commun.* 10:5449. doi: 10.1038/s41467-019-13423-8
- Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J. W., et al. (2016). Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* 7:10812. doi: 10.1038/ncomms10812
- Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.* 19:217. doi: 10.1186/s13059-018-1596-1599

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Belokopytova and Fishman. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



ERBB1/2/3 Expression, Prognosis, and Immune Infiltration in Cutaneous Melanoma

Shougang Liu^{1†}, Rong Geng^{2,3†}, Eryi Lin¹, Peizhen Zhao¹ and Yongfeng Chen^{1*}

¹ Department of Dermatology, Dermatology Hospital, Southern Medical University, Guangzhou, China, ² Department of Gynecology and Obstetrics, The First Affiliated Hospital, Jinan University, Guangzhou, China, ³ Department of Gynecology, Affiliated Foshan Maternity and Child Healthcare Hospital, Southern Medical University, Foshan, China

OPEN ACCESS

Edited by:

Tatiana V. Tatarinova,
University of La Verne, United States

Reviewed by:

Vijaykumar Muley,
Universidad Nacional Autónoma
de México, Mexico
Balaji Banoth,
St. Jude Children's Research
Hospital, United States

*Correspondence:

Yongfeng Chen
gdnpicyf@126.com

[†] These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 02 September 2020

Accepted: 28 January 2021

Published: 01 March 2021

Citation:

Liu S, Geng R, Lin E, Zhao P and
Chen Y (2021) ERBB1/2/3
Expression, Prognosis, and Immune
Infiltration in Cutaneous Melanoma.
Front. Genet. 12:602160.
doi: 10.3389/fgene.2021.602160

Background: The four ERBB tyrosine kinase family members [ERBB1 (epidermal growth factor receptor, EGFR), ERBB2 (HER2), ERBB3 (HER3), and ERBB4 (HER4)] (ERBB receptor family) have been shown, according to previous studies, to be related to the cutaneous melanoma. ERBB3 is the only member of the ERBBs that lacks tyrosine kinase activity and thus needs to dimer with other tyrosine kinases receptors to trigger the signaling pathway, while ERBB3 may dimer with all members of the ERBB family. Melanoma progression depends on activation of ERBB signaling, especially the ERBB3/ERBB2 cascade. There are lymphocytes and T cell infiltrates in melanoma. Numerous pieces of evidences indicate that local immune status plays an important role in the formation of anti-tumor immune responses. However, the relationship between the ERBBs and prognosis and immune infiltration in cutaneous melanoma is not completely clear.

Methods: The expression of the ERBBs was analyzed through the Oncomine database, Gene Expression Profiling Interactive Analysis (GEPIA), respectively. Immunohistochemistry of ERBBs was obtained from the Human Protein Atlas is increased before HPA database. ERBBs genes expression and mutation analysis in cutaneous melanoma from the cBioPortal. Functional annotation and Kyoto Encyclopedia of Genes and Genomes is increased before KEGG pathway enrichment analysis from the Metascape. Correlations between ERBBs and 31 genes that were close to each other and frequently altered were explored by GEPIA. Using the GEPIA database, we also investigated the relationship between ERBBs and myeloid-derived suppressor cells (MDSC) in cutaneous melanoma. The disease-free survival and different tumor stages of ERBBs were evaluated by GEPIA. The correlation of ERBBs and tumor-infiltrating immune cells and prognostic(5 years survival rates) was tested by the Tumor Immune Estimation Resource (TIMER).

Results: In general, the expression levels of ERBB1/2 in cutaneous melanoma were lower than those in normal skin tissue. By contrast, the ERBB3 expression level was higher in cutaneous melanoma than in normal skin tissue. Low expression of ERBB1/2 and high expression of ERBB3 were detrimental to the 5 years survival of cutaneous melanoma patients (ERBB1: log-rank P: 0.03; ERBB2: log-rank P: 0.008; ERBB3: log-rank P: 0.039). ERBB4 expression may not affect the prognosis of patients

with cutaneous melanoma. ERBBs may not play a role in the tumor stage and disease-free survival in cutaneous melanoma patients. The relationship between the ERBB family and 31 genes that were close to each other and frequently altered is demonstrated as the genes regulated by the ERBB family being mainly concentrated in the RAS/RAF/MEK/ERK signaling pathway. ERBB2 can induce infiltration of CD8⁺ T cells and B cells, while ERBB3 can induce infiltration of CD4⁺ T cells, CD8⁺ T cells, and Neutrophil cells. ERBBs are more significantly associated with M1 macrophages, dendritic cells, Th1, Th2, Th17, and Treg cellular immune markers (Cor > 0.2). ERBB2/3 were related to MDSC in cutaneous melanoma, including human mononuclear myeloid-derived suppressor cells (M-MDSC) and polymorphonuclear myeloid-derived suppressor cells (PMN-MDSC), and may influence the progression of cutaneous melanoma through MDSC, but the conclusion needs further probing.

Conclusion: This study investigated the prognosis and immune infiltration of the ERBB family in cutaneous melanoma. Our results suggest that ERBB1/2/3 may serve as early prognostic markers and potential therapeutic targets in cutaneous melanoma.

Keywords: ERBB family, immunochemistry, immune infiltrates, prognostic, cutaneous melanoma

INTRODUCTION

Melanoma originates from melanocytes and accounts for the highest proportion of skin cancer-related deaths. At present, the treatment of melanoma metastasis is still difficult, although researchers have explored a variety of methods. The 5 years overall survival (OS) of patients with early melanoma is higher, ranging from 24 to 29%, but the 5-year OS of patients with stage IIIC and IV is only 10–19%. In recent years, biological immunotherapy has gradually emerged, which can prolong the survival period of patients (Homet Moreno and Ribas, 2015). Therefore, further searching for potential pathogenic factors and pathogenic mechanisms can help identify potential prognostic markers and drug targets in melanoma.

The ERBB family belongs to the tyrosine kinase I subfamily and is composed of four closely related transmembrane tyrosine kinase receptors, all of which are encoded by the proto-oncogene HER1–4: ERBB1 (EGFR), ERBB2 (HER2), ERBB3 (HER3), and ERBB4 (HER4). They all have tyrosine kinase activity and play a key role in signal transduction. The tyrosine kinase active domain of each member of the ERBBs family is highly conserved, with a high degree of homology in structure and function. It is mainly expressed during human embryonic development and regulates the growth, survival, transformation, and apoptosis of normal (Jørgensen and Hersom, 2012; Griffin and Ramirez, 2017).

The overexpression and activation of the ERBB family are closely related to the clinicopathological characteristics and prognosis of various cancers, including melanoma, lung cancer, gastric cancer, breast cancer, etc. (Bittoni et al., 2015; Oudard et al., 2015). It is generally believed that abnormally functional carcinogenic pathways lead to melanoma. These pathways may include the EMT signaling pathway (Pietraszek-Gremplewicz et al., 2019), PI3K signaling pathway (Mujoo et al., 2014; Song et al., 2015), and so on. ERBBs can enhance the ability of tumor cells to migrate and invade, promote tumor angiogenesis,

and inhibit tumor cell apoptosis (Iqbal and Iqbal, 2014) through these pathways.

Previous studies have shown that ERBBs disorders are closely related to the clinicopathological characteristics and prognosis of human cutaneous melanoma. However, the potential role of ERBBs family members in cutaneous melanoma remains unintelligible. In this study, we analyzed the expression, mutation, prognosis, and immune infiltration of ERBBs family members in cutaneous melanoma through a variety of databases.

MATERIALS AND METHODS

Oncomine Database Analysis

Oncomine¹ is an online bioinformatics analysis tool that includes 18,000 cancer gene expression microarrays (Rhodes et al., 2007). Through the Gene Summary view in the Oncomine database, we determined the expression level of ERBB in cutaneous melanoma. Use the following values: *P*-value 0.01, fold change 2, top 10% of the gene ranking, and mRNA data type to determine the threshold.

GEPIA Database Analysis

Online database GEPIA² is an online database utilized to analyze the expression data of RNA sequencing in TCGA and GTEx projects. It can also generate gene Expression profiles, the expression in the box plot, and the main stages of pathology (Tang et al., 2017). The expression of ERBB was determined by the SKCM data set of GEPIA. The following values were used to determine the threshold: the *P*-value was 0.01, the multiple changes were 2, and it matched the normal value of TCGA and GTEx data.

¹<https://www.oncomine.org/resource/login.html>

²<http://gepia.cancer-pku.cn/index.html>

Human Protein Atlas Database Analysis

The HPA (version 19.3)³ (Navani, 2016) is a large-scale research project, the database will help researchers to explore protein expression in human tissue and cells. In this study, immunohistochemical images were used to analyze ERBBs protein expression in cutaneous melanoma and normal tissues.

cBioPortal Database Analysis

cBioPortal⁴ was used to further analyze the expression of ERBB (Gao et al., 2013) through the skin melanoma data set, which included 479 pathological reports. Co-expression and network analysis were performed based on the online instructions of cBioPortal.

Functional Annotation and KEGG Pathway Enrichment Analysis

Gene function annotation and KEGG pathway analysis were used to uncover the underlying mechanism of cutaneous melanoma. MetaScape⁵ (Zhou et al., 2019) was updated in 2018 and is a web-based tool that provides gene function annotation and enrichment analysis.

TIMER Database Analysis

The TIMER⁶ is used to study the expression characteristics of tumor-immune interaction genes in more than 30 cancer types (Li et al., 2017) to evaluate various the clinical impact of different immune cells of immune type. We analyzed the relationship between the expression of ERBBs and immune infiltration in cutaneous melanoma through the TIMER database.

Statistical Analysis

The results we generate using Oncomine were displayed by *P*-values, fold changes, and ranks ($p < 0.05$, fold change > 2). *P*-value and fold change were used to show the outcomes of GEPIA ($p < 0.01$, fold change > 2). Also, the Spearman correlation analysis is used to evaluate the relationship between genes and determine the strength of the correlation between genes by absolute value. $P < 0.05$ were considered statistically significant.

RESULTS

Transcriptional Levels of ERBBs in Cutaneous Melanoma and Other Cancers

The transcription levels of ERBBs were compared in cutaneous melanoma and normal samples by the ONCOMINE database (Figure 1A and Table 1). ERBB3 was upregulated in cutaneous melanoma in two datasets. In Riker's and Talantov's datasets (Talantov et al., 2005; Riker et al., 2008), ERBB3 is overexpressed compared with the normal samples: cutaneous melanoma with

a fold change of 4.667 and 2.264, respectively. In Riker's dataset (Riker et al., 2008), ERBB1 (EGFR) is under-expressed in cutaneous melanoma with a fold change of -8.110 . Talantov's dataset (Talantov et al., 2005) revealed another mRNA expression ERBB1 with a fold change of -7.657 ; that is, ERBB2 has a fold change of -3.952 and -2.229 in cutaneous melanoma compared with normal skin tissues in Riker's and Talantov's datasets, respectively. Based on the information afforded by the GEPIA database, a comparative study of ERBBs mRNA expression in normal skin tissue and cutaneous melanoma tissue was carried out. The results showed that, compared with normal skin tissue, cutaneous melanoma tissue had lower expression levels of ERBB1/2 and a higher expression level of ERBB3 (Figure 1B). In the immunohistochemistry supplied by the HPA data set, we discovered that ERBB1 was moderately expressed, ERBB2 was lowly expressed, and the ERBB3/4 proteins were highly expressed in cutaneous melanoma (Figure 2).

Further, we explored the expression of the ERBBs in other cancers models by ONCOMINE database (Figure 1A) showed that ERBB1 mRNA level was significantly higher in the bladder, brain and CNS, head and neck, kidney, lung, lymphoma cancer tissues, and markedly lower in the breast, colorectal, gastric, myeloma, ovarian, and sarcoma cancers compared with the corresponding normal tissues. The ERBB2 mRNA level was observably higher in the bladder, brain and CNS, breast, and pancreatic cancer tissues and signally lower in the colorectal, esophageal, head and neck, kidney, lung, lymphoma, and sarcoma cancers compared with the corresponding normal tissues. The ERBB3 mRNA level was dramatically higher in bladder, colorectal, kidney, lung, ovarian, and prostate cancer tissues and significantly lower in the brain and CNS, esophageal, head and neck, lymphoma, myeloma, and sarcoma cancers compared with the corresponding normal tissues. The ERBB4 mRNA level was higher in the breast and colorectal cancer tissues and observably lower in the bladder, brain and CNS, head and neck, kidney, and lung cancers compared with the corresponding normal tissues. These results show that ERBB1/2 are of low expression, and ERBB3 is of high expression, which suggests that ERBB1/2/3 may be potential prognostic markers in cutaneous melanoma.

ERBBs Mutation Rates, Their Influence on Neighboring Genes, and Their Correlation in Cutaneous Melanoma

cBioPortal is used to analyze the changes and network of ERBBs in cutaneous melanoma. In 479 cases, ERBBs were changed in 158 samples (33%), and two or more changes were detected in 53 samples (11%) (Figure 3A). Besides, we analyzed the mRNA expression of ERBBs (RNA Seq V2 RSEM) through Pearson correlation to calculate the relationship between ERBBs, and the results showed that ERBB1 and ERBB2 are positively correlated (Figure 3B). Subsequently, we established a network of ERBBs through the 31 most frequently changing neighbor genes [The Network TAB provides interaction analysis and network visualization of cancer changes in cBioPortal, the network includes Pathway, HPRD (human reference protein database), Reactome, NCI (National Cancer Institute)—Nature

³<https://www.proteinatlas.org/>

⁴<https://www.cbioportal.org/>

⁵<http://metascape.org>

⁶<https://cistrome.shinyapps.io/timer/>

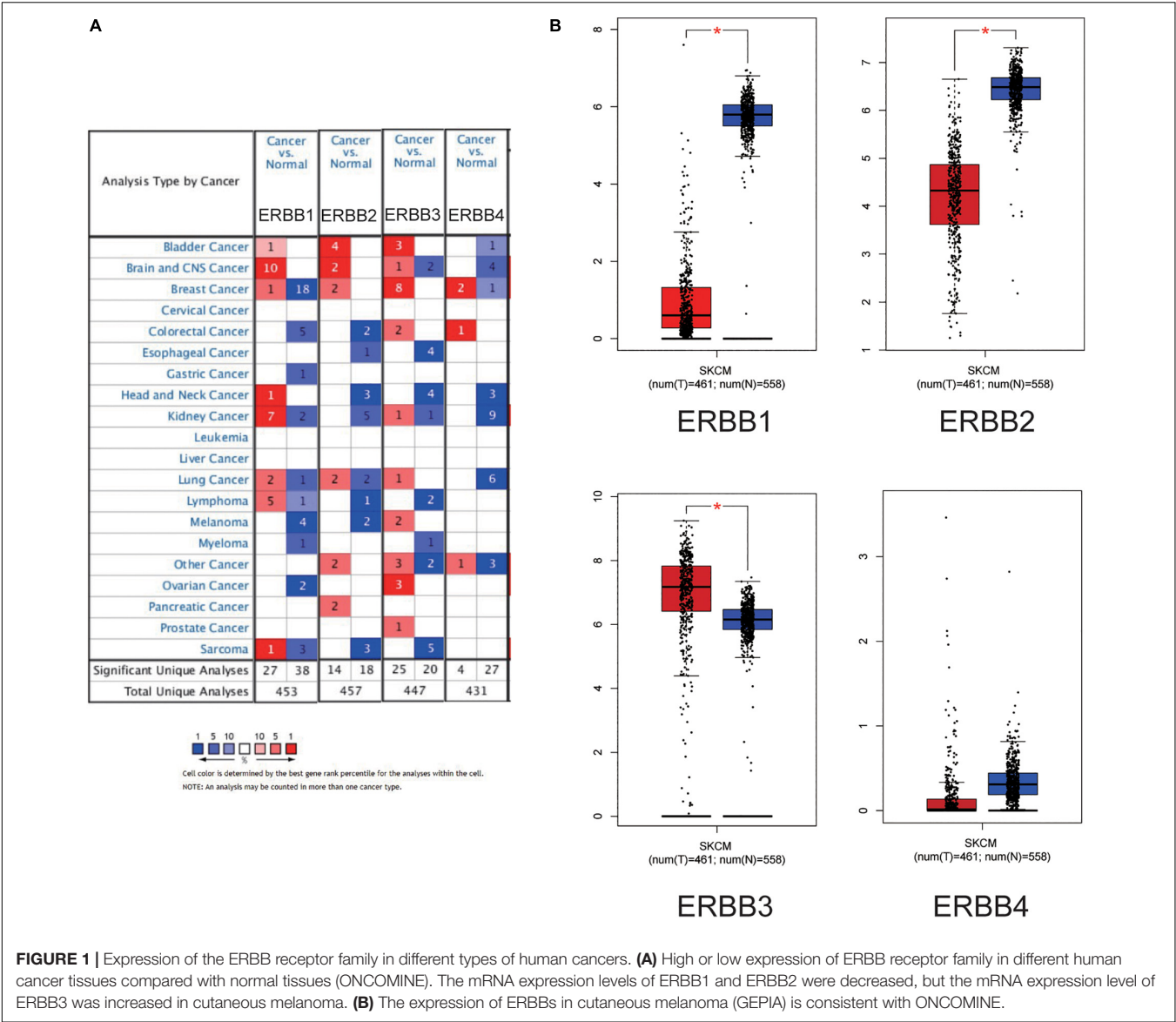


TABLE 1 | Conspicuous variation of ERBBs expression in transcription level between cutaneous melanoma and skin tissues (ONCOMINE database).

	Type vs. normal	Fold change	P-value	t-test	Reference
ERBB1	Cutaneous melanoma vs. normal	−8.110	2.44E-8	−10.259	Riker
	Cutaneous melanoma vs. normal	−7.657	1.84E-12	−16.418	Talantov
ERBB2	Cutaneous melanoma vs. normal	−3.952	2.70E-6	−6.762	Riker
	Cutaneous melanoma vs. normal	−2.229	5.86E-5	−6.116	Talantov
ERBB3	Cutaneous melanoma vs. normal	4.667	4.73E-9	12.658	Talantov
	Cutaneous melanoma vs. normal	2.264	0.006	2.891	Riker
ERBB4	NA	NA	NA	NA	NA

Fold change >> 2, P-value< < 0.05.

Pathway Interaction Database, Memorial Sloan Kettering Cancer Center (MSKCC), map of cancer cells⁷, From The Open Source Pathway Commons Project. By default, cBioportal

⁷<http://cancer.cellmap.org>

automatically generates a network containing all queries about the genetic neighbors (adjacent nodes), sequenced according to the alternating frequency of the genome of the selected cancer]. This shows that there is a close pertinence between changes of ERBBs and cell proliferation or differentiation, Fibroblast

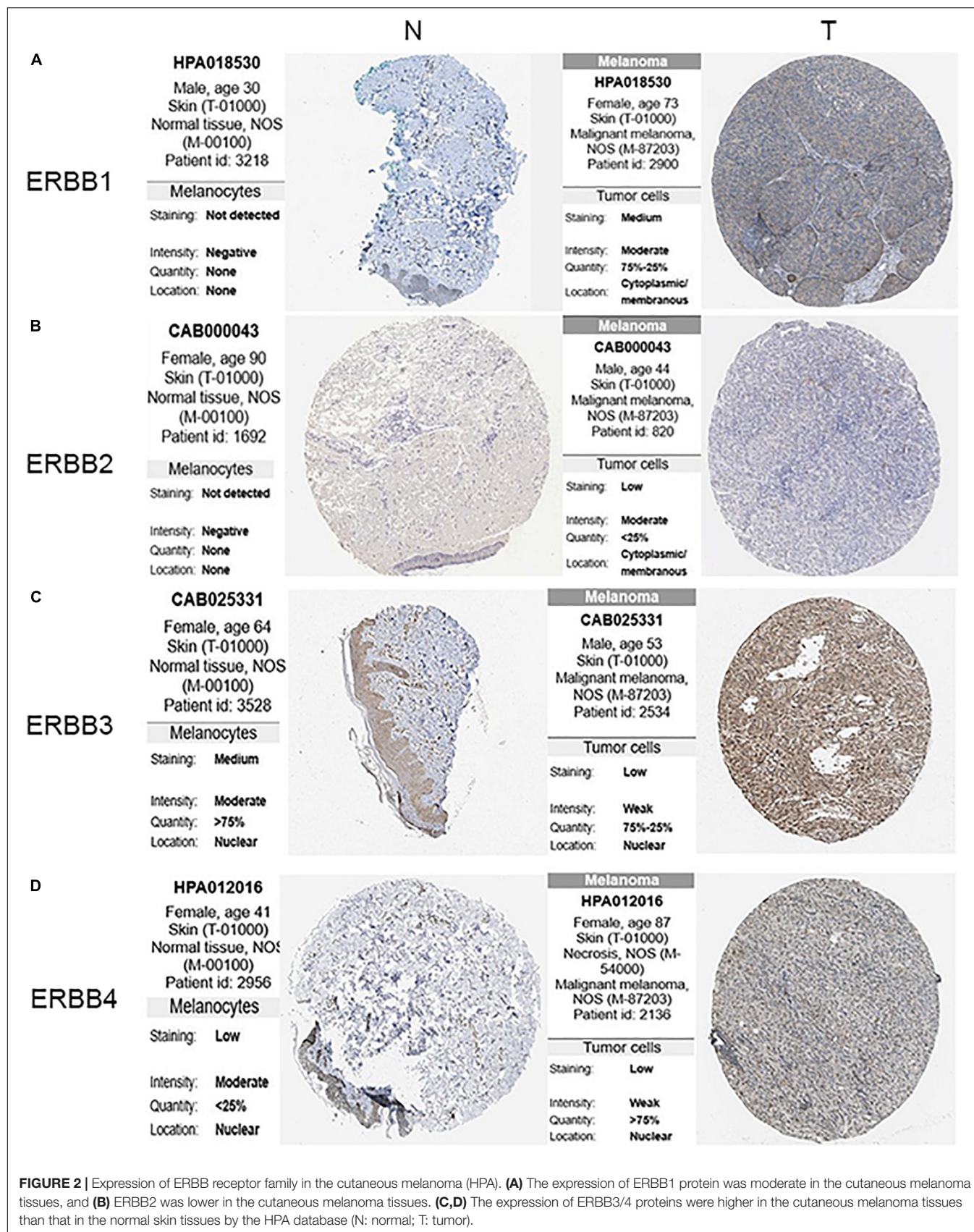
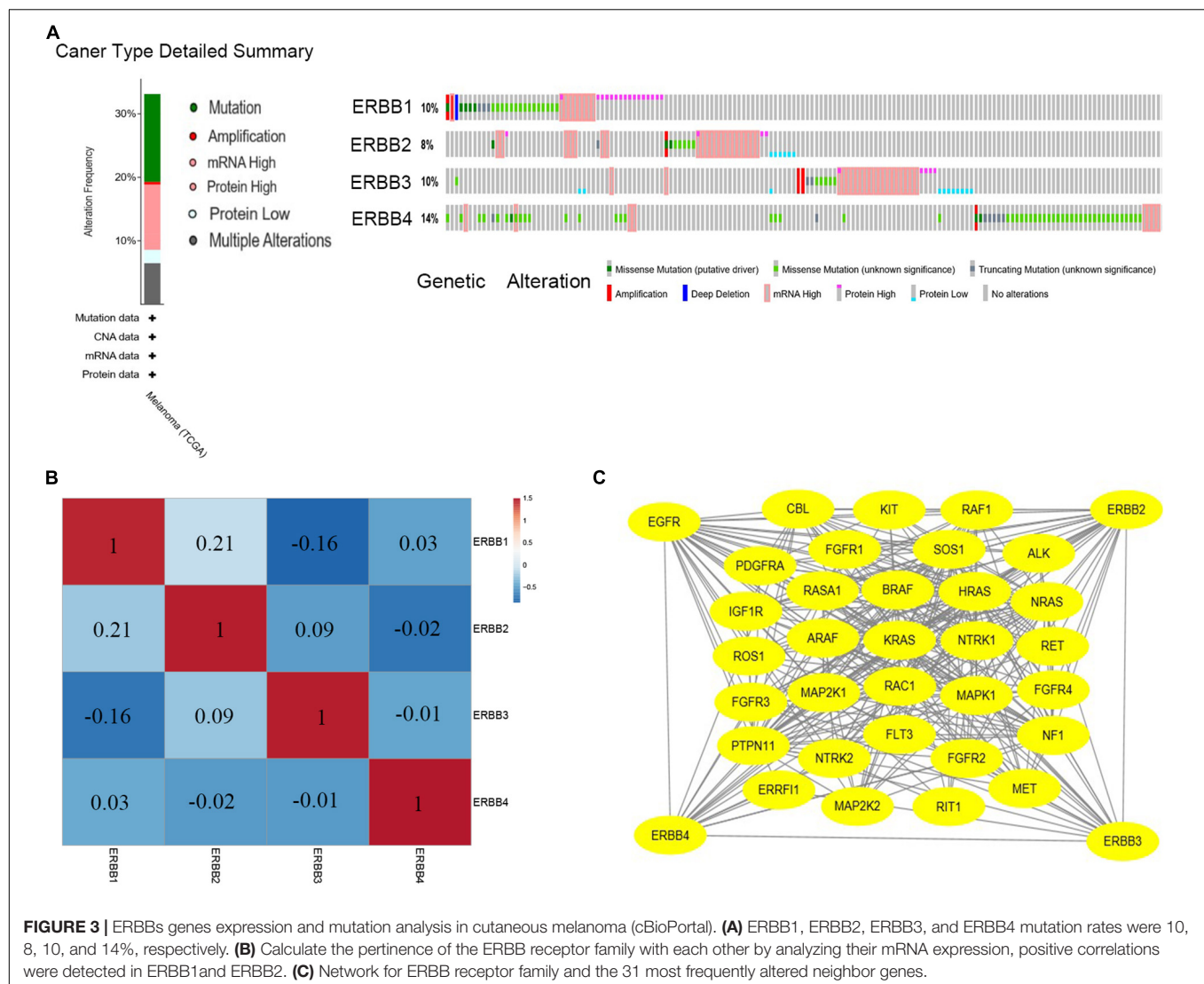


FIGURE 2 | Expression of ERBB receptor family in the cutaneous melanoma (HPA). **(A)** The expression of ERBB1 protein was moderate in the cutaneous melanoma tissues, and **(B)** ERBB2 was lower in the cutaneous melanoma tissues. **(C,D)** The expression of ERBB3/4 proteins were higher in the cutaneous melanoma tissues than that in the normal skin tissues by the HPA database (N: normal; T: tumor).



growth factor receptor (FGFR), including FGFR1/2/3/4, as well as MAPK pathways-related genes, including MAPK1, MAP2K1, and MAP2K2 (**Figure 3C**). Evaluate the function of clustering genes through GO and pathway analysis, and these enriched pathways are closely interrelated to each other. **Figure 4** provides the consequence of the functions and pathways of the markedly enriched genes. We discovered that the enriched genes are related to multiple pathways, such as MAPK family signal cascade, MAPK cascade regulation, EGFR tyrosine kinase inhibitor resistance, receptor tyrosine kinase signal transduction, and transmembrane receptor protein tyrosine kinase signal transduction pathway, etc.

Further, through the GEPIA database (**Supplementary Table 1**), in cutaneous melanoma, we examined the association between the ERBBs and 31 genes were close to each other and frequently altered and observed that all the genes except KIT, ALK, NTRK1, FGFR4, and MET were affected by the changes of some ERBB family members. In this network, ERBB-regulated genes are mainly concentrated in the RAS-RAF-MEK-ERK

signaling pathway, while the ERBB family may have no direct regulatory relationship with SOS, NF1, PDGFRA, FGFR1, RASA1, ROS1, FGFR3, RAC1, PTPN11, NTRK2, FLT3, FGFR2, ERRF1, and RIT1 genes.

Immune Cell Infiltration of ERBBs in Cutaneous Melanoma

ERBBs affect the clinical prognosis of patients by participating in the inflammatory response and immune cell infiltration. TIMER displays the relationship between ERBBs and immune infiltrating cells. There was a positive pertinence between ERBB1 expression and the infiltration of CD4⁺ T cells (Cor = 0.152, $p = 1.30 \times 10^{-3}$), Macrophages (Cor = 0.258, $p = 2.66 \times 10^{-8}$), and Neutrophils (Cor = 0.143, $p = 2.36 \times 10^{-3}$; **Figure 5A**). ERBB2 expression was negatively related to the infiltration of B cells (Cor = -0.116, $p = 1.42 \times 10^{-2}$) and CD8⁺ T cells (Cor = -0.109, $p = 2.24 \times 10^{-2}$) and positively associated with the infiltration of CD4⁺ T cells (Cor = 0.19, $p = 5.39 \times 10^{-5}$;

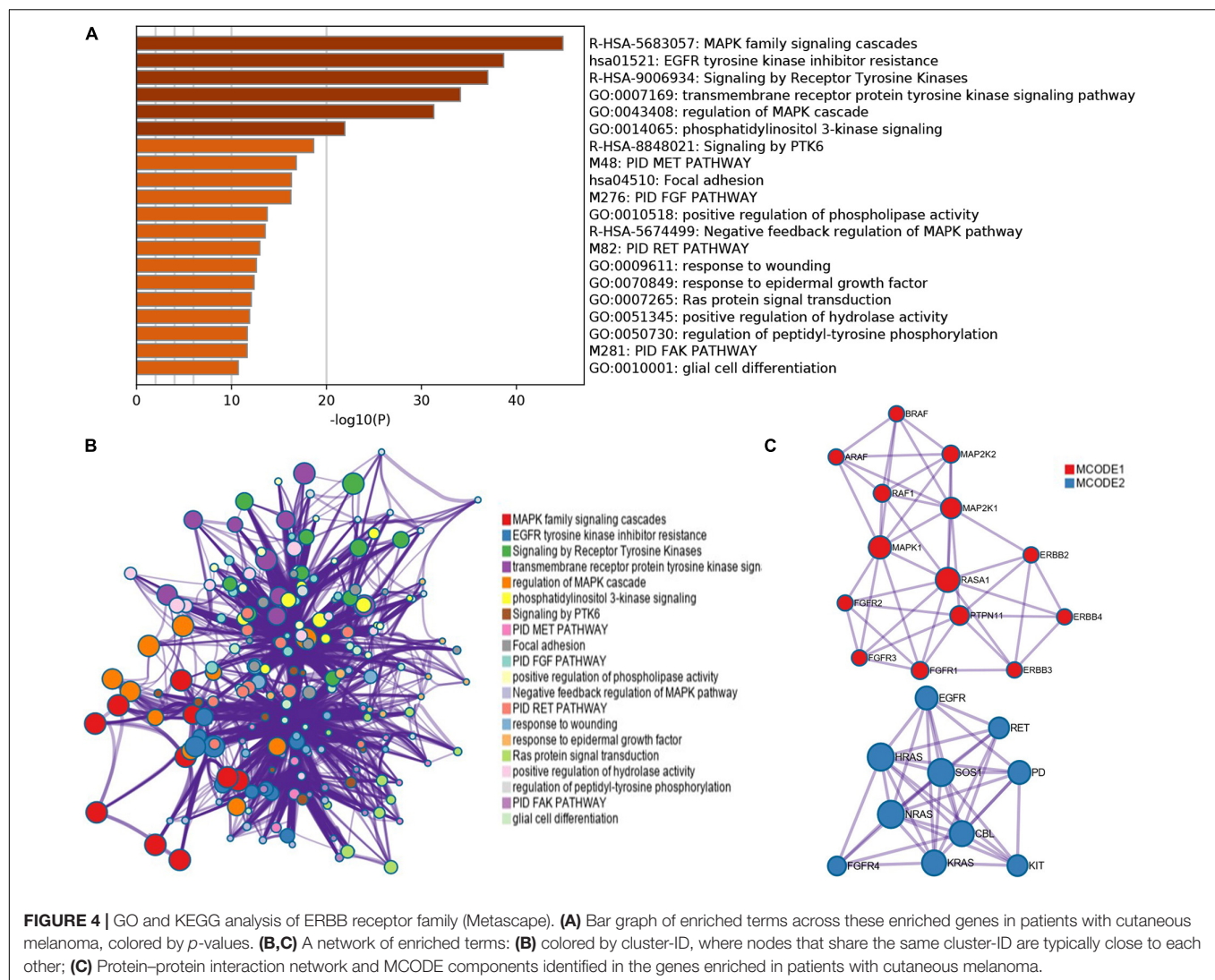


FIGURE 4 | GO and KEGG analysis of ERBB receptor family (Metascape). **(A)** Bar graph of enriched terms across these enriched genes in patients with cutaneous melanoma, colored by p -values. **(B,C)** A network of enriched terms: **(B)** colored by cluster-ID, where nodes that share the same cluster-ID are typically close to each other; **(C)** Protein-protein interaction network and MCODE components identified in the genes enriched in patients with cutaneous melanoma.

Figure 5B). ERBB3 expression was positively interrelated to the infiltration of CD8⁺ T cells (Cor = 0.121, $p = 1.15\text{e-}02$) and CD4⁺ T cells (Cor = 0.148, $p = 1.72\text{e-}03$) and Neutrophil (Cor = 0.147, $p = 1.69\text{e-}03$; **Figure 5C**). There was a positive correlation between ERBB4 expression and the infiltration of CD8⁺ T cells (Cor = 0.141, $p = 3.12\text{e-}03$) and neutrophils (Cor = 0.122, $p = 9.21\text{e-}03$; **Figure 5D**). These results reveal that ERBB2/3 are more closely related to immune infiltration in cutaneous melanoma, which suggests that the role of ERBB2/3 in regulating tumor immunity.

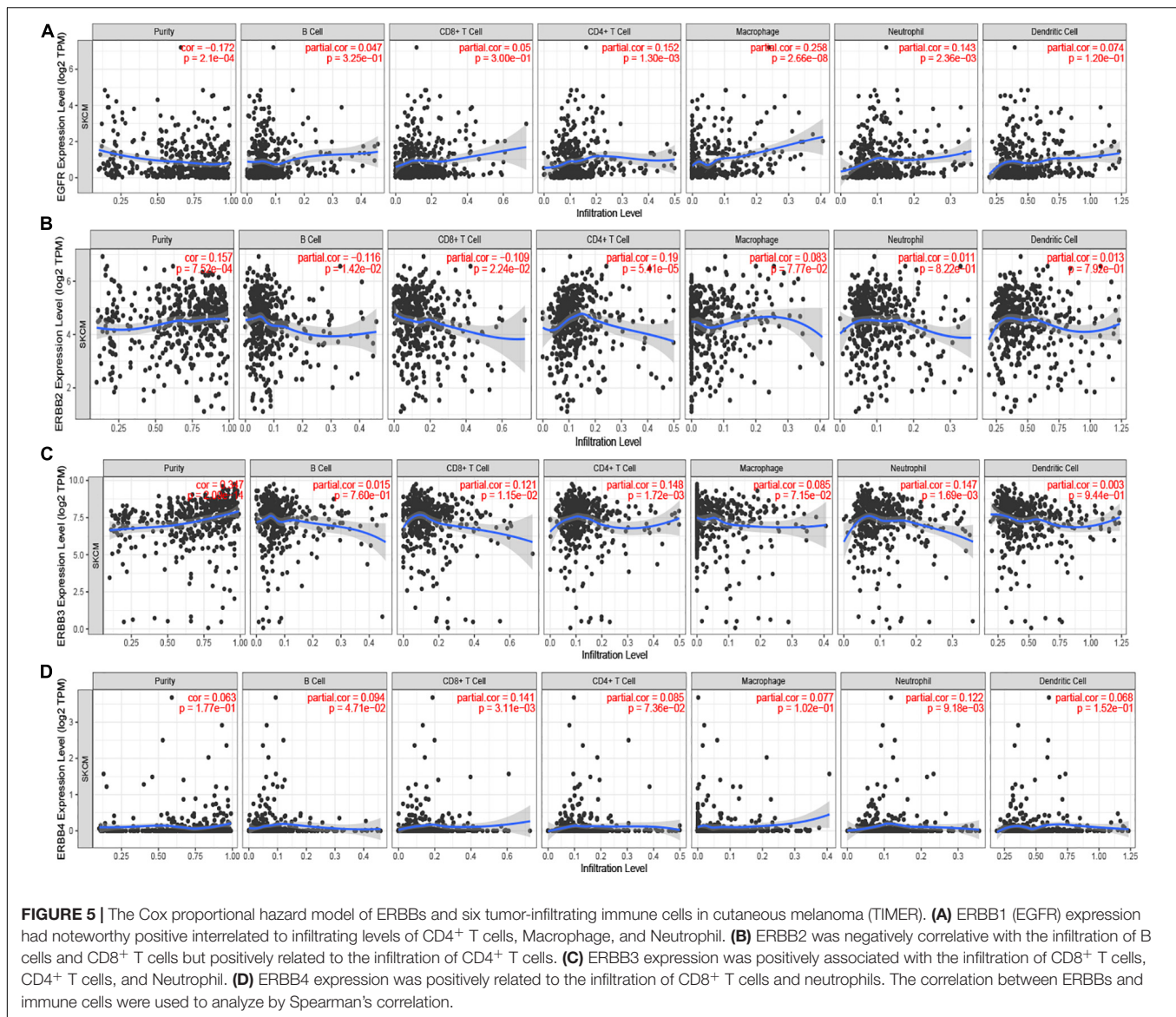
ERBBs CNV Is Correlated With Immune Infiltration Levels in Cutaneous Melanoma

ERBB CNV has a signal related to infiltrating levels of B cells, CD8⁺ T cells, CD4⁺ T cells, Macrophages, Neutrophils, and Dendritic Cells. ERBB 1 (EGFR) induces the infiltrating levels of B cells, CD4⁺ T cells, Neutrophils, and Dendritic Cells in cutaneous melanoma (**Figure 6A**). ERBB2 induces

the infiltrating levels of B cells, CD8⁺ T cells, CD4⁺ T cells, Macrophages, Neutrophils, and Dendritic Cells (**Figure 6B**). ERBB3 induces the infiltrating levels of B cells, CD4⁺ T cells, Macrophages, Neutrophils, and Dendritic Cells (**Figure 6C**). ERBB4 induces the infiltrating levels of B cells, CD4⁺ T cells, and Dendritic Cells (**Figure 6D**). ERBB1/2/4 CNV are correlated with arm-level deletion. ERBB3 CNV is connected with arm-level gain. These results suggest that ERBBs CNV induces immune infiltration in cutaneous melanoma.

Correlation Analysis Between ERBBs and Markers of Immune Cells in Cutaneous Melanoma

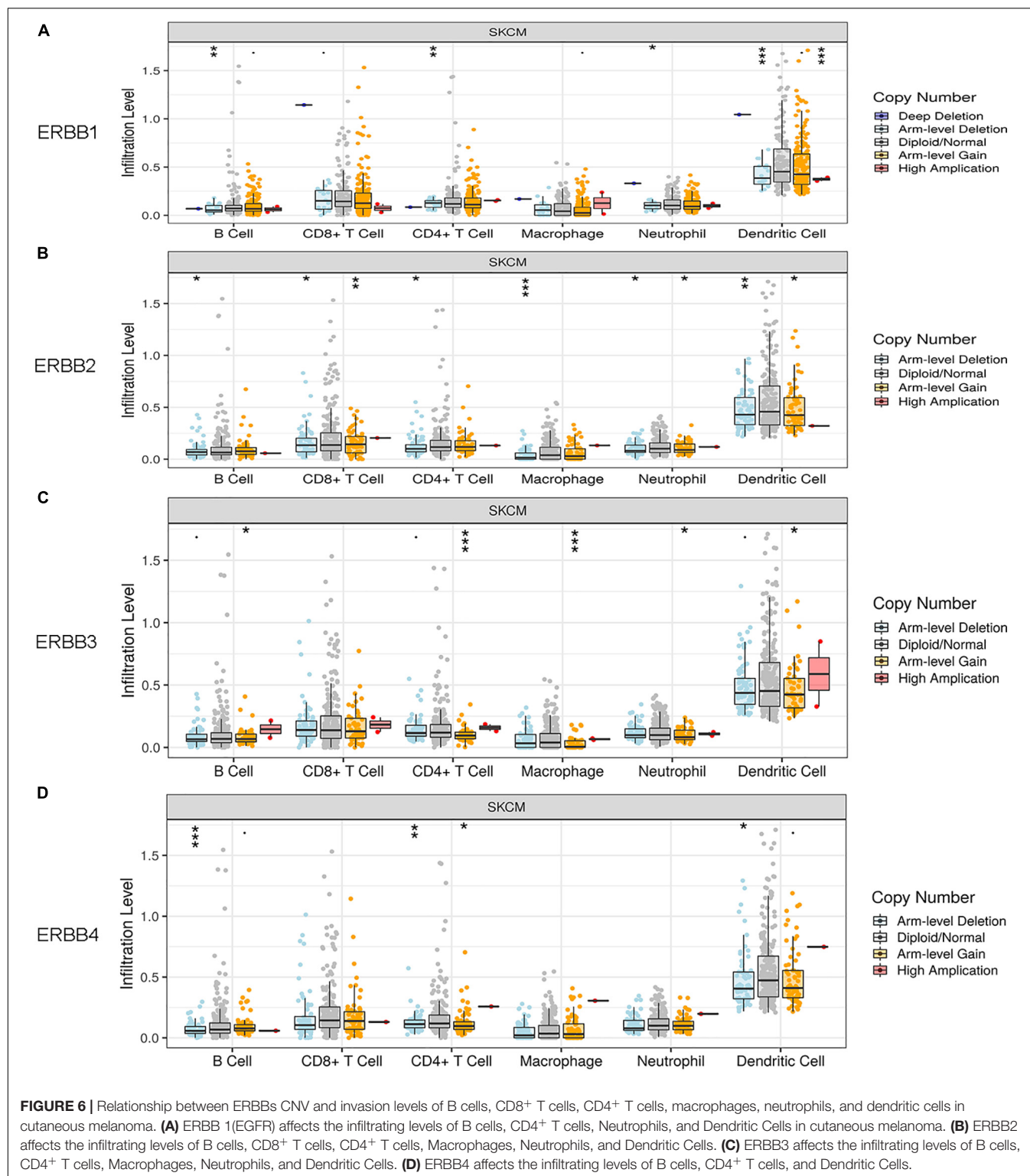
We further evaluated the relationship between the ERBBs levels and immune infiltrating cells through the TIMER database and according to the expression levels of immune marker genes in cutaneous melanoma tissues. The immune cells analyzed by the TIMER database include CD8⁺ T cells, CD4⁺ T



cells, B cells, monocytes, tumor-associated macrophages (TAM), M1 and M2 macrophages, neutrophils, dendritic cells, and different subgroups of T cells, namely, T helper 1 (Th1), Th2, Tfh, Th17, regulatory T (Tregs), and T cell exhaustion. Because tumor purity will affect the level of immune infiltration of clinical samples, the purity of the relevant analysis was adjusted (Table 2).

Specifically, ERBB1 expression displayed dramatically interrelated to the expression of specific immune cells markers, such as B cell marker, CD79A, Monocyte marker, CD115, TAM marker, CCL2, M1 macrophage marker, iNOS and COX2, M2 Macrophage marker, CD163, Neutrophils marker, CD11B, Dendritic cell marker, BDCA-1 and BDCA-4. Th1 markers, STAT4 and TNF- α , Th2 markers, GATA3 and STAT6, the Tfh marker, BCL6, Th17 and STAT3, Treg markers, CCR8, STAT5 and TGF β , T cell exhaustion markers, PD-1, LAG3, and GZMB. The memorable pertinence between ERBB2

expression and the expression of specific immune cells markers is shown, such as the B-cell marker, CD19 and CD79a, Monocyte marker, CD115, Dendritic cell marker, HLA-DPB1 and HLA-DPA1, Th1 marker IFN- γ , Th2 marker, STAT5A, Tfh marker, BCL6, Th17 marker, STAT3, Treg marker, CCR8, STAT5B, and TGF β , T cell exhaustion marker, CTLA4, and GZMB. ERBB3 expression indicated significantly related to the expression of specific immune cells markers, such as the CD8⁺ T cell marker, CD8A, T cell (general) marker, CD2, TAM marker, CD68, Neutrophils, CD66b, Dendritic cell, HLA-DPB1, HLA-DRA and HLA-DPA1, Th1 marker, STAT4, STAT1 and IFN- γ , Th2 marker, STAT6 and STAT5, Th17 marker, STAT3, Treg marker, CCR8, STAT5B, and TGF β . ERBB4 expression manifested memorably relevant to the expression of specific immune cells markers, such as the CD8⁺ T cell marker, CD8A, T cell (general) marker, CD2, B cell marker, CD19, CD79A, Monocyte marker, CD86, CD115,



TAM marker, CCL2, M1 Macrophage marker, COX2, M2 Macrophage marker, CD163, VSIG4, MS4A4A, Neutrophils marker, CD11b, and CCR7, Dendritic cell marker, BDCA-1 and BDCA-4, Th1 marker, T-bet, STAT4 and STAT1, Th2 marker, GATA3 and STAT5A, Tfh marker, BCL6 and IL21,

Th17 marker, STAT3, Treg marker, CCR8 and STAT5B, T cell exhaustion marker, and TIM-3. The study shows that the ERBBs and M1 Macrophage, Dendritic cell, Th1, Th2, Th17, and Treg cells have a more significant relationship ($\text{Cor} > 0.2$) in cutaneous melanoma.

TABLE 2 | Correlation analysis between ERBBs and related markers of immune cells.

Description	Gene markers	SKCM							
		Purity							
		ERBB1		ERBB2		ERBB3		ERBB4	
		Cor	P	Cor	P	Cor	P	Cor	P
CD8 ⁺ T cell	CD8A	−0.053	2.62e-1	−0.045	3.37e-01	0.133	**	0.1	*
	CD8B	−0.078	9.74e-02	−0.066	1.61e-01	0.086	6.62e-02	0.086	6.65e-02
T cell (general)	CD3D	−0.041	3.85e-01	−0.088	5.90e-02	0.047	3.18e-01	0.08	8.59e-02
	CD3E	−0.029	5.32e-01	−0.05	2.85e-01	0.047	3.11e-01	0.087	6.26e-02
B cell	CD2	−0.015	7.48e-01	−0.066	1.58e-01	0.102	*	0.107	*
	CD19	−0.007	8.85e-01	−0.117	*	0.021	6.57e-01	0.155	***
	CD79A	−0.094	*	−0.111	*	0	9.98e-01	0.13	**
Monocyte	CD86	0.072	1.22e-01	−0.001	9.85e-01	0.062	1.85e-01	0.151	**
	CD115	0.175	***	0.102	*	0.08	8.57e-02	0.123	**
TAM	CCL2	0.128	**	0.009	8.52e-01	−0.021	6.57e-01	0.11	*
	CD68	−0.015	7.50e-01	−0.041	3.77e-01	−0.147	**	−0.054	2.49e-01
	IL10	0.063	1.77e-01	−0.016	7.27e-01	0.002	9.63e-01	0.056	2.31e-01
M1 Macrophage	iNOS	−0.205	***	0.012	8.01e-01	0.044	3.43e-01	−0.006	8.92e-01
	IRF5	0.073	1.21e-01	0.033	4.86e-01	−0.076	1.07e-01	−0.006	8.94e-01
	COX2	0.215	***	0.043	3.57e-01	−0.006	8.94e-01	0.127	**
M2 Macrophage	CD163	0.124	**	0.045	3.40e-01	0.047	3.17e-01	0.148	**
	VSIG4	0.063	1.77e-01	0.073	1.20e-01	0.05	2.84e-01	0.106	*
	MS4A4A	0.091	5.30e-02	−0.007	8.74e-01	0.044	3.43e-01	0.139	**
Neutrophils	CD66b	0.049	2.94e-01	−0.062	1.88e-01	−0.116	*	0	9.94e-01
	CD11b	0.104	*	0.062	1.84e-01	0.085	7.00e-02	0.102	*
	CCR7	0.042	3.73e-01	−0.076	1.07e-01	0.06	2.01e-01	0.1	*
Dendritic cell	HLA-DPB1	0.036	4.42e-01	0.113	*	0.107	*	0.052	2.65e-01
	HLA-DQB1	0.05	2.91e-01	0.039	4.11e-01	0.066	1.58e-01	0.068	1.46e-01
	HLA-DRA	0.046	3.26e-01	0.055	2.42e-01	0.115	*	0.084	7.19e-02
	HLA-DPA1	0.027	5.67e-01	0.1	*	0.119	*	0.064	1.69e-01
	BDCA-1	0.214	***	−0.01	8.26e-01	0.084	7.32e-02	0.162	***
	BDCA-4	0.434	***	0.079	9.06e-02	0.053	2.59e-01	0.182	***
	CD11c	0.011	8.17e-01	−0.029	5.41e-01	−0.059	208e-01	0.067	1.54e-01
Th1	T-bet	−0.017	7.22e-01	−0.06	2.02e-01	0.066	1.56e-01	0.121	**
	STAT4	−0.114	*	0.007	8.83e-01	0.181	***	0.172	***
	STAT1	−0.01	8.25e-01	0.043	3.56e-01	0.242	***	0.107	*
	IFN- γ	−0.068	1.45e-01	−0.102	*	0.107	*	0.084	7.22e-02
Th2	TNF- α	0.106	*	0.026	5.73e-01	0.046	3.25e-01	0.026	5.85e-01
	GATA3	0.327	***	0.089	5.76e-02	0.015	7.42e-01	0.144	**
	STAT6	0.141	**	0.228	8.32e-07	0.337	***	0.069	1.40e-01
	STAT5A	−0.045	3.33e-01	0.159	***	0.326	***	0.103	**
Tfh	IL13	0.049	2.93e-01	−0.052	2.68e-01	−0.018	7.06e-01	0.007	8.80e-01
	BCL6	0.255	***	0.158	***	0.023	6.27e-01	0.158	***
	IL21	0.08	8.66e-02	0.077	1.01e-01	0.057	2.27e-01	0.179	***
Th17	STAT3	0.282	***	0.331	***	0.29	***	0.195	***
Treg	IL17A	0.034	4.63e-01	0.079	9.02e-02	0.003	9.50e-01	−0.037	4.28e-01
	FOXP3	0.072	1.22e-01	−0.013	7.76e-01	−0.004	9.28e-01	0.044	3.44e-01
	CCR8	0.138	**	−0.039	4.09e-01	0.145	**	0.133	**
T cell exhaustion	STAT5B	0.174	***	0.235	***	0.331	***	0.174	***
	TGF β	0.276	***	0.177	***	−0.1	*	0.048	3.02e-01
	PD-1	−0.099	*	−0.048	3.01e-01	0.065	1.66e-01	0.063	1.78e-01
	CTLA4	−0.086	6.49e-02	−0.171	***	−0.076	1.05e-01	0.071	1.29e-01
	LAG3	−0.129	**	−0.034	4.65e-01	0.058	2.19e-01	0.053	2.59e-01
	TIM-3	0.036	4.42e-01	0	9.94e-01	0.069	1.41e-01	0.118	**
	GZMB	−0.097	*	−0.132	**	−0.011	8.17e-01	0.016	7.31e-01

TAM, tumor-associated macrophage; Th, T helper cell; Tfh, Follicular helper T cell; Treg, regulatory T cell; Cor, R-value of Spearman's correlation; Purity, correlation adjusted by purity. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

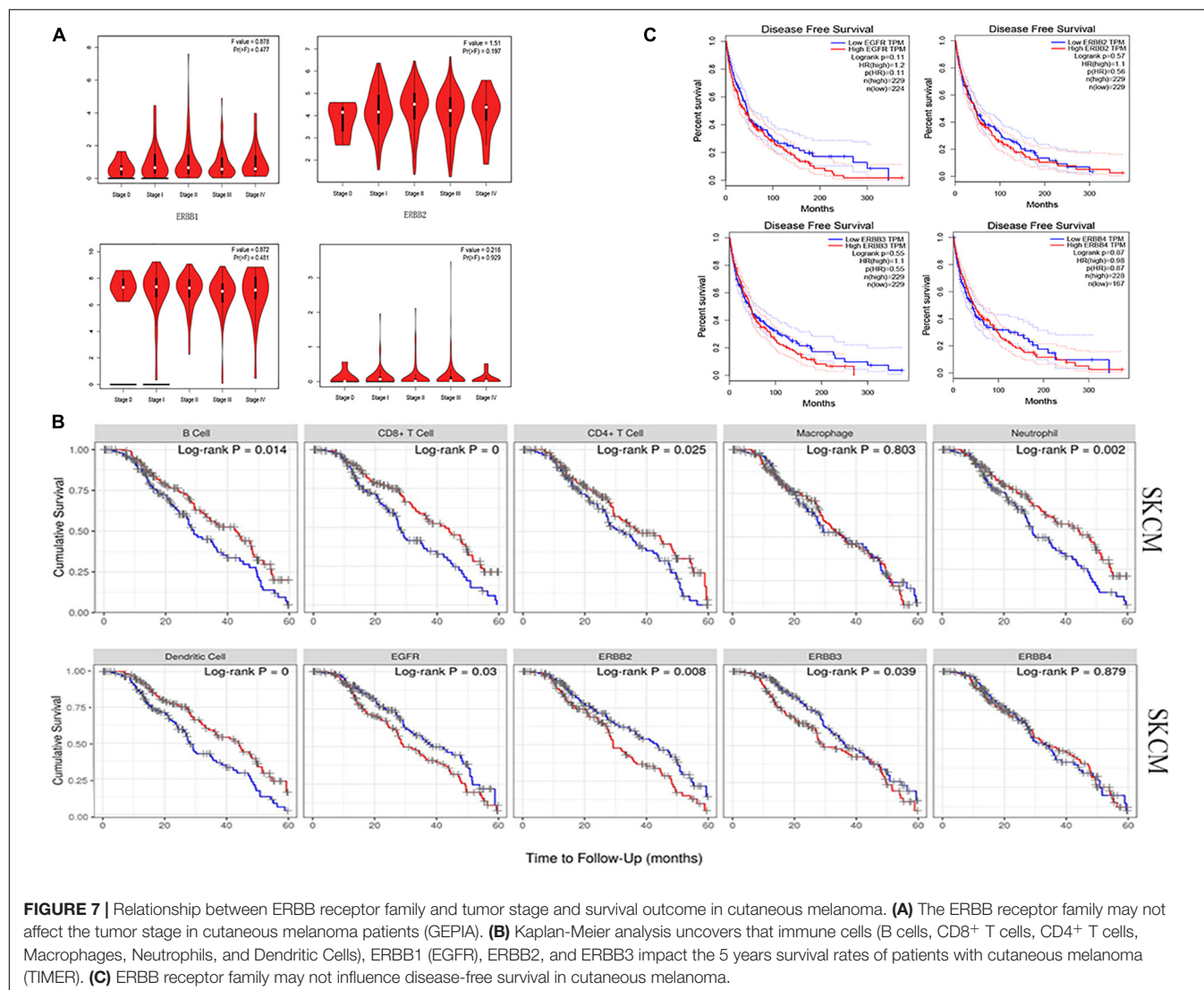


FIGURE 7 | Relationship between ERBB receptor family and tumor stage and survival outcome in cutaneous melanoma. **(A)** The ERBB receptor family may not affect the tumor stage in cutaneous melanoma patients (GEPIC). **(B)** Kaplan-Meier analysis uncovers that immune cells (B cells, CD8+ T cells, CD4+ T cells, Macrophages, Neutrophils, and Dendritic Cells), ERBB1 (EGFR), ERBB2, and ERBB3 impact the 5 years survival rates of patients with cutaneous melanoma (TIMER). **(C)** ERBB receptor family may not influence disease-free survival in cutaneous melanoma.

TABLE 3 | ERBB2/3 is associated with myeloid-derived suppressor cells (MDSC) in cutaneous melanoma.

		SKCM							
Description	Gene markers	ERBB1		ERBB2		ERBB3		ERBB4	
		Cor	P	Cor	P	Cor	P	Cor	P
Human M-MDSC	CD11b	0.038	0.41	0.068	0.14	-0.093	*	0.061	0.19
	CD14	0.026	0.57	0.099	*	-0.19	***	0.023	0.61
Human PMN-MDSC	CD66b	0.032	0.49	-0.032	0.49	-0.026	0.58	-0.023	0.62
	CD33	-0.0095	0.84	0.072	0.12	-0.013	0.77	0.076	0.1
	CD62L	-0.0065	0.89	-0.14	**	-0.19	***	-0.0089	0.85
	CD54	-0.077	0.1	-0.022	0.64	-0.054	0.25	0.15	**
	CD63	-0.13	**	-0.12	**	0.26	***	-0.075	0.11
	CD274	0.024	0.6	-0.035	0.46	-0.12	*	-0.0075	0.87
	ARG1	0.14	**	0.28	***	-0.019	0.68	-0.03	0.52
	CXCR2	0.21	***	0.28	***	-0.15	**	-0.051	0.28
	CXCR4	0.015	0.75	-0.083	0.076	-0.19	***	0.038	0.41

*P<0.05, **P<0.01, ***P<0.001.

Expression of ERBBs With Tumor Stage and Survival Outcome in Cutaneous Melanoma

We analyzed the relationship between ERBBs expression and tumor stage and disease-free survival, whereas ERBBs may not affect tumor stage and disease-free survival in cutaneous melanoma (Figures 7A,C). Also, we found that ERBB1/2/3 and immune infiltration cells (B cells, CD4⁺ T cells of CD8⁺ T cells, macrophages, neutrophils, and dendritic cells) influenced the 5 years survival of patients with cutaneous melanoma, while ERBB4 may not affect the 5-year survival of cutaneous melanoma patients (Figure 7B). These results show that the expression of ERBB1/2/3 in early cutaneous melanoma have prognostic significance.

ERBB2/3 Are Associated With Myeloid-Derived Suppressor Cells (MDSC) in Cutaneous Melanoma

In the GEPIA database, we found that ERBB2 was positively correlated with Human monocytic myeloid-derived suppressor cell(s) (M-MDSC) markers CD14 (Table 3), while ERBB3 was negatively related to Human M-MDSC markers CD11b and CD14. ERBB1 was positively correlated with Human PMN-MDSC markers ARG1, CXCR2, and negatively relevant to CD63. ERBB2 was negatively connected with Human PMN-MDSC markers CD62L and CD63 and positively correlated with ARG1 and CXCR2. ERBB3 was negatively correlated with Human PMN-MDSC markers, CD62L, CD274, CXCR2, and CXCR4 and positively correlated with CD63. ERBB4 was positively associated with Human PMN-MDSC marker CD54. These results suggest that ERBB2/3 may be closely related to MDSC (Human M-MDSC and Human PMN-MDSC) in cutaneous melanoma. In terms of specific mechanisms, further research is needed in the future.

DISCUSSION

Previous studies have shown (Alaoui-Jamali et al., 2015; Elster et al., 2015) that ERBB family members can encode a type I transmembrane protein with common structural properties and be activated by homo- or hetero-dimerization with ERBB family members. For example, when ERBB3 binds to its specific ligand NRG1, it can form a heterodimer with other ERBB partners. Ng et al. (2014) found synergistic activation of ERBB2 and ERBB3 as well as synergistic activation of ERBB3 and ERBB4 in cutaneous melanoma cell lines after NRG1 stimulation. This dimerization activates PI3K/AKT and MAPK/ERK signal transduction pathways. It has been widely recognized that ERBB family members participate in tumorigenesis and tumor therapeutic resistance by activating PI3K and MAPK signaling pathways (Elster et al., 2015). ERBB3 is often highly expressed in primary melanoma and metastatic tumors, and high levels of ERBB2 and ERBB3 are often detected in BRAF WT and mutant cells. Other studies have shown that ERBB4 detected in human melanoma cells (Gordon-Thomson et al., 2005) is mainly a truncated type (120 kD) rather than a full-length protein. So far,

ERBB4 has not been widely studied in melanoma. Although the significance of ERBBs has been confirmed, it is still necessary to clarify the function of ERBBs in cutaneous melanoma.

ERBB1 expression is upregulated in many cancers, but gene expression is inconsistent in cutaneous melanoma. Some studies have stated that ERBB1 over-expression often occurs in advanced stages of melanoma. The expression of ERBB1 in cutaneous melanoma tissues was lower than that in normal tissues. However, in immunohistochemistry, ERBB1 showed moderate expression. ERBB1 autophosphorylation is the key to the PI3K/AKT and MAPK pathways. ERBB1 participates in regulating cell proliferation, apoptosis, and promoting cell invasion through these cascade reactions. We found that ERBB1 is the low expression and may not be associated with tumor stage and disease-free survival but is with 5 years survival.

ERBB2 is a HER2-receptor tyrosine kinase that can cause uncontrolled cell proliferation and tumorigenesis through various mechanisms (Hynes and Lane, 2005). ERBB2 is a vital cancer marker, and no ligand for ERBB2 has been found yet. ERBB2 forms a heterodimer with another ERBB family member to form a more stable and strong signaling function and is considered to be an important therapeutic target for cancer (Giroux, 2013; Milik et al., 2017; Khanjani et al., 2018). ERBB2 is associated with poor clinical prognosis, even though its expression is low (Gilcrease et al., 2009). In our study, it was found that ERBB2 is the low expression in cutaneous melanoma tissues, but this expression may not affect tumor stage and disease-free survival; however, it is related to the 5 years survival rates of cutaneous melanoma patients.

ERBB3 plays a meaningful role in cell proliferation and survival (Yarden and Pines, 2012). Because ERBB3 lacks tyrosine kinase activity, which cannot initiate a signal cascade through autophosphorylation, it must heterodimerize with ERBB1 or ERBB2 to phosphorylate tyrosine in the C-terminal domain of ERBB3 (Campbell et al., 2010). Ueno et al. (2008) found that ERBB3 and ERBB1 dimerized in melanoma cells and promoted the metastasis of melanoma to a certain extent. ERBB3 is overexpressed and activated in a variety of cancers. The regulation of ERBB3 expression and signaling involves many HER3 interacting proteins, including PI3K, Shc, and E3 ubiquitin ligases NEDD4 and Nrdp1 (Mujoo et al., 2014). In our report, we confirmed that ERBB3 exhibits high expression and influences the 5 years survival rates of cutaneous melanoma patients but may not play role in the tumor stage and disease-free survival.

ERBB4 has been shown to have a crucial role in the normal growth of the central nervous system, breast, and fetus (Chuu et al., 2008; Liu et al., 2010; Iwakura and Nawa, 2013). However, the role of ERBB4 in human cancer is still debatable. ERBB4 tyrosine kinase activates through ligand-bound dimerization and induces activation of mitogen-activated protein kinases (MAPKs) and phosphatidylinositol kinase (PI3K)/AKT pathway (Prickett et al., 2009). Prickett et al. (2009) reported that somatic mutations of ERBB4 in malignant melanoma are widespread and proved that ERBB4 mutations are new drug targets for the treatment of metastatic melanoma. However, Zhou et al. (2013) and others reported that ERBB4 hotspot mutations were not detected in melanoma patients in southern China, suggesting that in Chinese

melanoma patients, ERBB4 mutations can only play a limited role. These conclusions prove that there may be geographical differences in mutations of susceptibility genes in melanoma (Casula et al., 2009). In this chapter, we found that ERBB4 is the highest mutation rate in the ERBB family, but it may not impact the tumor stage, disease-free survival, and 5 years survival rates of melanoma patients.

We analyzed the relationship between the ERBB family and 31 genes that are close and frequently altered. In melanoma, KIT, ALK, NTRK1, FGFR4, and MET may not be regulated by the ERBB family, while the genes regulated by the ERBB family are mainly concentrated in the RAS-RAF-MEK-ERK signaling pathway (Wee and Wang, 2017), which is one of the core pathways in the pathogenesis of melanoma (McCubrey et al., 2007) and also a pharmacological target for cancer treatment. CBL, as a ubiquitin ligase, can ubiquitinate ERBB1 to activate downstream signaling pathways (Levkowitz et al., 1998; Grøvdal et al., 2004). The heterodimer combination of ERBB2 and ERBB1 or ERBB3 has strong signal activity (Pinkas-Kramarski et al., 1996), which activates downstream Ras/Raf/MEK/ERK cascade reaction pathway under the stimulation of external factors (Djerf Severinsson et al., 2011). Among them, RAS family genes include HRAS, NRAS, and KRAS; RAF family genes include ARAF, BRAF, and RAF1. ERK signaling pathway includes MAPK1, MAPK2P1, and MAPK2P2. Besides, ERBB1/IGF-1R/CRAF can reduce the proliferation of melanoma cells by inhibiting MAPK and/or PI3K/AKT signaling pathways (Sun et al., 2016). In cutaneous melanoma, the direct association of the ERBB family with other genes (SOS, NF1, PDGFRA, FGFR1, RASA1, ROS1, FGFR3, RAC1, PTPN11, NTRK2, FLT3, FGFR2, ERRFI1, and RIT1) had not been reported, and so the relationship between ERBB family and these genes remains to be further studied and determined.

In some studies, the presence of tumor-infiltrating lymphocytes (TIL) in melanoma is associated with better prognosis and has been interpreted as an indicator that the host promotes a more effective immune response to the tumor (Clemente et al., 1996; Anichini et al., 2012). However, the significance of TIL remains to be further demonstrated. Gooden et al. (2011) conducted a meta-analysis on the impact of TIL on cancer prognosis in 2011 and found that the presence of CD3⁺ and CD8⁺ cells had a beneficial impact on the survival of patients. In our study, ERBB2 can induce infiltration of CD8⁺ T cells and B cells, while ERBB3 can induce infiltration of CD4⁺ T cells, neutrophils cells, and CD8⁺ T cells. CD8⁺ T cells combine antigens and MHC I molecules to form complexes. Once CD8⁺ T cells are fully activated, they can induce the apoptosis of melanoma cells by releasing perforin and granules (Giavina-Bianchi et al., 2017; Zhang et al., 2017). Forsthuber et al. (2019) found that neutrophils play an environmentally dependent role in melanoma in a mouse tumor transplantation model and can be actively switched to an anti-tumor model. CD4⁺ T cells with MHC class II molecules present antigens; the combination of cytokines in the microenvironment under the action of cells can differentiate into various kinds of effects and can be turned into CD4⁺ T helper cells (Th) to activate CD8⁺ T

cells, B cells, and natural killer cells (NK) to tumor cells to play an antitumor immune response (Giavina-Bianchi et al., 2017).

We investigated the relationship between ERBB and immune-infiltrating cell markers (Cor > 0.2). The downstream of EGFR, the MAPK pathway, stimulates the activation of NF-KappaB heteromorphs and homodimers to drive the expression of iNOS, thus supporting the occurrence of melanoma. COX2, as the downstream of the NF-kappa B pathway, plays a pro-oncogenic role in the NF-kB-iNOS-COX-2 signaling pathway (Uffort et al., 2009). Neuropilin-1 (NRP1, BDCA-4) induces a c-Jun N-terminal kinase (JNK)-dependent signaling cascade that leads to the upregulation of EGFR or IGF1R, thereby promoting cancer cell growth (Rizzolio et al., 2020). Studies have shown that the TGF- β pathway can initiate EGFR expression (Sun et al., 2014), making EGFR fully pathogenic. Epidermal growth factor binds to receptors and triggers a variety of signal transduction pathways, one of which activates signal transduction and transcriptional activator (STAT) (Olayioye et al., 1999). From the table we can see that ERBBs are closely related to STAT1, STAT3, STAT5A, STAT5B, and STAT6 (Cor > 0.2). ERBB1 induced low expression of STAT3, STAT5B, and STAT6. ERBB2 induced low expression of STAT3, STAT5A, and STAT5B. ERBB3 induced high expression of STAT1, STAT3, STAT5A, and STAT6. ERBB1/2 can induce low expression of STAT3 and STAT5B may be related to the formation of heterodimer combination. In malignant melanoma, Insulin-like growth factor binding protein 2 (IGFBP2) regulates the expression of PD-L1 by activating the EGFR-STAT3 signaling pathway (Li et al., 2020). The irreversible inhibition of Canertinib on ERBB1-3 was more effective in inhibiting Akt, ERK1/2, and STAT3 signaling pathways (Djerf Severinsson et al., 2011). Phosphorylated STAT5 is regulated by rEGF in melanoma, and inhibition of STAT5B expression can significantly reduce the expression of BCL-2, resulting in decreased cell survival rate and increased apoptosis (Mirmohammadsadegh et al., 2006). Studies have shown that high STAT1, STAT3, and STAT5B expression and low STAT6 expression are associated with better prognosis in SKCM patients. These studies suggest that ERBBs may be closely related to the STAT signaling pathway in cutaneous melanoma. Also, the direct relationship between ERBBs and immune-infiltrating cell markers (COR > 0.2) BDCA-1, GATA3, and BCL6 has not been supported in the literature and requires further study. In conclusion, our study shows that ERBBs and M1 Macrophage, Dendritic cell, Th1, Th2, Th17, and Treg cells have a more significant relationship (Cor > 0.2) in cutaneous melanoma. Together, these findings uncover that ERBBs may play an essential role in recruitment and supervision.

MDSCs are the heterogeneous population of immature bone marrow cells derived from bone marrow (Dumitru et al., 2012), which are composed of bone marrow progenitors, immature macrophages, immature granulocytes, and immature dendritic cells. MDSCs play a strong immunosuppressive role (Goh et al., 2013; Ostrand-Rosenberg et al., 2017) through their involvement in infection, inflammation, and cancer and have a significant ability to inhibit the T cell response (Gabrilovich and Nagaraj, 2009; Scapini et al., 2016). In addition to inhibiting the adaptive immune response, MDSCs also regulate the innate immune response by regulating the

production of cytokines in macrophages. MDSC also has non-immune functions that promote tumor growth and metastasis by paracrine stimulation of tumor cell proliferation, movement, and angiogenesis (Veglia et al., 2018). Macrophages and myeloid-derived suppressor cells (MDSC), further subdivided into monocytic MDSC (M-MDSC) and polymorphonuclear MDSC (PMN-MDSC) (Cassetta et al., 2019). Human M-MDSC is present in the same density fraction as monocytes but differs from monocytes by the low presence or absence of HLA-DR expression. They are further characterized as lymphocyte lineage marker negative cells with the following phenotype $CD11b^+HLA^-DR^-CD14^+CD15^-$. It is possible to use a CD33 myeloid cell marker instead of a CD11b. Human PMN-MDSC are typically described as $CD66b^+CD15^+CD14^{-/dim}CD33^{dim}HLA-DR^-$ cells. CD66b or CD15 can be used as lineage markers. PMN-MDSC have been shown to also express other markers, including chemokine markers (e.g., CXCR2, CXCR4), activation markers (e.g., Markers including CD274/ PD-L1, CD54/ICAM-1, CD62L, CD63), and functional markers [e.g., arginase 1(ARG1)], at variable levels depending on the disease type and severity (Dumitru et al., 2012; Scapini et al., 2016; Cassetta et al., 2019). MDSCs are associated with a poor prognosis of human melanoma (Meyer et al., 2014). In this study, we found that ERBB2/3 may be closely related to MDSC (Human M-MDSC and Human PMN-MDSC) in cutaneous melanoma. Further experimental studies are needed to support this conclusion, which may be studied by our laboratory in appropriate circumstances, due to the lack of evidence in literature for the direct effects of ERBBs and MDSCs markers in cutaneous melanoma.

In summary, we systematically analyzed ERBBs expression, prognosis, immune infiltration, and its relationship with MDSC. In cutaneous melanoma, ERBB3 high expression and ERBB1/2 low expression were strongly associated with 5 years survival rates of cutaneous melanoma patients but may not affect

tumor stage or disease-free survival. ERBB1/2/3 are associated with infiltration of multiple immune cells, especially the M1 Macrophage, Dendritic cell, Th1, Th2, Th17, and Treg cells, which suggests that ERBBs may affect survival rate in cutaneous melanoma patients by affecting immune cell infiltration. Besides, ERBB2/3 are closely related to MDSC, but the role of ERBB2/3 in the cutaneous melanoma population with MDSC remains to be further studied. According to current studies, ERBB1/2/3 may serve as potential therapeutic targets in cutaneous melanoma.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YC, RG, and SL conceived and designed this research. SL and RG executed the analysis procedure, contributed analysis tools, and analyzed the results. EL and PZ contributed to the writing of the manuscript. All authors reviewed the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81901453).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.602160/full#supplementary-material>

REFERENCES

- Alaoui-Jamali, M. A., Morand, G. B., and da Silva, S. D. (2015). ERBB polymorphisms: insights and implications for response to targeted cancer therapeutics. *Front. Genet.* 6:17. doi: 10.3389/fgene.2015.00017
- Anichini, A., Molla, A., Vegetti, C., Bersani, I., Zappasodi, R., Arienti, F., et al. (2012). Tumor-infiltrating lymphocyte grade is an independent predictor of sentinel lymph node status and survival in patients with cutaneous melanoma. *J. Clin. Oncol.* 30, 2678–2683. doi: 10.1200/jco.2011.37.8539
- Bittoni, A., Mandolesi, A., Andrikou, K., Santoni, M., Alfonsi, S., Lanese, A., et al. (2015). HER family receptor expression and prognosis in pancreatic cancer. *Int. J. Biol. Mark.* 30, e327–e332.
- Campbell, M. R., Amin, D., and Moasser, M. M. (2010). HER3 comes of age: new insights into its functions and role in signaling, tumor biology, and cancer therapy. *Clin. Cancer Res.* 16, 1373–1383. doi: 10.1158/1078-0432.ccr-09-1218
- Cassetta, L., Baekkevold, E. S., Brandau, S., Bujko, A., Cassatella, M. A., Dorhoi, A., et al. (2019). Deciphering myeloid-derived suppressor cells: isolation and markers in humans, mice and non-human primates. *Cancer Immunol. Immunother.* 68, 687–697.
- Casula, M., Muggiano, A., Cossu, A., Budroni, M., Caracò, C., Ascierto, P. A., et al. (2009). Role of key-regulator genes in melanoma susceptibility and pathogenesis among patients from South Italy. *BMC Cancer* 9:352. doi: 10.1186/1471-2407-9-352
- Chuu, C. P., Chen, R. Y., Barkinge, J. L., Ciaccio, M. F., and Jones, R. B. (2008). Systems-level analysis of ErbB4 signaling in breast cancer: a laboratory to clinical perspective. *Mol. Cancer Res.* 6, 885–891. doi: 10.1158/1541-7786.mcr-07-0369
- Clemente, C. G., Mihm, M. C. Jr., Bufalino, R., Zurrida, S., Collini, P., and Cascinelli, N. (1996). Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer* 77, 1303–1310. doi: 10.1002/(sici)1097-0142(19960401)77:7<1303::aid-cnrc12>3.0.co;2-5
- Djerf Severinsson, E. A., Trinks, C., Gréen, H., Abdiu, A., Hallbeck, A. L., Stål, O., et al. (2011). The pan-ErbB receptor tyrosine kinase inhibitor canertinib promotes apoptosis of malignant melanoma in vitro and displays anti-tumor activity in vivo. *Biochem. Biophys. Res. Commun.* 414, 563–568. doi: 10.1016/j.bbrc.2011.09.118
- Dumitru, C. A., Moses, K., Trellakis, S., Lang, S., and Brandau, S. (2012). Neutrophils and granulocytic myeloid-derived suppressor cells: immunophenotyping, cell biology and clinical relevance in human oncology. *Cancer Immunol. Immunother.* 61, 1155–1167. doi: 10.1007/s00262-012-1294-5
- Elster, N., Collins, D. M., Toomey, S., Crown, J., Eustace, A. J., and Hennessy, B. T. (2015). HER2-family signalling mechanisms, clinical implications and targeting

- in breast cancer. *Breast Cancer Res. Treat* 149, 5–15. doi: 10.1007/s10549-014-3250-x
- Forsthuber, A., Lipp, K., Andersen, L., Ebersberger, S., Grana, C., Ellmeier, W., et al. (2019). CXCL5 as regulator of neutrophil function in cutaneous melanoma. *J. Invest. Dermatol.* 139, 186–194. doi: 10.1016/j.jid.2018.07.006
- Gabrilovich, D. I., and Nagaraj, S. (2009). Myeloid-derived suppressor cells as regulators of the immune system. *Nat. Rev. Immunol.* 9, 162–174. doi: 10.1038/nri2506
- Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:11.
- Giavina-Bianchi, M. H., Giavina-Bianchi, P. F. Jr., and Neto, C. F. (2017). Melanoma: tumor microenvironment and new treatments. *An. Bras. Dermatol.* 92, 156–166. doi: 10.1590/abd1806-4841.20176183
- Gilcrease, M. Z., Woodward, W. A., Nicolas, M. M., Corley, L. J., Fuller, G. N., Esteva, F. J., et al. (2009). Even low-level HER2 expression may be associated with worse outcome in node-positive breast cancer. *Am. J. Surg. Pathol.* 33, 759–767. doi: 10.1097/pas.0b013e31819437f9
- Giroux, S. (2013). Overcoming acquired resistance to kinase inhibition: the cases of EGFR, ALK, and BRAF. *Bioorgan. Med. Chem. Lett.* 23, 394–401. doi: 10.1016/j.bmcl.2012.11.037
- Goh, C., Narayanan, S., and Hahn, Y. S. (2013). Myeloid-derived suppressor cells: the dark knight or the joker in viral infections? *Immunol. Rev.* 255, 210–221. doi: 10.1111/immr.12084
- Gooden, M. J., de Bock, G. H., Leffers, N., Daemen, T., and Nijman, H. W. (2011). The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis. *Br. J. Cancer* 105, 93–103. doi: 10.1038/bjc.2011.189
- Gordon-Thomson, C., Jones, J., Mason, R. S., and Moore, G. P. M. (2005). ErbB receptors mediate both migratory and proliferative activities in human melanocytes and melanoma cells. *Melanoma Res.* 15, 21–28. doi: 10.1097/00008390-200502000-00005
- Griffin, R., and Ramirez, R. A. (2017). Molecular targets in non-small cell lung cancer. *Ochsner J.* 17, 388–392.
- Grøvdal, L. M., Stang, E., Sorkin, A., and Madhus, I. H. (2004). Direct interaction of Cbl with pTyr 1045 of the EGF receptor (EGFR) is required to sort the EGFR to lysosomes for degradation. *Exp. Cell Res.* 300, 388–395. doi: 10.1016/j.yexcr.2004.07.003
- Homet Moreno, B., and Ribas, A. (2015). Anti-programmed cell death protein-1/ligand-1 therapy in different cancers. *Br. J. Cancer* 112, 1421–1427. doi: 10.1038/bjc.2015.124
- Hynes, N. E., and Lane, H. A. (2005). ERBB receptors and cancer: the complexity of targeted inhibitors. *Nat. Rev. Cancer* 5, 341–354. doi: 10.1038/nrc1609
- Iqbal, N., and Iqbal, N. (2014). Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. *Mol. Biol. Int.* 2014:852748.
- Iwakura, Y., and Nawa, H. (2013). ErbB1-4-dependent EGF/neuregulin signals and their cross talk in the central nervous system: pathological implications in schizophrenia and Parkinson's disease. *Front. Cell. Neurosci.* 7:4. doi: 10.3389/fncel.2013.00004
- Jørgensen, J. T., and Hersom, M. (2012). HER2 as a prognostic marker in gastric cancer – a systematic analysis of data from the literature. *J. Cancer* 3, 137–144. doi: 10.7150/jca.4090
- Khanjani, F., Sajedi, R. H., and Hasannia, S. (2018). Rapid screening of drug candidates against EGFR/HER2 signaling pathway using fluorescence assay. *Anal. Bioanal. Chem.* 410, 7827–7835. doi: 10.1007/s00216-018-1403-1
- Levkowitz, G., Waterman, H., Zamir, E., Kam, Z., Oved, S., Langdon, W. Y., et al. (1998). c-Cbl/Sli-1 regulates endocytic sorting and ubiquitination of the epidermal growth factor receptor. *Genes* 12, 3663–3674. doi: 10.1101/gad.12.23.3663
- Li, T., Fan, J., Wang, B., Traugh, N., Chen, Q., Liu, J. S., et al. (2017). TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* 77, e108–e110.
- Li, T., Zhang, C., Zhao, G., Zhang, X., Hao, M., Hassan, S., et al. (2020). IGFBP2 regulates PD-L1 expression by activating the EGFR-STAT3 signaling pathway in malignant melanoma. *Cancer Lett.* 477, 19–30. doi: 10.1016/j.canlet.2020.02.036
- Liu, W., Purevdorj, E., Zscheppang, K., von Mayersbach, D., Behrens, J., Brinkhaus, M. J., et al. (2010). ErbB4 regulates the timely progression of late fetal lung development. *Biochim. Biophys. Acta* 1803, 832–839. doi: 10.1016/j.bbamcr.2010.03.003
- McCubrey, J. A., Steelman, L. S., Chappell, W. H., Abrams, S. L., Wong, E. W., Chang, F., et al. (2007). Roles of the Raf/MEK/ERK pathway in cell growth, malignant-transformation and drug resistance. *Biochim. Biophys. Acta* 1773, 1263–1284. doi: 10.1016/j.bbamcr.2006.10.001
- Meyer, C., Cagnon, L., Costa-Nunes, C. M., Baumgaertner, P., Montandon, N., Leyvraz, L., et al. (2014). Frequencies of circulating MDSC correlate with clinical outcome of melanoma patients treated with ipilimumab. *Cancer Immunol. Immunother.* 63, 247–257. doi: 10.1007/s00262-013-1508-5
- Milik, S. N., Lasheen, D. S., Serya, R. A. T., and Abouezid, K. A. M. (2017). How to train your inhibitor: design strategies to overcome resistance to epidermal growth factor receptor inhibitors. *Eur. J. Med. Chem.* 142, 131–151. doi: 10.1016/j.ejmech.2017.07.023
- Mirmohammadsadeh, A., Hassan, M., Bardenheuer, W., Marini, A., Gustrau, A., Nambiar, S., et al. (2006). STAT5 phosphorylation in malignant melanoma is important for survival and is mediated through SRC and JAK1 kinases. *J. Invest. Dermatol.* 126, 2272–2280. doi: 10.1038/sj.jid.5700385
- Mujoo, K., Choi, B. K., Huang, Z., Zhang, N., and An, Z. (2014). Regulation of ERBB3/HER3 signaling in cancer. *Oncotarget* 5, 10222–10236. doi: 10.18632/oncotarget.2655
- Navani, S. (2016). Manual evaluation of tissue microarrays in a high-throughput research project: the contribution of Indian surgical pathology to the Human Protein Atlas (HPA) project. *Proteomics* 16, 1266–1270. doi: 10.1002/pmic.201500409
- Ng, Y. K., Lee, J. Y., Supko, K. M., Khan, A., Torres, S. M., Berwick, M., et al. (2014). Pan-erbB inhibition potentiates BRAF inhibitors for melanoma treatment. *Melanoma Res.* 24, 207–218. doi: 10.1097/cmr.0000000000000060
- Olayioye, M. A., Beuvink, I., Horsch, K., Daly, J. M., and Hynes, N. E. (1999). ErbB receptor-induced activation of stat transcription factors is mediated by Src tyrosine kinases. *J. Biol. Chem.* 274, 17209–17218. doi: 10.1074/jbc.274.24.17209
- Ostrand-Rosenberg, S., Sinha, P., Figley, C., Long, R., Park, D., Carter, D., et al. (2017). Frontline Science: myeloid-derived suppressor cells (MDSCs) facilitate maternal-fetal tolerance in mice. *J. Leukoc. Biol.* 101, 1091–1101. doi: 10.1189/jlb.1h1016-306rr
- Oudard, S., Culine, S., Vano, Y., Goldwasser, F., Théodore, C., Nguyen, T., et al. (2015). Multicentre randomized phase II trial of gemcitabine+platinum, with or without trastuzumab, in advanced or metastatic urothelial carcinoma overexpressing Her2. *Eur. J. Cancer* 51, 45–54. doi: 10.1016/j.ejca.2014.10.009
- Pietraszek-Gremplewicz, K., Simiczjyew, A., Dratkiewicz, E., Podgórska, M., Styczeń, I., Matkowski, R., et al. (2019). Expression level of EGFR and MET receptors regulates invasiveness of melanoma cells. *J. Cell. Mol. Med.* 23, 8453–8463. doi: 10.1111/jcmm.14730
- Pinkas-Kramarski, R., Soussan, L., Waterman, H., Levkowitz, G., Alroy, I., Klapper, L., et al. (1996). Diversification of Neu differentiation factor and epidermal growth factor signaling by combinatorial receptor interactions. *EMBO J.* 15, 2452–2467. doi: 10.1002/j.1460-2075.1996.tb00603.x
- Prickett, T. D., Agrawal, N. S., Wei, X., Yates, K. E., Lin, J. C., Wunderlich, J. R., et al. (2009). Analysis of the tyrosine kinase in melanoma reveals recurrent mutations in ERBB4. *Nat. Genet.* 41, 1127–1132. doi: 10.1038/ng.438
- Rhodes, D. R., Kalyana-Sundaram, S., Mahavisno, V., Varambally, R., Yu, J., Briggs, B. B., et al. (2007). Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia* 9, 166–180. doi: 10.1593/neo.07112
- Riker, A. I., Enkemann, S. A., Fodstad, O., Liu, S., Ren, S., Morris, C., et al. (2008). The gene expression profiles of primary and metastatic melanoma yields a transition point of tumor progression and metastasis. *BMC Med. Genomics* 1:13. doi: 10.1186/1755-8794-1-13
- Rizzolio, S., Corso, S., Giordano, S., and Tamagnone, L. (2020). Autocrine signaling of NRP1 ligand galectin-1 elicits resistance to BRAF-targeted therapy in melanoma cells. *Cancers* 12:2218. doi: 10.3390/cancers12082218
- Scapini, P., Marini, O., Tecchio, C., and Cassatella, M. A. (2016). Human neutrophils in the saga of cellular heterogeneity: insights and open questions. *Immunol. Rev.* 273, 48–60. doi: 10.1111/immr.12448

- Song, X., Fan, P. D., Bantikassegn, A., Guha, U., Threadgill, D. W., Varmus, H., et al. (2015). ERBB3-independent activation of the PI3K pathway in EGFR-mutant lung adenocarcinomas. *Cancer Res.* 75, 1035–1045. doi: 10.1158/0008-5472.can-13-1625
- Sun, C., Wang, L., Huang, S., Heynen, G. J., Prahallad, A., Robert, C., et al. (2014). Reversible and adaptive resistance to BRAF (V600E) inhibition in melanoma. *Nature* 508, 118–122. doi: 10.1038/nature13121
- Sun, X., Li, J., Sun, Y., Zhang, Y., Dong, L., Shen, C., et al. (2016). MiR-7 reverses the resistance to BRAFi in melanoma by targeting EGFR/IGF-1R/CRAF and inhibiting the MAPK and PI3K/AKT signaling pathways. *Oncotarget* 7, 53558–53570. doi: 10.18632/oncotarget.10669
- Talantov, D., Mazumder, A., Yu, J. X., Briggs, T., Jiang, Y., Backus, J., et al. (2005). Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin. Cancer Res.* 11, 7234–7242. doi: 10.1158/1078-0432.ccr-05-0683
- Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, w98–w102.
- Ueno, Y., Sakurai, H., Tsunoda, S., Choo, M. K., Matsuo, M., Koizumi, K., et al. (2008). Heregulin-induced activation of ErbB3 by EGFR tyrosine kinase activity promotes tumor growth and metastasis in melanoma cells. *Int. J. Cancer* 123, 340–347. doi: 10.1002/ijc.23465
- Uffort, D. G., Grimm, E. A., and Ellerhorst, J. A. (2009). NF-kappaB mediates mitogen-activated protein kinase pathway-dependent iNOS expression in human melanoma. *J. Invest. Dermatol.* 129, 148–154. doi: 10.1038/jid.2008.205
- Veglia, F., Perego, M., and Gabrilovich, D. (2018). Myeloid-derived suppressor cells coming of age. *Nat. Immunol.* 19, 108–119. doi: 10.1038/s41590-017-0022-x
- Wee, P., and Wang, Z. (2017). Epidermal growth factor receptor cell proliferation signaling Pathways. *Cancers* 9:52. doi: 10.3390/cancers9050052
- Yarden, Y., and Pines, G. (2012). The ERBB network: at last, cancer therapy meets systems biology. *Nat. Rev. Cancer* 12, 553–563. doi: 10.1038/nrc3309
- Zhang, Y., Kurupati, R., Liu, L., Zhou, X. Y., Zhang, G., Hudaihed, A., et al. (2017). Enhancing CD8⁺ T cell fatty acid catabolism within a metabolically challenging tumor microenvironment increases the efficacy of melanoma immunotherapy. *Cancer Cell* 32, 377–391.e9.
- Zhou, Q. M., Li, W., Guan, Y. X., Zhang, X., Chen, X. C., Ding, Y., et al. (2013). The absence of the ERBB4 hotspot mutations in melanomas in patients from southern China. *Chin. J. Cancer* 32, 410–414. doi: 10.5732/cjc.012.10121
- Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A. H., Tanaseichuk, O., et al. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* 10:1523.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Liu, Geng, Lin, Zhao and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Identification and Analysis of Potential Key Genes Associated With Hepatocellular Carcinoma Based on Integrated Bioinformatics Methods

Zhuolin Li¹, Yao Lin², Bizhen Cheng¹, Qiaoxin Zhang¹ and Yingmu Cai^{1*}

¹ Department of Clinical Laboratory, The First Affiliated Hospital of Shantou University Medical College, Shantou, China,

² Department of Plastic Surgery and Burn Center, The Second Affiliated Hospital of Shantou University Medical College, Shantou, China

OPEN ACCESS

Edited by:

Ancha Baranova,
George Mason University,
United States

Reviewed by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia
Evgenii Chekalin,
Michigan State University,
United States

*Correspondence:

Yingmu Cai
st_ymcai@163.com

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 10 June 2020

Accepted: 18 February 2021

Published: 09 March 2021

Citation:

Li Z, Lin Y, Cheng B, Zhang Q and
Cai Y (2021) Identification
and Analysis of Potential Key Genes
Associated With Hepatocellular
Carcinoma Based on Integrated
Bioinformatics Methods.
Front. Genet. 12:571231.
doi: 10.3389/fgene.2021.571231

Background: Hepatocellular carcinoma (HCC) is a type of primary liver tumor with poor prognosis and high mortality, and its molecular mechanism remains incompletely understood. This study aimed to use bioinformatics technology to identify differentially expressed genes (DEGs) in HCC pathogenesis, hoping to identify novel biomarkers or potential therapeutic targets for HCC research.

Methods: The bioinformatics analysis of our research mostly involved the following two datasets: Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). First, we screened DEGs based on the R packages (limma and edgeR). Using the DAVID database, the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of DEGs were carried out. Next, the protein-protein interaction (PPI) network of the DEGs was built in the STRING database. Then, hub genes were screened through the cytoHubba plug-in, followed by verification using the GEPIA and Oncomine databases. We demonstrated differences in levels of the protein in hub genes using the Human Protein Atlas (HPA) database. Finally, the hub genes prognostic values were analyzed by the GEPIA database. Additionally, using the Comparative Toxicogenomics Database (CTD), we constructed the drug-gene interaction network.

Results: We ended up with 763 DEGs, including 247 upregulated and 516 downregulated DEGs, that were mainly enriched in the epoxygenase P450 pathway, oxidation-reduction process, and metabolism-related pathways. Through the constructed PPI network, it can be concluded that the P53 signaling pathway and the cell cycle are the most obvious in module analysis. From the PPI, we filtered out eight hub genes, and these genes were significantly upregulated in HCC samples, findings consistent with the expression validation results. Additionally, survival analysis showed that high level gene expression of CDC20, CDK1, MAD2L1, BUB1, BUB1B, CCNB1, and CCNA2 were connected with the poor overall survival of HCC patients.

Toxicogenomics analysis showed that only topotecan, oxaliplatin, and azathioprine could reduce the gene expression levels of all seven hub genes.

Conclusion: The present study screened out the key genes and pathways that were related to HCC pathogenesis, which could provide new insight for the future molecularly targeted therapy and prognosis evaluation of HCC.

Keywords: hepatocellular carcinoma, bioinformatics, differentially expressed genes, survival, biomarker, GEO, TCGA

INTRODUCTION

Accounting for 75–85% of all primary liver cancer, hepatocellular carcinoma (HCC) is the main histological classification of liver cancer, which is the fourth most frequent cause of cancer-related death globally (Harris et al., 2019; Yang J.D. et al., 2019). The liver is the second most common cancer-prone organ, after the lungs, as was shown by the recent cancer study in China (Fu and Wang, 2018). On the whole, the estimated morbidity of HCC per 100,000 world standard population is 40.0 in males and 15.3 in females (Zhu et al., 2016). Major risk factors for HCC include genetic predisposition, epigenetic variation, chronic hepatitis B infection, hepatitis C virus infection, smoking, obesity, aflatoxin exposure, and diabetes (Puszyk et al., 2013; Baecker et al., 2018). Transplantation is the most useful way to treat HCC; however, after the transplantation process, the tumor recurrence and metastasis rates are high (Au and Chok, 2018; Aufhauser et al., 2018). More than 70% of patients at advanced stage are not suitable for transplantation, whether due to the tumor burden or liver dysfunction (Wang et al., 2019). Therefore, it is urgent to recognize new biomarkers that can act as molecular targets for therapy, and predictors of the prognosis of HCC. With the development of times and technological progress, microarray and high-throughput sequencing technologies have matured and become more reliable, and public databases are improving, such as the Gene Expression Omnibus (GEO)¹ and the Cancer Genome Atlas (TCGA)². The advancement of microarray (Yang X. et al., 2018) and high throughput sequencing technologies (Weinstein et al., 2013) has provided a highly efficient tools to explore key genetic or epigenetic changes in disease to identify biological markers that can be applied to disease diagnosis, therapy, and prognosis (Weinstein et al., 2013; Wang et al., 2018; Yang X. et al., 2018; Li et al., 2019). Additionally, the application of integrated bioinformatics methods in cancer research can solve the problem of different results due to errors caused by different technical platforms or small sample size, thus finding much valuable biological information (Liu X. et al., 2018; Deng et al., 2019; Yan et al., 2019; Yang K. et al., 2019).

In this research, by analyzing and distinguishing genes in human HCC samples and normal hepatocyte samples using TCGA and GEO datasets, we firstly screened out differentially expressed genes (DEGs). Then, GO and KEGG pathway enrichment analyses were applied in the further exploration of

the main biological functions, which regulated by the DEGs. After that, the final step is to utilize a protein–protein interaction (PPI) network, survival analyses and drug–gene interaction network analyses to ascertain crucial genes and pathways which affecting the pathogenic mechanism and prognosis of HCC patients.

MATERIALS AND METHODS

Gene Expression Datasets

The microarray gene expression dataset of GSE121248, which comprises 70 hepatocellular carcinoma samples and 37 normal liver samples, was obtained from the GEO website and exploited as discovery dataset to identify DEGs. The included dataset met the following criteria: (1) dataset included human HCC samples and normal liver samples. (2) they contained at least ten samples. (3) dataset was obtained from the Affymetrix Human Genome U133 Plus 2.0 Array [HG-U133_Plus_2] microarray platform. The raw RNA sequencing data, which comprises 374 HCC samples and 50 normal liver tissue samples, was selected from the TCGA liver hepatocellular carcinoma (TCGA-LIHC) dataset and used as a validation dataset.

Identification of DEGs

We used the R language to analyze the original CEL files of the GSE121248 dataset.

The preprocessing procedures: using the “affy” R package to RMA background correction, Log2conversion, Quantile normalization, and Median polish algorithm summarization (Bolstad et al., 2003; Gautier et al., 2004). Using the bioconductor annotation package to convert microarray data probes into gene symbol. If multiple probes were mapped to a gene symbol, take the average value as the final expression value of the gene (Zhang et al., 2018). Next, $|\log_2\text{fold change (FC)}| > 1$ and adjusted p value < 0.05 were used to select the DEGs between tumor and normal tissues using the LIMMA package (Ritchie et al., 2015; Nagy et al., 2018).

DEGs Validation Using the TCGA Dataset

The DEGs from the GSE121248 dataset were validated using the TCGA-LIHC dataset.

The edgeR package of R software was applied to normalize and analyze the TCGA-LIHC dataset (Robinson et al., 2010). $|\log_2\text{fold change (FC)}| > 1$ and p -value < 0.05 were considered significant differences. The overlapping DEGs between GSE121248 and TCGA-LIHC datasets were clustered using the

¹<https://www.ncbi.nlm.nih.gov/geo/>

²<http://tcga-data.nci.nih.gov>

heatmap and were retained for further study. The overlapping DEGs were analyzed using VennDiagram and ggplot2 packages in R software to draw Venn diagrams and volcano plots, to visualize the identified DEGs (Chen and Boutros, 2011).

Functional Enrichment Analysis of Overlapping DEGs

We used the Database for Annotation, Visualization and Integrated Discovery (DAVID version 6.8)³ to elucidate potential GO function [including biological processes (BP), molecular functions (MF), cellular components (CC)] and signaling pathways (KEGG) related to the overlapping DEGs (Dennis et al., 2003; Kanehisa et al., 2017). We used threshold *p*-value 0.05.

Protein–Protein Interaction Network Construction and Module Analysis

The Search Tool for the Retrieval of Interacting Genes (STRING version 11)⁴ database was one of the largest online databases of known protein-protein interactions covering the largest number of species (Szklarczyk et al., 2017). The parameter of interactions was set with a confidence score >0.7. The confidence score refers to the strength of data support in terms of the thickness of the line. Confidence score >0.7 means high confidence. Overlapping DEGs were entered into Cytoscape software (version 3.7.2)⁵ to construct and analyze PPI network (Shannon et al., 2003). Moreover, the Cytoscape plug-in MCODE was used to screen crucial clustering modules in the entire network (Bader and Hogue, 2003).

Identification of Hub Genes

The Cytoscape plug-in CytoHubba was used to calculate the protein node degree (Chin et al., 2014; Cao et al., 2018). The top three methods [(Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), and Density of Maximum Neighborhood Component (DMNC)] were selected to provide the analyzed results. Each method displayed their top ten genes. A Venn diagram was generated to visualize common hub genes based on these three methods.

Expression Analysis of Hub Genes in Multiple Databases

The hub genes mRNA expression levels were finally validated in two databases, Gene Expression Profiling Interactive Analysis (GEPIA)⁶ (Tang et al., 2017) and Oncomine. Oncomine (Version4.5)⁷ is an online database that has the comprehensive cancer mutation spectrum, gene expression data and related clinical information, which can be used to discover new biomarkers or new therapeutic targets (Rhodes et al., 2004). In addition to detecting the mRNA expression levels of the hub genes, we also investigated the protein levels in HCC tissues and

normal liver tissues using the human protein atlas database (HPA v19)⁸ (Thul and Lindskog, 2018).

Survival Analysis

Gene Expression Profiling Interactive Analysis is a newly developed online database for cancer and normal gene expression profiling. In the current study, the overall survival of each hub gene was analyzed using LIHC dataset in the GEPIA database. The patients were divided into two groups (the high- and low-expression group) according to the median expression level of each hub gene. This division method could evaluate the difference in overall survival probability between these two groups. We were drawn the overall survival curves of each hub gene using the GEPIA database, with a *p*-value <0.05.

Drug-Gene Interaction Network Analysis

The Comparative Toxicogenomics Database (CTD)⁹, an online database providing information on the interactions between gene products and chemotherapeutic drugs, and their relationships to diseases) was used to construct the chemotherapeutic drug-gene interaction network (Davis et al., 2019). The networks were visualized by Cytoscape software 3.7.2¹⁰.

RESULTS

Identification of DEGs

The gene expression dataset of GSE121248, which contains 70 LIHC samples and 37 normal liver samples, was analyzed in the limma package using $|\log FC| > 1$ and corrected *p*-value <0.05 of R software. In total, 1,518 DEGs (557 high expression genes and 961 low expression genes) were identified between HCC tissue samples and normal liver tissue samples. The volcano map and heatmap of all DEGs are shown in **Figures 1A,C**. Additionally, compared with normal liver tissues in the TCGA-LIHC dataset, 2,898 DEGs were obtained in LIHC tissues, comprising 1,299 upregulated genes and 1,599 downregulated genes (**Figure 1B**). Furthermore, 763 overlapping DEGs (247 high expression genes and 516 low expression genes) were identified between the GSE121248 and TCGA-LIHC datasets using a Venn diagram (**Figure 1D**). **Figure 1E** shows clustering analysis results of the 763 overlapping DEGs based on the TCGA-LIHC dataset.

Enrichment Analysis of Overlapping DEGs

We conducted GO and KEGG pathway enrichment analysis to further elucidate potential biological functions associated with the 763 overlapping DEGs of HCC. The GO analysis results of the DEGs were classified into molecular functions, biological processes and cellular components. For molecular functions, the overlapping DEGs were mainly associated with oxidoreductase activity, monooxygenase activity, heme binding and oxygen binding (**Figure 2A**). In the BP category, the epoxide hydrolase P450

³<https://david.ncifcrf.gov/>

⁴<http://string-db.org/>

⁵<http://www.cytoscape.org/>

⁶<http://gepia.cancer-pku.cn/>

⁷<https://www.oncomine.org/>

⁸<https://www.proteinatlas.org/>

⁹<http://ctdbase.org/>

¹⁰<https://cytoscape.org/>

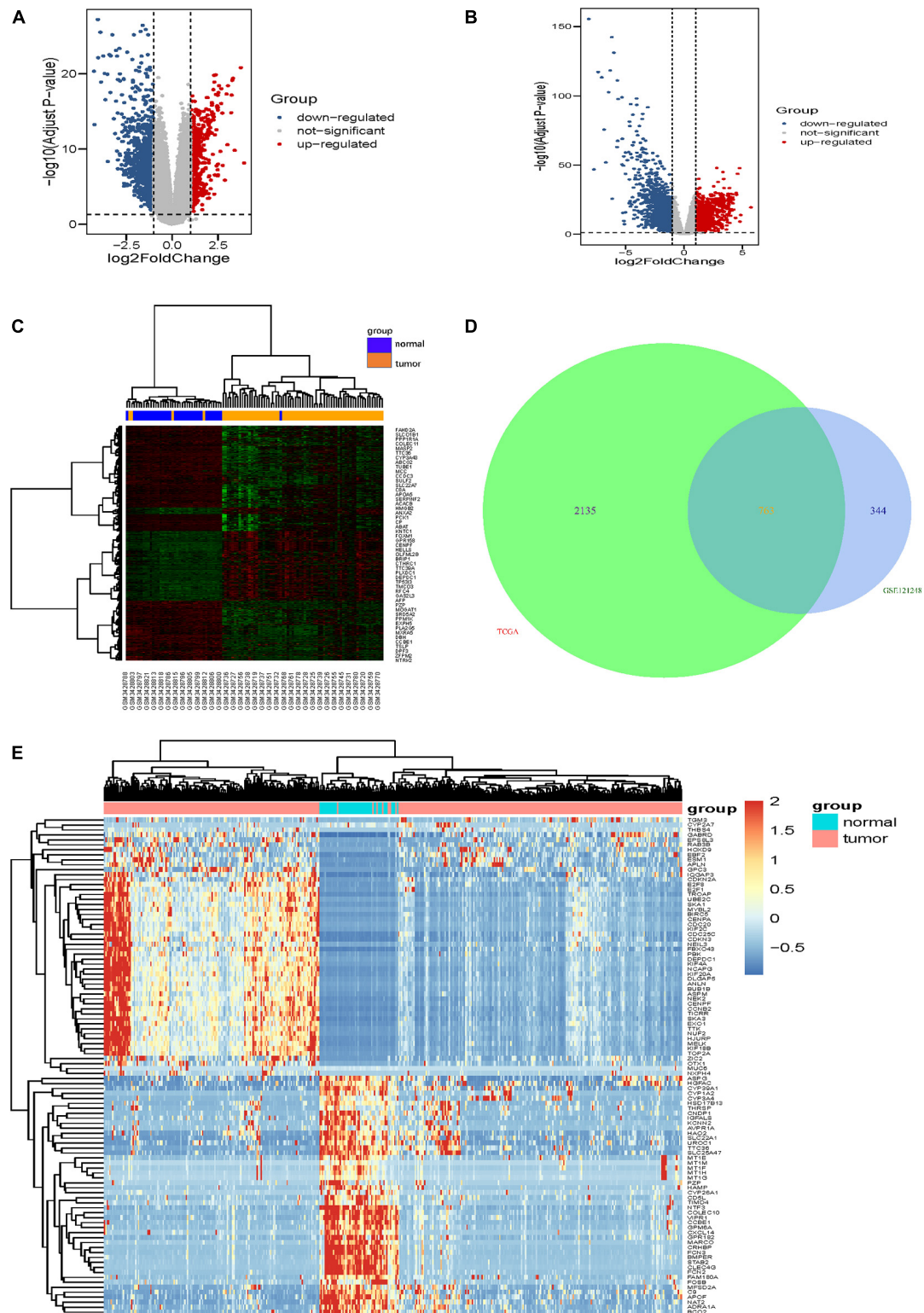
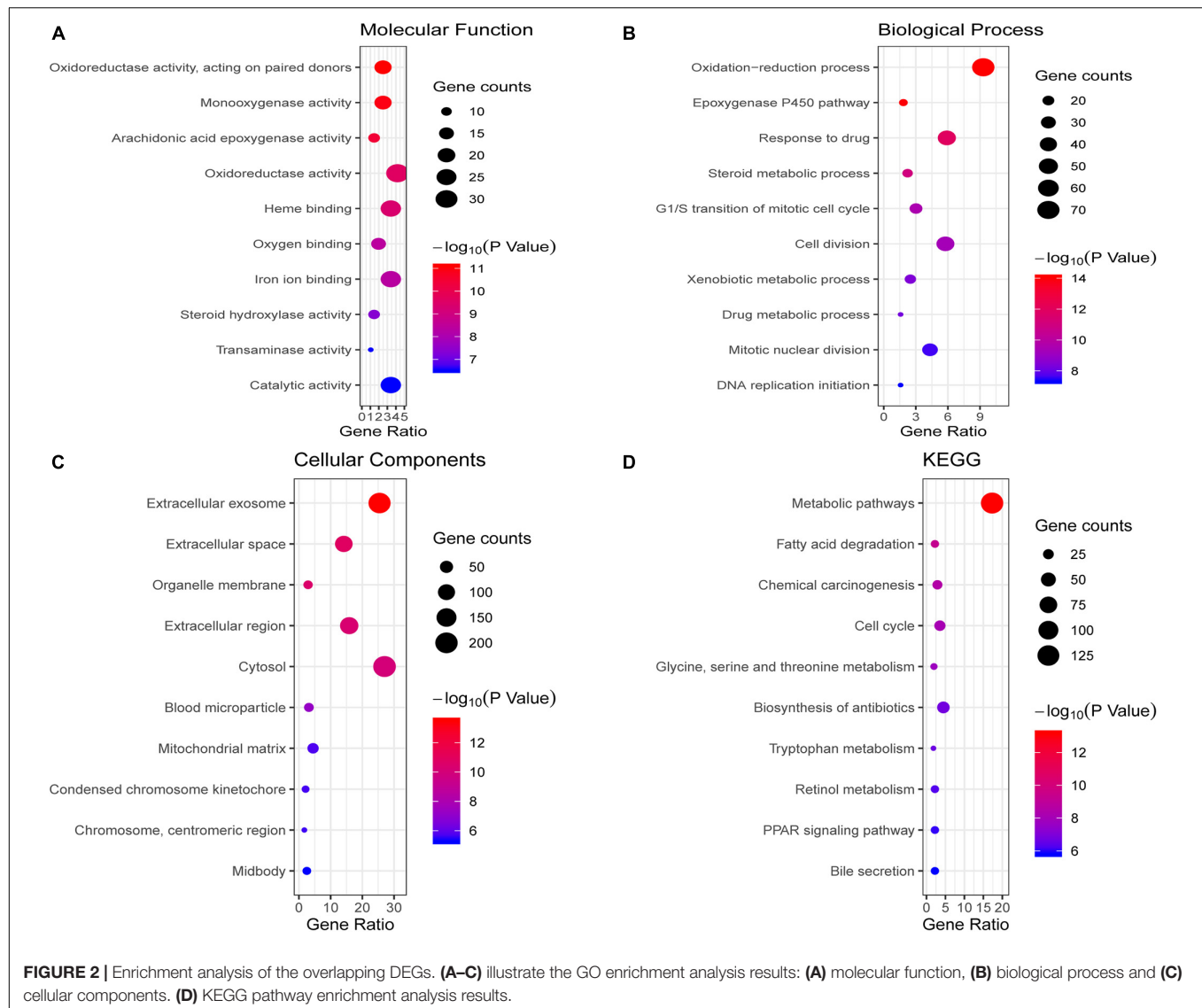


FIGURE 1 | Identification of DEGs. (A,B) show the volcano maps of DEGs for (A) GSE121248 dataset, (B) TCGA-LIHC dataset. (C) The heatmap of the top 50 DEGs in dataset GSE121248. The green color and red color in the heatmap indicate low and high expression of DEGs. (D) Venn diagrams of the DEGs between the GSE121248 dataset and the TCGA-LIHC dataset. (E) The heatmap of the top 100 overlapping DEGs according to the value of $|\log_2\text{FC}|$ in TCGA-LIHC dataset. The color in heatmaps from green to red shows the progression from down-regulation to up-regulation.



pathway, oxidation-reduction process, response to drug and cell division were enriched (Figure 2B). In the CC category, they were enriched in extracellular regions, such as extracellular exosomes and the extracellular space (Figure 2C). The pathway enrichment analysis results showed that overlapping DEGs mainly participated in multiple metabolism pathways, such as fatty acid degradation, glycine, serine and threonine metabolism, and tryptophan metabolism (Figure 2D).

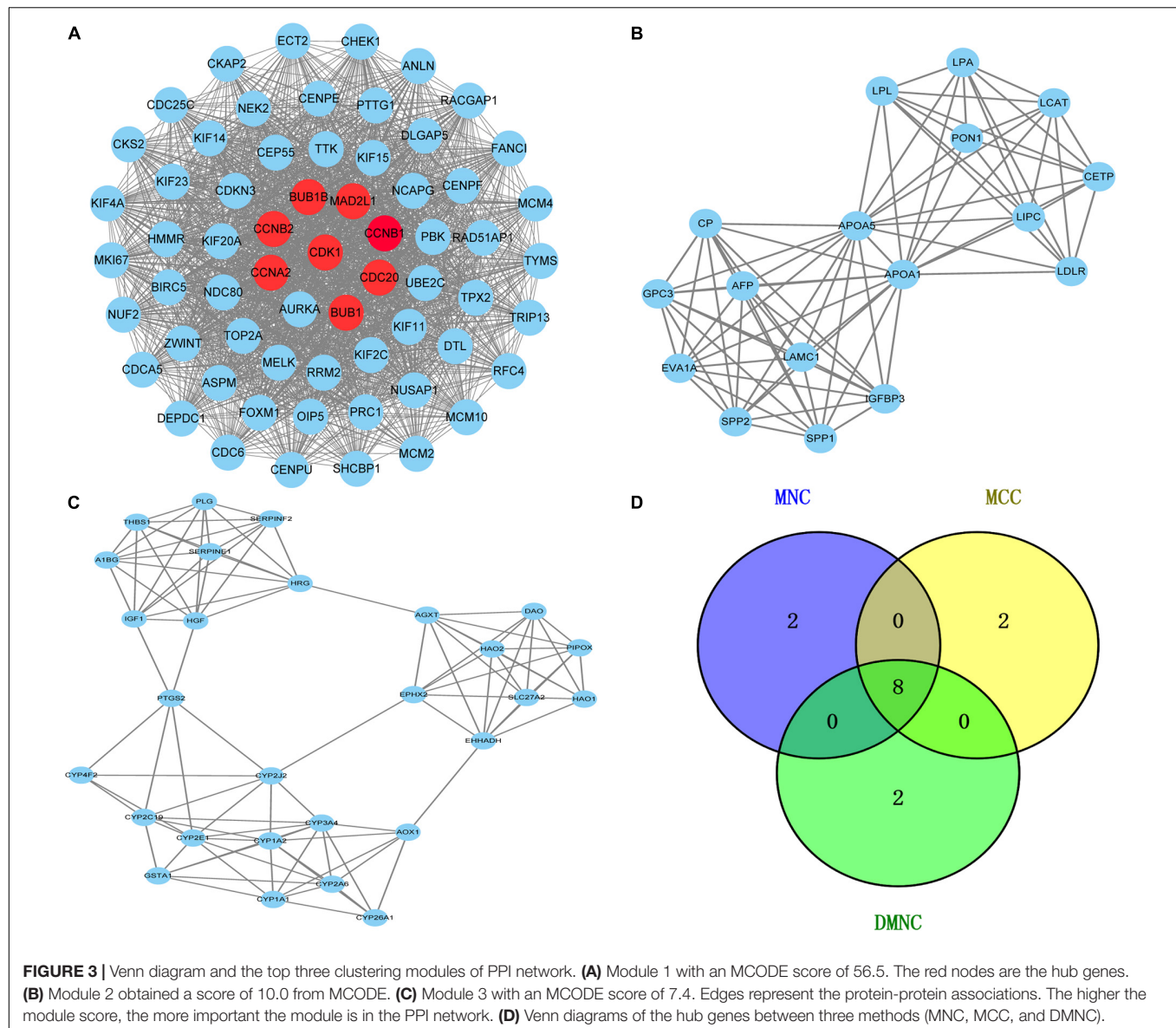
PPI Network Establishment and Module Analysis

To further reveal the potential relationships between proteins encoded by DEGs, a PPI network was constructed using the STRING database. Network analysis of overlapping DEGs revealed 526 nodes and 4,173 edges in the PPI network. Additionally, we conducted module analysis using the MCODE plug-in to detect crucial clustering modules. In total, 29 clusters

were obtained in MCODE, and the top three modules with the highest scores were selected as hub modules. Module 1 contained 63 nodes and 1,752 edges with the highest score of 56.516 and was mainly enriched in cell cycle, oocyte meiosis, P53 signaling pathway and progesterone-mediated oocyte maturation (Figure 3A). Module 2 contained 17 nodes and 80 edges with a score of 10 and mainly participated in PPAR signaling pathway and glycerolipid metabolism (Figure 3B). Module 3 comprised 28 nodes and 100 edges with a score of 7.407 and was mainly implicated in Chemical carcinogenesis, Peroxisome, Metabolic pathways and Drug metabolism cytochrome P450 (Figure 3C).

Hub Genes Selection From the PPI Network

The Cytoscape plug-in cytoHubba including the top three algorithms (MCC, MNC, and DMNC) was applied to select hub genes, and the top 10 genes were selected by each of the three



methods. The Venn diagram identified eight overlapping hub genes based on these three methods (**Figure 3D**): cell division cycle protein 20 homolog (CDC20), cyclin-dependent kinase1 (CDK1), mitotic spindle assembly checkpoint protein MAD2A (MAD2L1), threonine-protein kinase BUB1 (BUB1), threonine-protein kinase BUB1 beta (BUB1B), mitotic-specific cyclin-B1 (CCNB1), mitotic-specific cyclin-B2 (CCNB2) and cyclin-A2 (CCNA2). These eight hub genes were used for further analysis.

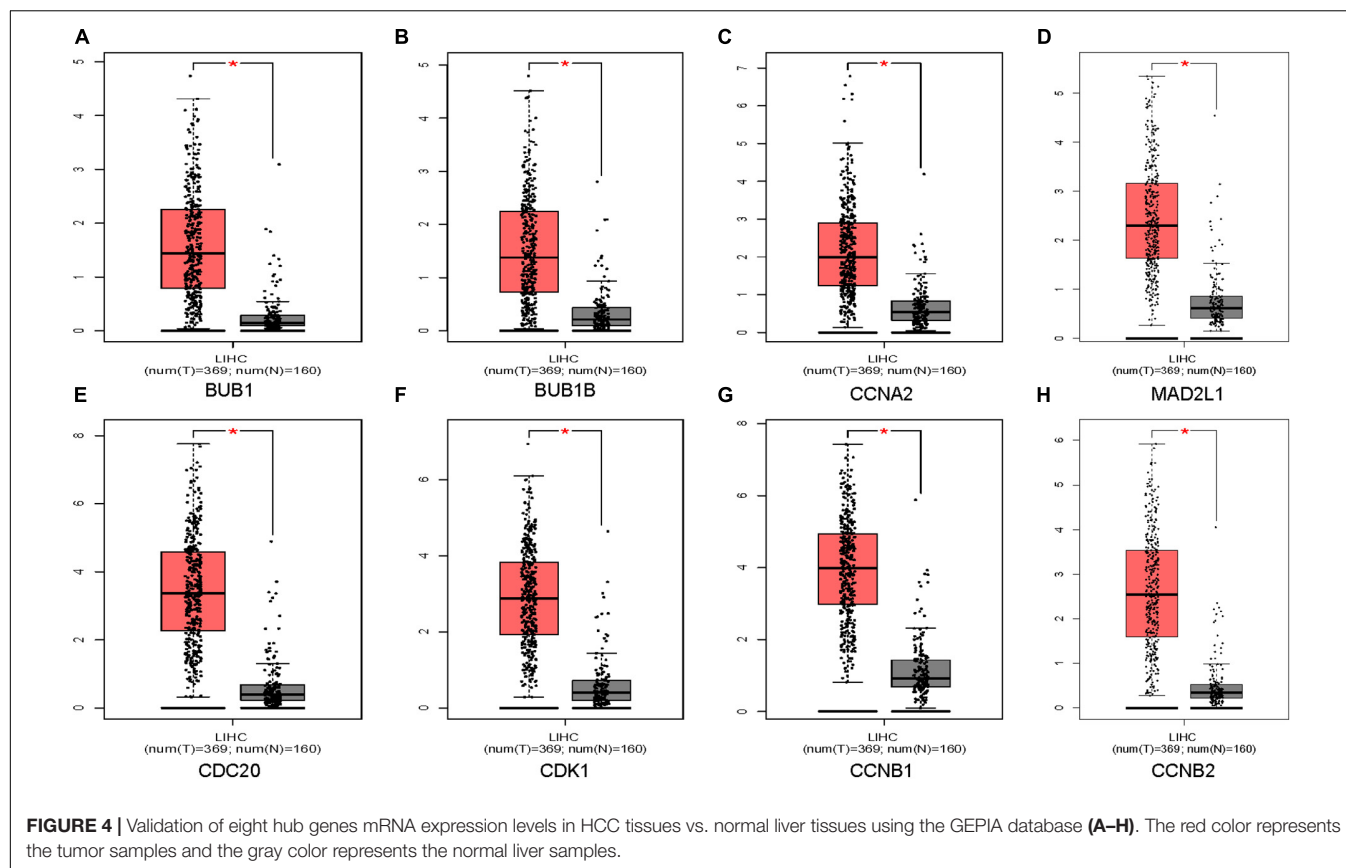
Validation of Hub Genes in Multiple Databases

Oncomine and GEPIA were applied to validate the differentially expression levels of 8 hub genes between HCC tissues and normal liver tissues in HCC. These eight hub genes were all remarkably overexpressed in HCC samples (**Figure 4**). Moreover, a summary of hub genes in multiple tumors

indicated that hub genes were significantly overexpressed in HCC (**Figure 5**). Furthermore, we also investigated the protein expression levels in HCC tissue samples and normal liver tissue samples using the human protein atlas database. Because the HPA dataset could not provide immunohistochemical information on BUB1 and BUB1B, we showed the results of the remaining six staining pairs in **Figure 6**. The protein expression levels of hub genes were agreed with the mRNA expression results, and most genes were overexpressed in HCC tissue (**Figure 7**). These findings indicate that the overexpression of these hub genes may play a critical role in HCC mechanism.

Survival Analysis

We further used the GEPIA database to analyze the prognostic value of these 8 hub genes in HCC patients. The survival analysis



of patients in the GEPIA database was based on the TCGA-LIHC data set. We used threshold p -value 0.05 and calculated the hazards ratio based on Cox PH Model (Xu et al., 2020). The relatively higher expression of CDC20 (HR = 2.3; $P = 3.8 \times 10^{-6}$), CDK1 (HR = 2; $P = 0.00017$), MAD2L1 (HR = 1.7; $P = 0.0047$), BUB1 (HR = 1.8; $P = 0.001$), BUB1B (HR = 1.7; $P = 0.0028$), CCNB1 (HR = 2; $P = 0.00015$), and CCNA2 (HR = 1.7; $P = 0.0037$) were associated with a poor prognosis in HCC patients, while only CCNB2 (HR = 1.4; $P = 0.052$) showed no statistical significance in the overall survival of patients (Figure 6).

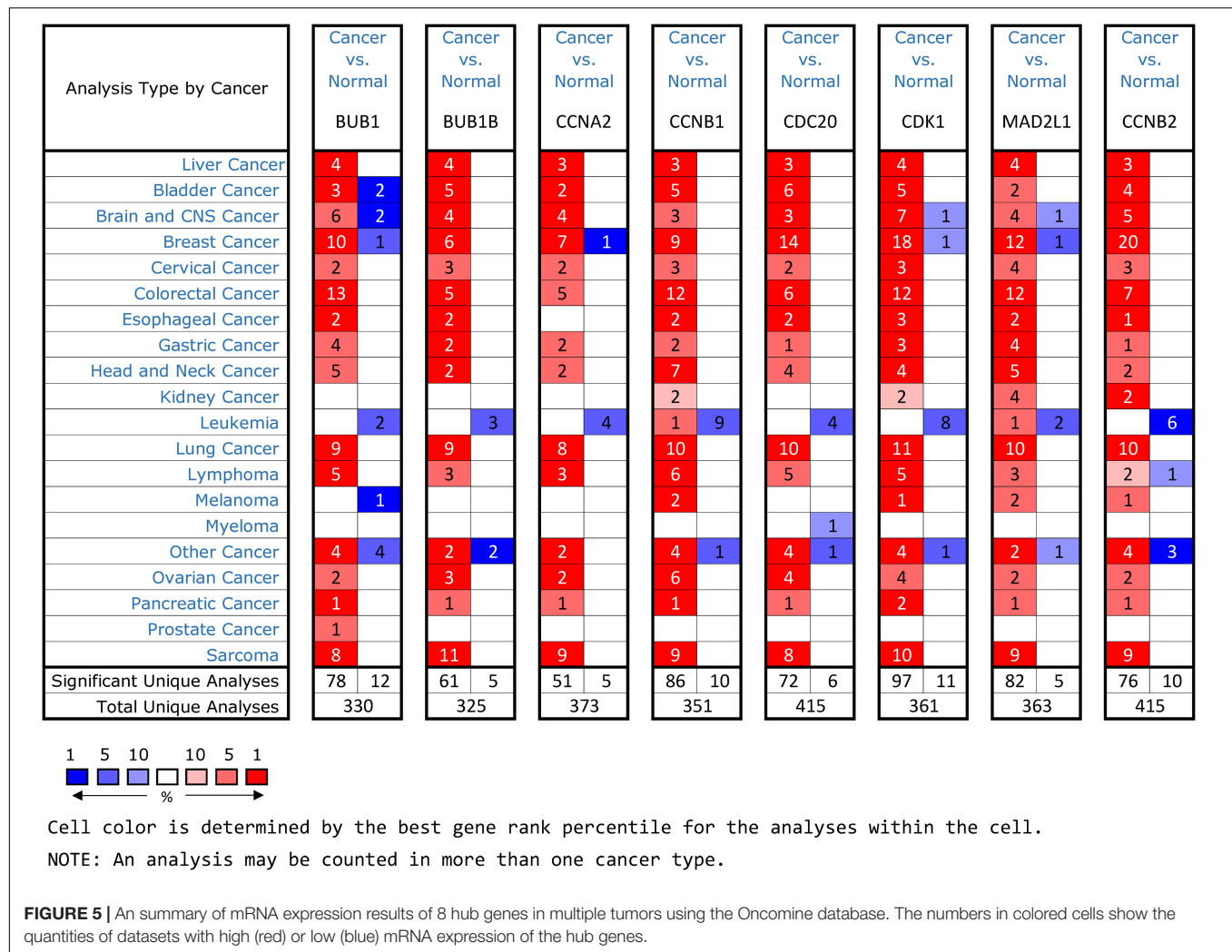
Drug-Gene Interaction Network Analysis

To investigate the potential information on the interactions between hub genes and cancer chemotherapeutics drugs, we used the CTD database to construct chemotherapeutics drug-gene interaction network. Various drugs could influence the mRNA expression level of seven hub genes, namely, CDC20, CDK1, MAD2L1, CCNA2, CCNB1, BUB1, and BUB1B (Figure 8). However, only topotecan, oxaliplatin and azathioprine could reduce expression levels of all seven hub genes.

DISCUSSION

Hepatocellular carcinoma is a type of primary liver tumor with poor prognosis and high mortality, and the progress in

its diagnosis and treatment has always attracted widespread attention from researchers around the world. Because the high recurrence and metastasis rate of HCC remains a challenge, identifying new molecules as biological markers is urgently needed. Integrated bioinformatics analysis, which focuses on screening of DEGs, discovering hub node of network-based and doing survival analysis, which has been diffusely used to recognize latent biological markers related to cancer diagnosis, therapy, and prognosis estimation. In recent years, increasing researches have demonstrated that abnormal gene expression is a factor in the tumorigenesis and development, so it is feasible to screen differential genes as biomarkers to assist diagnosis and treatment. In 2017, by developing an integrated approach including GO and KEGG analysis, PPI network creation, hub gene identification, and overall survival analysis, Li L. et al. (2017) picked out 16 hub genes for HCC from three GEO datasets, five of which may be playing a part in the occurrence, development, invasion, metastasis or recurrence of HCC. In 2018, Zhang L. et al. (2018) used bioinformatics methods to select 10 genes from the GEO dataset GSE64041 for the identification of hub genes and pathways of HCC. Gu et al. (2020) recognized 13 crucial genes correlated with progression and prognosis of HCC from the TCGA-LIHC dataset by weighted gene coexpression network analysis. Compared with previous similar studies, our study not only integrated a large sample size of mRNA expression data from the GEO database but also analyzed RNA sequencing result and clinical data from the TCGA-LIHC database to screen

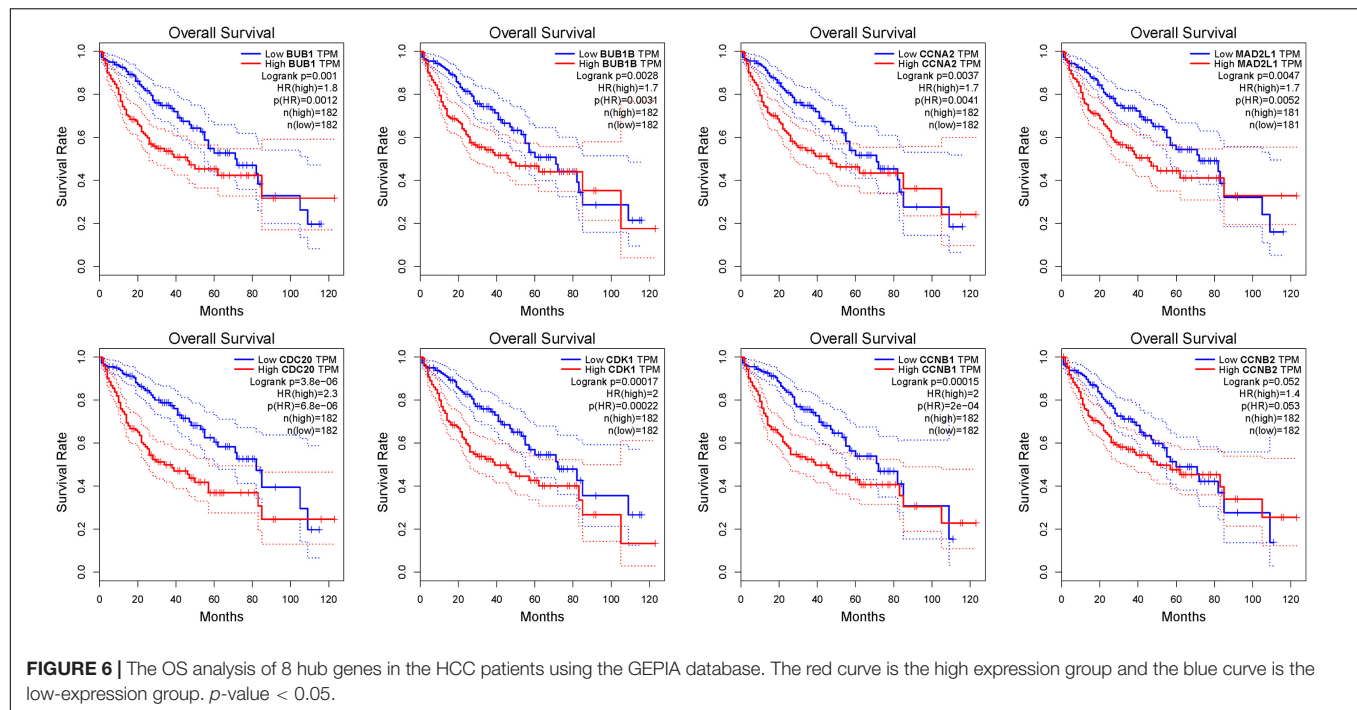


out potential hub genes in HCC. And in the second place, this study validated the DEGs through multiple databases. Finally, we explored the relationship between seven hub genes and existing drugs for cancer therapy, which may provide some guidance for the molecular targeting therapy of HCC in the future.

In our research, DEGs in HCC based on the GEO expression profile of GSE121248 (70 HCC samples and 37 normal samples) and TCGA-LIHC RNA sequencing data (374 HCC samples and 50 normal samples) were identified by bioinformatics analysis. In total, 763 significantly robust DEGs, including 247 upregulated DEGs and 516 downregulated DEGs, were identified. The enrichment analysis results of GO indicated that the DEGs were mostly relevant to “oxidoreductase activity, acting on paired donors,” “monooxygenase activity,” “arachidonic acid epoxygenase activity,” “oxidation-reduction process,” “epoxygenase P450 pathway,” “response to drug,” “extracellular exosome,” “extracellular region,” and “cytosol.” The analysis of KEGG pathway showed that the DEGs were mainly concentrated in the following: “fatty acid degradation pathway,” “metabolic pathways,” “chemical carcinogenesis pathway,” “cell cycle pathway,” and “biosynthesis of antibiotics pathway.”

Previous studies have reported that the arachidonic acid-derived metabolites and cytochrome P450 epoxygenase CYP2J2 possibly play vital roles in regulating malignant tumor, stimulating tumor cell growth, and inhibiting tumor cell apoptosis (Liu L. et al., 2011; Xu et al., 2011; Yarla et al., 2016). Additionally, metabolic pathways are important for cancer cell survival because the metabolic demands of cancer cells are often expressed as increased, and HCC shows a significant alteration in lipid metabolism (Pope et al., 2019). Moreover, dysregulation of the cell cycle processes and mitotic cell cycle plays a vital role in the tumorigenesis and progression (Williams and Stoeber, 2012; Wlodarchak and Xing, 2016). These theories are consistent with our results in GO and KEGG enrichment analysis.

Through building PPI network and analyzing it, we identified crucial hub genes in the PPI network, including CDC20, CDK1, MAD2L1, BUB1, BUB1B, CCNB1, CCNB2, and CCNA2. Using Oncomine and GEPIA validation, the mRNA expression of these eight hub genes in HCC samples was higher than normal liver samples, the finding that was in accord with the microarray results. Subsequently, HPA database data displayed that the protein and mRNA expression of hub genes were consistent,



and most genes were overexpressed in HCC tissue. To inquire prognostic biological markers of HCC, we applied the GEPIA to analyze the influence of hub genes expression level on survival of HCC patients and found that, except CCNB2, the high level gene expression of CDC20, CDK1, MAD2L1, BUB1, BUB1B, CCNB1, and CCNA2 were related to HCC patients poor overall survival. Therefore, these seven genes may be functional in HCC occurrence and development.

It was reported that high expression of CDC20 (cell division cycle protein 20) is associated with poor survival in astrocytoma (Ding et al., 2017), cutaneous squamous cell carcinoma (Chu et al., 2019) and pancreatic ductal adenocarcinoma (Dong et al., 2019). CDC20 promotes the progression of prostate cancer by stabilizing hypo-catenin in tumor-like dry cells (Zhang et al., 2019). However, the expression of cell division cycle protein 20 in HCC still lacks accurate experimental data. As a part of the Ser/Thr protein kinase family, CDK1 (cyclin-dependent kinase 1) is a key molecule that controls the eukaryotic cell cycle. By phosphorylating Bora, Cyclin A/cdk1 could facilitate the phosphorylation, activation and mitotic entry of Aurora A-dependent Plk1 (Vigneron et al., 2018). It is reported that CDK1 overexpression has been found in colorectal cancer, pancreatic ductal adenocarcinoma and thyroid cancer (Zhang P. et al., 2018; Piao et al., 2019; Zheng et al., 2019). It was also reported that CDK1 amplification rate in HCC tissues was usually up to 46% (18/39), which was meaningfully related to poor overall survival ($p = 0.008$) (Wu et al., 2018). These results were in accord with our study findings.

As a pro-oncogene upregulated in gastric cancer, MAD2L1 (mitotic arrest deficient 2-like protein 1) can be downregulated expression by miR-30a-3p, resulting in inhibition of the proliferation of gastric cancer cells (Wang et al., 2019). Besides,

by restraining MAD2L1, miR-200c-5p can inhibit HCC cells proliferation, migration and invasion (Li Y. et al., 2017), suggesting that MAD2L1 can be used in HCC patients prognostic evaluation and targeted therapy. As a cyclin controlling the G1/S and G2/M phases in the cell cycle, CCNA2 (cyclin-A2) is more expressed in CRC samples than in normal samples. The reduction of CCNA2 gene expression would disrupt cell cycle progression and induce apoptosis, thus significantly inhibiting the growth of CRC cells (Gan et al., 2018). By maintaining the expression of CCNA2 protein and the production of arginine, arginine metabolic enzyme argininosuccinate lyase (ASL) can promote the production of nitric oxide synthase, thus promoting the formation of HCC (Hung et al., 2017).

As a mitotic checkpoint serine/threonine kinase, BUB1 is related to tumorigenesis in many cancers. shRNA silencing inhibits the expression of BUB1 gene in glioblastoma tumor cells, thereby reducing the proliferation and tumorigenicity of tumor cells *in vivo* and *in vitro* (Yu et al., 2019). Increased BUB1 expression signally facilitates cell proliferation, while decreased BUB1 expression restrains liver cancer cells proliferation (Zhu et al., 2020). The proliferation, migration, and invasion of PCa cell lines can be enhanced via BUB1B overexpression (Fu et al., 2016). Worse OS and DFS of HCC patients can be predicted by the high expression of BUB1B (Zhuang et al., 2018). CCNB1, an important protein regulating the G2/M (mitotic) cell cycle, is activated by Chk1, exerting its oncogenic role in colorectal cancer cells growth *in vivo* and *in vitro* (Fang et al., 2014). Abnormal FOXM1 expression can transcriptionally activate CCNB1 expression, thereby promoting the proliferation of HCC cells (Chai et al., 2018).

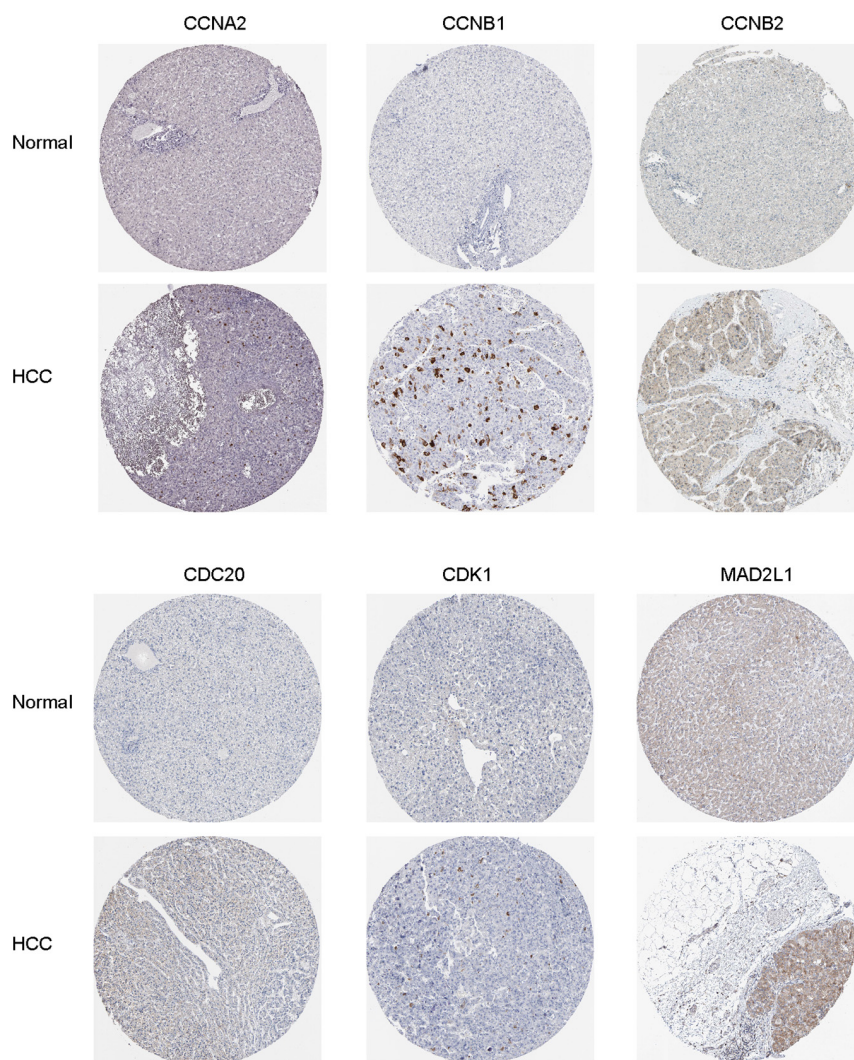


FIGURE 7 | Immunohistochemical staining analysis of hub genes (CCNA2, CCNB1, CCNB2, CDC20, CDK1, and MAD2L1) in HCC tissues and normal liver tissues.

After exploring the potential information about the interactions between the seven hub genes and existing chemotherapeutic drugs, we found that various drugs could influence the expression levels of these hub genes. However, only topotecan, oxaliplatin and azathioprine could simultaneously reduce the expression level all seven hub genes. And it should be noted that further experiments are needed to support whether HCC patients with hub gene overexpression can benefit from hub gene inhibition or whether these key genes may be targets of drug treatment of tumor need ulteriorly biological experiments support.

In the current study, we have discussed that the development of HCC is associated with the overexpression of seven hub genes, which lead to poor overall survival, indicating that they may be considered as potential prognostic biomarkers for HCC. However, our study has several limitations: (1) some important clinical information (for example, different age, tumor size, TNM stage and grade) were not considered; (2) biological experiments

must be carried out in the future to verify the results of our research; (3) the molecular mechanism of hub gene upregulation remains unclear. Therefore, the verification of hub genes will be the focus of our next work.

CONCLUSION

Adopting a series of bioinformatics analysis methods, the current study identified 763 DEGs and seven hub genes (CDC20, CDK1, MAD2L1, BUB1, BUB1B, CCNB1, and CCNA2) that may be involved in hepatocellular carcinoma tumorigenesis and progression. Additionally, multiple database analysis and survival analysis demonstrated that these seven hub genes may regard as a latent prognostic biomarker and the overexpression of these seven hub genes might lead to reduced overall survival in HCC patients. These results provide a theoretical basis for the molecularly targeted therapy and prognosis evaluation of HCC.

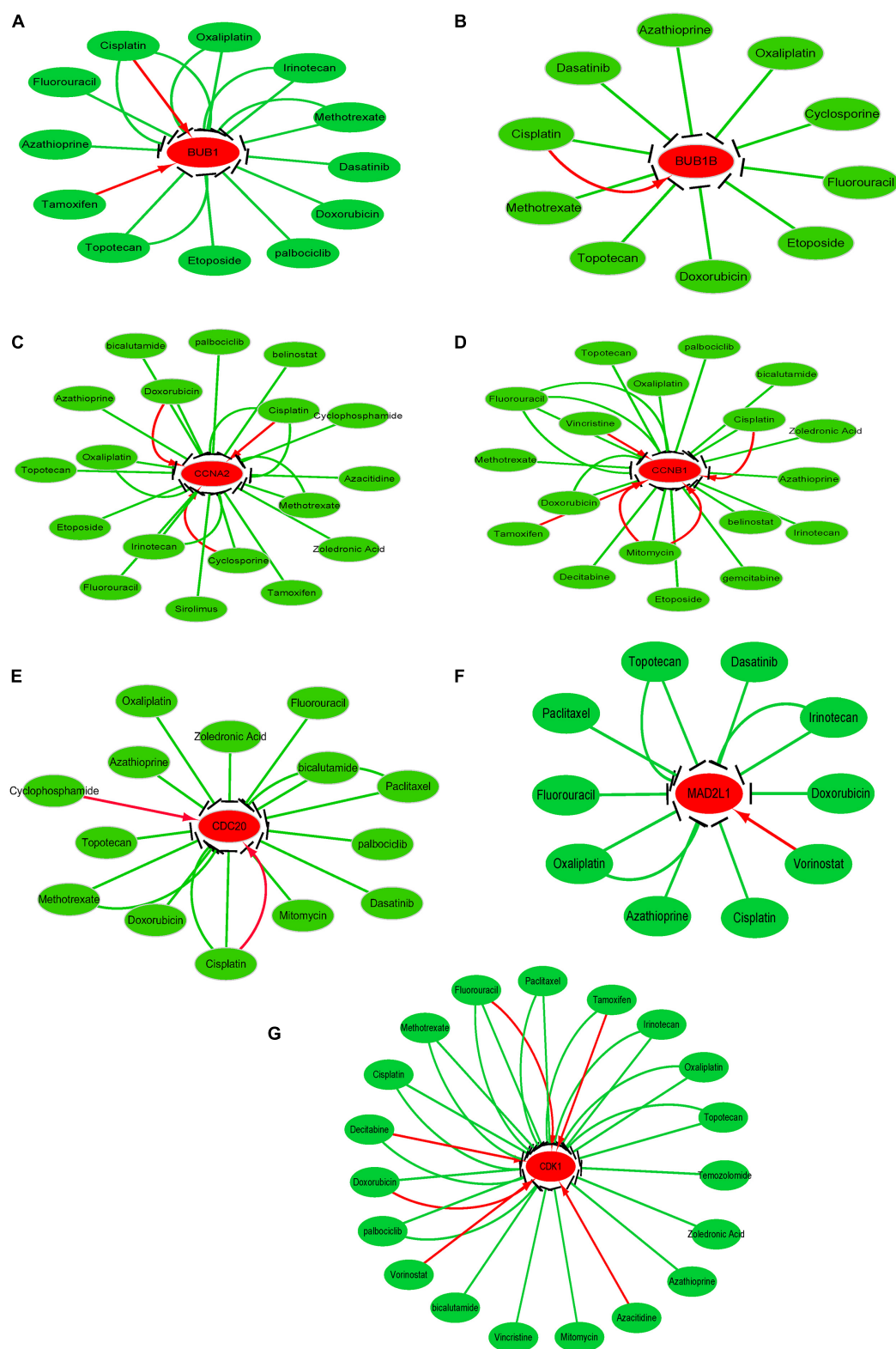


FIGURE 8 | Drug-gene interactions network with chemotherapeutic drugs and seven hub genes was constructed using the CTD database. **(A–G)** shows the relationship between existing chemotherapeutic drugs and the expression levels of hub genes. **(A)** BUB1, **(B)** BUB1B, **(C)** CCNA2, **(D)** CCNB1, **(E)** CDC20, **(F)** MAD2L1, and **(G)** CDK1. The red and green arrows represent that the chemotherapy drugs will increase or decrease the expression of the hub genes. The number of arrows between hub genes and chemotherapy drugs indicates the number of references supported by previous studies.

DATA AVAILABILITY STATEMENT

The data in this study were obtained from the GEO (GSE121248) and TCGA public databases, and the acquisition and application methods complied with the corresponding database guidelines and policies.

AUTHOR CONTRIBUTIONS

ZL conceived the study, analyzed the raw data, and wrote the manuscript. YL prepared the diagrams in the manuscript. BC and QZ provided the useful suggestions in methodology. YC designed and instructed the research. All the authors have read and approved the final manuscript.

FUNDING

This study was funded by the Shantou Medical and health plan (Grant No. 180404094011013).

REFERENCES

- Au, K. P., and Chok, K. S. H. (2018). Multidisciplinary approach for post-liver transplant recurrence of hepatocellular carcinoma: a proposed management algorithm. *World J. Gastroenterol.* 24, 5081–5094. doi: 10.3748/wjg.v24.i45.5081
- Aufhauser, D. D. Jr., Sadot, E., Murken, D. R., Eddinger, K., Hoteit, M., Abt, P. L., et al. (2018). Incidence of occult intrahepatic metastasis in hepatocellular carcinoma treated with transplantation corresponds to early recurrence rates after partial hepatectomy. *Ann. Surg.* 267, 922–928. doi: 10.1097/sla.0000000000002135
- Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2
- Baecker, A., Liu, X., La Vecchia, C., and Zhang, Z. F. (2018). Worldwide incidence of hepatocellular carcinoma cases attributable to major risk factors. *Eur. J. Cancer Prev.* 27, 205–212. doi: 10.1097/cej.0000000000000428
- Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185
- Cao, L., Chen, Y., Zhang, M., Xu, D. Q., Liu, Y., Liu, T., et al. (2018). Identification of hub genes and potential molecular mechanisms in gastric cancer by integrated bioinformatics analysis. *PeerJ* 6:e5180. doi: 10.7717/peerj.5180
- Chai, N., Xie, H. H., Yin, J. P., Sa, K. D., Guo, Y., Wang, M., et al. (2018). FOXM1 promotes proliferation in human hepatocellular carcinoma cells by transcriptional activation of CCNB1. *Biochem. Biophys. Res. Commun.* 500, 924–929. doi: 10.1016/j.bbrc.2018.04.201
- Chen, H., and Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12:35. doi: 10.1186/1471-2105-12-35
- Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., and Lin, C. Y. (2014). cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst. Biol.* 8 Suppl. 4(Suppl. 4):S11. doi: 10.1186/1752-0509-8-s4-s11
- Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wieggers, J., et al. (2019). The comparative toxicogenomics database: update 2019. *Nucleic Acids Res.* 47, D948–D954. doi: 10.1093/nar/gky868

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.571231/full#supplementary-material>

Supplementary Table 1 | List of differentially expressed genes in GSE121248.

Supplementary Table 2 | List of differentially expressed genes between HCC tissue and normal liver tissue in the TCGA dataset.

Supplementary Table 3 | List of common differentially expressed genes both in GSE121248 and the TCGA-LIHC dataset.

Supplementary Table 4 | Gene ontology enrichment analysis of the common differentially expressed genes both in GSE121248 and the TCGA-LIHC dataset.

Supplementary Table 5 | Pathway enrichment analysis of the common DEGs both in GSE121248 and the TCGA-LIHC dataset.

Supplementary Table 6 | STRING significance scores of common differentially expressed genes.

- Deng, J. L., Xu, Y. H., and Wang, G. (2019). Identification of potential crucial genes and key pathways in breast cancer using bioinformatic analysis. *Front. Genet.* 10:695. doi: 10.3389/fgene.2019.00695
- Dennis, G. Jr., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et al. (2003). DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.* 4:3.
- Ding, Y., Yu, S., Bao, Z., Liu, Y., and Liang, T. (2017). CDC20 with malignant progression and poor prognosis of astrocytoma revealed by analysis on gene expression. *J. Neurooncol.* 133, 87–95. doi: 10.1007/s11060-017-2434-8
- Dong, S., Huang, F., Zhang, H., and Chen, Q. (2019). Overexpression of BUB1B, CCNA2, CDC20, and CDK1 in tumor tissues predicts poor survival in pancreatic ductal adenocarcinoma. *Biosci. Rep.* 39, 1–10. doi: 10.1042/bsr20182306
- Fang, Y., Yu, H., Liang, X., Xu, J., and Cai, X. (2014). Chk1-induced CCNB1 overexpression promotes cell proliferation and tumor growth in human colorectal cancer. *Cancer Biol. Ther.* 15, 1268–1279. doi: 10.4161/cbt.29691
- Fu, J., and Wang, H. (2018). Precision diagnosis and treatment of liver cancer in China. *Cancer Lett.* 412, 283–288. doi: 10.1016/j.canlet.2017.10.008
- Fu, X., Chen, G., Cai, Z. D., Wang, C., Liu, Z. Z., Lin, Z. Y., et al. (2016). Overexpression of BUB1B contributes to progression of prostate cancer and predicts poor outcome in patients with prostate cancer. *Onco Targets Ther.* 9, 2211–2220. doi: 10.2147/ott.S101994
- Gan, Y., Li, Y., Li, T., Shu, G., and Yin, G. (2018). CCNA2 acts as a novel biomarker in regulating the growth and apoptosis of colorectal cancer. *Cancer Manag. Res.* 10, 5113–5124. doi: 10.2147/cmar.S176833
- Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy-analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405
- Gu, Y., Li, J., Guo, D., Chen, B., Liu, P., Xiao, Y., et al. (2020). Identification of 13 key genes correlated with progression and prognosis in hepatocellular carcinoma by weighted gene co-expression network analysis. *Front. Genet.* 11:153. doi: 10.3389/fgene.2020.00153
- Harris, P. S., Hansen, R. M., Gray, M. E., Massoud, O. I., McGuire, B. M., and Shoreibah, M. G. (2019). Hepatocellular carcinoma surveillance: an evidence-based approach. *World J. Gastroenterol.* 25, 1550–1559. doi: 10.3748/wjg.v25.i13.1550
- Hung, Y. H., Huang, H. L., Chen, W. C., Yen, M. C., Cho, C. Y., Weng, T. Y., et al. (2017). Argininosuccinate lyase interacts with cyclin A2 in cytoplasm and modulates growth of liver tumor cells. *Oncol. Rep.* 37, 969–978. doi: 10.3892/or.2016.5334

- Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092
- Li, C., Qin, F., Hong, H., Tang, H., Jiang, X., Yang, S., et al. (2019). Identification of Flap endonuclease 1 as a potential core gene in hepatocellular carcinoma by integrated bioinformatics analysis. *PeerJ* 7:e7619. doi: 10.7717/peerj.7619
- Li, L., Lei, Q., Zhang, S., Kong, L., and Qin, B. (2017). Screening and identification of key biomarkers in hepatocellular carcinoma: evidence from bioinformatic analysis. *Oncol. Rep.* 38, 2607–2618. doi: 10.3892/or.2017.5946
- Li, Y., Bai, W., and Zhang, J. (2017). MiR-200c-5p suppresses proliferation and metastasis of human hepatocellular carcinoma (HCC) via suppressing MAD2L1. *Biomed. Pharmacother.* 92, 1038–1044. doi: 10.1016/j.biopha.2017.05.092
- Liu, L., Chen, C., Gong, W., Li, Y., Edin, M. L., Zeldin, D. C., et al. (2011). Epoxyeicosatrienoic acids attenuate reactive oxygen species level, mitochondrial dysfunction, caspase activation, and apoptosis in carcinoma cells treated with arsenic trioxide. *J. Pharmacol. Exp. Ther.* 339, 451–463. doi: 10.1124/jpet.111.180505
- Liu, X., Wu, J., Zhang, D., Bing, Z., Tian, J., Ni, M., et al. (2018). Identification of potential key genes associated with the pathogenesis and prognosis of gastric cancer based on integrated bioinformatics analysis. *Front. Genet.* 9:265. doi: 10.3389/fgene.2018.00265
- Nagy, Á., Lánckzy, A., Menyhárt, O., and Györfy, B. (2018). Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci. Rep.* 8:9227. doi: 10.1038/s41598-018-27521-y
- Piao, J., Zhu, L., Sun, J., Li, N., Dong, B., Yang, Y., et al. (2019). High expression of CDK1 and BUB1 predicts poor prognosis of pancreatic ductal adenocarcinoma. *Gene* 701, 15–22. doi: 10.1016/j.gene.2019.02.081
- Pope, E. D. III, Kimbrough, E. O., Vemireddy, L. P., Surapaneni, P. K., Copland, J. A. III, and Mody, K. (2019). Aberrant lipid metabolism as a therapeutic target in liver cancer. *Expert Opin. Ther. Targets* 23, 473–483. doi: 10.1080/14728222.2019.1615883
- Puszyk, W. M., Trinh, T. L., Chapple, S. J., and Liu, C. (2013). Linking metabolism and epigenetic regulation in development of hepatocellular carcinoma. *Lab. Invest.* 93, 983–990. doi: 10.1038/labinvest.2013.94
- Rhodes, D. R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., et al. (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6, 1–6. doi: 10.1016/s1476-5586(04)80047-2
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303
- Szkarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937
- Tang, Z., Li, C., Kang, B., Gao, G., Li, C., and Zhang, Z. (2017). GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* 45, W98–W102. doi: 10.1093/nar/gkx247
- Thul, P. J., and Lindskog, C. (2018). The human protein atlas: a spatial map of the human proteome. *Protein Sci.* 27, 233–244. doi: 10.1002/pro.3307
- Vigneron, S., Sundermann, L., Labbé, J.-C., Pintard, L., Radulescu, O., Castro, A., et al. (2018). Cyclin A-cdk1-dependent phosphorylation of bora is the triggering factor promoting mitotic entry. *Dev. Cell* 45, 637–650.e7. doi: 10.1016/j.devcel.2018.05.005
- Wang, J., Li, H., Xia, C., Yang, X., Dai, B., Tao, K., et al. (2019). Downregulation of CENPK suppresses hepatocellular carcinoma malignant progression through regulating YAP1. *Onco Targets Ther.* 12, 869–882. doi: 10.2147/ott.S190061
- Wang, X. B., Cui, N. H., Liu, X. N., Ma, J. F., Zhu, Q. H., Guo, S. R., et al. (2018). Identification of promoter hypermethylation as a biomarker for intra-epithelial lesion and cervical cancer: a meta-analysis of published studies, TCGA, and GEO datasets. *Front. Genet.* 9:258. doi: 10.3389/fgene.2018.00258
- Wang, Y., Wang, F., He, J., Du, J., Zhang, H., Shi, H., et al. (2019). miR-30a-3p targets MAD2L1 and regulates proliferation of gastric cancer cells. *Onco Targets Ther.* 12, 11313–11324. doi: 10.2147/ott.S222854
- Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome Atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764
- Williams, G. H., and Stoerber, K. (2012). The cell cycle and cancer. *J. Pathol.* 226, 352–364. doi: 10.1002/path.3022
- Wlodarchak, N., and Xing, Y. (2016). PP2A as a master regulator of the cell cycle. *Crit. Rev. Biochem. Mol. Biol.* 51, 162–184. doi: 10.3109/10409238.2016.1143913
- Wu, C. X., Wang, X. Q., Chok, S. H., Man, K., Tsang, S. H. Y., Chan, A. C. Y., et al. (2018). Blocking CDK1/PDK1/β-Catenin signaling by CDK1 inhibitor RO3306 increased the efficacy of sorafenib treatment by targeting cancer stem cells in a preclinical model of hepatocellular carcinoma. *Theranostics* 8, 3737–3750. doi: 10.7150/thno.25487
- Xu, J. S., Wang, X., He, J., Liao, K. L., and Wang, X. Z. (2020). Combining bioinformatics techniques to explore the molecular mechanisms involved in pancreatic cancer metastasis and prognosis. *J. Cell. Mol. Med.* 24, 14128–14138. doi: 10.1111/jcmm.16023
- Xu, X., Zhang, X. A., and Wang, D. W. (2011). The roles of CYP450 epoxygenases and metabolites, epoxyeicosatrienoic acids, in cardiovascular and malignant diseases. *Adv. Drug Deliv. Rev.* 63, 597–609. doi: 10.1016/j.addr.2011.03.006
- Yan, H., Zheng, G., Qu, J., Liu, Y., Huang, X., Zhang, E., et al. (2019). Identification of key candidate genes and pathways in multiple myeloma by integrated bioinformatics analysis. *J. Cell. Physiol.* 234, 23785–23797. doi: 10.1002/jcp.28947
- Yang, J. D., Hainaut, P., Gores, G. J., Amadou, A., Plymoth, A., and Roberts, L. R. (2019). A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat. Rev. Gastroenterol. Hepatol.* 16, 589–604. doi: 10.1038/s41575-019-0186-y
- Yang, K., Zhang, S., Zhang, D., Tao, Q., Zhang, T., Liu, G., et al. (2019). Identification of SERPINE1, PLA1 and ACTA1 as biomarkers of head and neck squamous cell carcinoma based on integrated bioinformatics analysis. *Int. J. Clin. Oncol.* 24, 1030–1041. doi: 10.1007/s10147-019-01435-9
- Yang, X., Zhu, S., Li, L., Zhang, L., Xian, S., Wang, Y., et al. (2018). Identification of differentially expressed genes and signaling pathways in ovarian cancer by integrated bioinformatics analysis. *Onco Targets Ther.* 11, 1457–1474. doi: 10.2147/ott.S152238
- Yarla, N. S., Bishayee, A., Sethi, G., Reddanna, P., Kalle, A. M., Dhananjaya, B. L., et al. (2016). Targeting arachidonic acid pathway by natural products for cancer prevention and therapy. *Semin. Cancer Biol.* 4, 48–81. doi: 10.1016/j.semcancer.2016.02.001
- Yu, H., Zhang, S., Ibrahim, A. N., Deng, Z., and Wang, M. (2019). Serine/threonine kinase BUB1 promotes proliferation and radio-resistance in glioblastoma. *Pathol. Res. Pract.* 215:152508. doi: 10.1016/j.prp.2019.152508
- Zhang, L., Huang, Y., Ling, J., Zhuo, W., Yu, Z., Shao, M., et al. (2018). Screening and function analysis of hub genes and pathways in hepatocellular carcinoma via bioinformatics approaches. *Cancer Biomark.* 22, 511–521. doi: 10.3233/cbm-171160
- Zhang, P., Kawakami, H., Liu, W., Zeng, X., Strebhardt, K., Tao, K., et al. (2018). Targeting CDK1 and MEK/ERK overcomes apoptotic resistance in BRAF-mutant human colorectal cancer. *Mol. Cancer Res.* 16, 378–389. doi: 10.1158/1541-7786.Mcr-17-0404
- Zhang, Q., Huang, H., Liu, A., Li, J., Liu, C., Sun, B., et al. (2019). Cell division cycle 20 (CDC20) drives prostate cancer progression via stabilization of beta-catenin in cancer stem-like cells. *EBioMedicine* 42, 397–407. doi: 10.1016/j.ebiom.2019.03.032
- Zhang, X., Feng, H., Li, Z., Li, D., Liu, S., Huang, H., et al. (2018). Application of weighted gene co-expression network analysis to identify key modules and hub genes in oral squamous cell carcinoma tumorigenesis. *Onco Targets Ther.* 11, 6001–6021. doi: 10.2147/ott.S171791
- Zheng, H. P., Huang, Z. G., He, R. Q., Lu, H. P., Dang, Y. W., Lin, P., et al. (2019). Integrated assessment of CDK1 upregulation in thyroid cancer. *Am. J. Transl. Res.* 11, 7233–7254.

- Zhu, L. J., Pan, Y., Chen, X. Y., and Hou, P. F. (2020). BUB1 promotes proliferation of liver cancer cells by activating SMAD2 phosphorylation. *Oncol. Lett.* 19, 3506–3512. doi: 10.3892/ol.2020.11445
- Zhu, R. X., Seto, W. K., Lai, C. L., and Yuen, M. F. (2016). Epidemiology of hepatocellular carcinoma in the Asia-Pacific region. *Gut Liver* 10, 332–339. doi: 10.5009/gnl15257
- Zhuang, L., Yang, Z., and Meng, Z. (2018). Upregulation of BUB1B, CCNB1, CDC7, CDC20, and MCM3 in tumor tissues predicted worse overall survival and disease-free survival in hepatocellular carcinoma patients. *Biomed. Res. Int.* 2018:7897346. doi: 10.1155/2018/7897346

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Li, Lin, Cheng, Zhang and Cai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Dynamical Modeling of the Core Gene Network Controlling Transition to Flowering in *Pisum sativum*

Polina Pavlinova¹, Maria G. Samsonova¹ and Vitaly V. Gursky^{2*}

¹Mathematical Biology and Bioinformatics Laboratory, Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russia, ²Theoretical Department, Ioffe Institute, Saint Petersburg, Russia

OPEN ACCESS

Edited by:

Tatiana V. Tatarinova,
University of La Verne, United States

Reviewed by:

Jianjun Zhao,
Agricultural University of Hebei, China
Hamed Bostan,
National Institute of Environmental
Health Sciences (NIEHS),
United States

*Correspondence:

Vitaly V. Gursky
gursky@math.ioffe.ru

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 06 October 2020

Accepted: 28 January 2021

Published: 11 March 2021

Citation:

Pavlinova P, Samsonova MG and
Gursky VV (2021) Dynamical
Modeling of the Core Gene Network
Controlling Transition to Flowering in
Pisum sativum.
Front. Genet. 12:614711.
doi: 10.3389/fgene.2021.614711

Transition to flowering is an important stage of plant development. Many regulatory modules that control floral transition are conservative across plants. This process is best studied for the model plant *Arabidopsis thaliana*. The homologues of *Arabidopsis* genes responsible for the flowering initiation in legumes have been identified, and available data on their expression provide a good basis for gene network modeling. In this study, we developed several dynamical models of a gene network controlling transition to flowering in pea (*Pisum sativum*) using two different approaches. We used differential equations for modeling a previously proposed gene regulation scheme of floral initiation in pea and tested possible alternative hypothesis about some regulations. As the second approach, we applied neural networks to infer interactions between genes in the network directly from gene expression data. All models were verified on previously published experimental data on the dynamic expression of the main genes in the wild type and in three mutant genotypes. Based on modeling results, we made conclusions about the functionality of the previously proposed interactions in the gene network and about the influence of different growing conditions on the network architecture. It was shown that regulation of the *PIM*, *FTa1*, and *FTc* genes in pea does not correspond to the previously proposed hypotheses. The modeling suggests that short- and long-day growing conditions are characterized by different gene network architectures. Overall, the results obtained can be used to plan new experiments and create more accurate models to study the flowering initiation in pea and, in a broader context, in legumes.

Keywords: pea, gene network, flowering initiation, differential equations, neural networks, dynamical model

INTRODUCTION

Flowering is associated with a significant physiological change in plant development which manifests the transition from vegetative growth to reproductive development. For the reproductive success of plants, it is important for this transition to occur at the most appropriate moment. Various exogenous and endogenous pathways contribute to the control for the flowering time, and these pathways are best studied for the model plant *Arabidopsis thaliana* (Srikanth and Schmid, 2011; Andrés and Coupland, 2012; Khan et al., 2014). The key factor in the activation of the photoperiodic pathway of flowering initiation in *Arabidopsis* is the protein encoded by

the *FLOWERING LOCUS T* (*FT*) gene. The *FT*, a phosphatidylethanolamine binding protein (PEBP), is a mobile signal transported from the leaves to the top of the shoot apex, where it promotes the plant's transition to flowering. Expression of the *FT* gene depends on the influence of external and internal signals, which allow the plant to regulate the flowering initiation time (Kardailsky et al., 1999; Kobayashi et al., 1999; Jaeger et al., 2013). After synthesis in the leaves, the *FT* protein moves to the shoot apical meristem and forms a complex with the bZIP-type transcription factor *FLOWERING LOCUS D* (*FD*; Abe et al., 2005), which belongs to the 14-3-3 protein family (Taoka et al., 2011). Main target genes of the *FT*-*FD* complex are the flower meristem identity gene *API* (*APETALA1*; Wigge et al., 2005) and *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS1* (*SOC1*; Yoo et al., 2016). The latter is an activator of the gene *LEAFY* (*LFY*), which also controls the transition of shoot apical meristems to flower meristems (Lee et al., 2008). The flower meristem identity genes *API* and *LFY* transcriptionally activate each other (Jaeger et al., 2013).

The balance between activation and repression of flowering initiation is important for plants with indeterminate inflorescence architecture, in which newly forming flowers do not stop further plant growth (Benlloch et al., 2015). The key repressor of flowering initiation in *Arabidopsis* is the gene *TERMINAL FLOWER1* (*TFL1*), which is a close relative of *FT* and encodes a protein belonging to the PEBP family. This protein is expressed during floral transition in the center of the shoot apical meristem and maintains it in the vegetative state by suppressing the expression of *LFY* and *API* (Jaeger et al., 2013; Goretti et al., 2020). In turn, *API* represses *TFL1* by directly binding its regulatory elements (Kaufmann et al., 2010). This mutual repression between *TFL1* and *LFY/API* explains the inflorescence meristem maintenance and flower meristem formation on its flanks (Benlloch et al., 2015). The minimal graph summarizing the genetic control of the photoperiod pathway in flower transition in *Arabidopsis* is shown in **Figure 1A**.

During evolution of legumes, the floral transition regulation has become more complex. This class of plants is characterized by the formation of a more complicated, the so-called compound, inflorescence architecture (Benlloch et al., 2015). In the process of growth, two meristems (primary and secondary) are successively formed. Moreover, multiple copies of the PEBP genes were identified in legumes homologous to *FT* and *TFL1*, associated with multiple genome duplication events during evolution (Hecht et al., 2011). The legume *FT*-like genes are subdivided into three subclasses: *FTa*, *FTb*, and *FTc*. Five *FT*-like genes from these subclasses were identified in pea (*Pisum sativum*; *FTa1*, *FTa2*, *FTb1*, *FTb2*, and *FTc*). These genes are characterized by variable expression patterns under different conditions. Under long day (LD) conditions, *FTa1* and *FTb2* are expressed in the leaves, while under short day (SD) conditions only decreased expression of *FTa1* is observed. In the plant apex, only *FTc* and *FTa1* are expressed. Such differences indicate distinct functions of the *FT* genes in floral initiation in pea (Hecht et al., 2011).

Pea homologues of the flower meristem identity genes *API* and *LFY* are *PROLIFERATING INFLORESCENCE MERISTEM*

(*PIM/PEAM4*) and *UNIFOLIATA* (*UNI*), respectively (Hofer et al., 1997; Taylor et al., 2002). Homologues of the floral repressor *TFL1* in pea include *DETERMINATE* (*DET*), which is a marker of the primary inflorescence meristem (Berbel et al., 2012), and *LATE FLOWERING* (*LF*), whose function is not entirely clear (Foucher et al., 2003). The secondary inflorescence meristem is under control of *VEGETATIVE1* (*VEG1*; Berbel et al., 2012). A pea homologue of *FD* is *VEGETATIVE2* (*VEG2*), which is thought to form the complex with *FTs* similarly to *Arabidopsis* (Sussmilch et al., 2015).

Hecht et al. (2011) qualitatively analyzed the expression of these genes in pea, both in the leaves and in the shoot apical meristem, under different growth conditions and genotypes. Later, Sussmilch et al. (2015) proposed a scheme for regulations underlying the compound inflorescence development and floral transition in pea, as depicted in **Figure 1B**. In our study, we apply modeling to test whether the proposed regulation scheme fits the expression data quantitatively.

Methods of mathematical modeling are widely applied to the analysis of gene networks. These methods include Boolean models, ordinary differential equations (ODEs), neural networks, Bayesian networks, and stochastic modeling (Chai et al., 2014; Le Novère, 2015). The choice between different modeling approaches depends on the type of data used to calibrate the model.

Various modeling techniques were used for the quantitative analysis of gene networks involved in plant growth and development (Haque et al., 2019), in particular, in the photoperiodic pathway of floral transition. A method of neural networks was applied to study the transition to flowering of *Arabidopsis* (Welch et al., 2003). This model had a prescribed neural network architecture and described the interaction of the main genes responsible for various pathways of flowering initiation in the plant. The model was trained on values of such phenotypic parameters as the daylight length and the number of days after sowing. Later, the main regulatory elements underlying the photoperiodic pathway of *Arabidopsis* transition to flowering were identified using a dynamical model based on differential equations, which was applied to the data on flowering time of the wild and mutant genotypes (Jaeger et al., 2013). It was shown that the dynamics of flowering initiation can be explained by dividing the gene network into several feedback and forward loops with specific functional roles (Pullen et al., 2013). A more advanced model was developed later by Leal Valentim et al. (2015), in which additional regulators (*SOC1* and *AGL24*) were added into the activation of *LFY* by the *FT*-*FD* complex, and the model was fitted to gene expression data. This approach allowed to test various hypotheses about *LFY* regulation by *SOC1* and *AGL24* and elucidated a nonlinear nature of the flowering network. Wang et al. (2014) investigated different approximations used to formulate model equations and compared their influence on the model performance in describing floral initiation in *Arabidopsis*. Apart from *Arabidopsis*, similar models of floral transition were also elaborated for chickpea (*Cicer arietinum*), which is a member of the legume family. Like pea, it has multiple homologues of the *FT* and *TFL1* genes (Ridge et al., 2017).

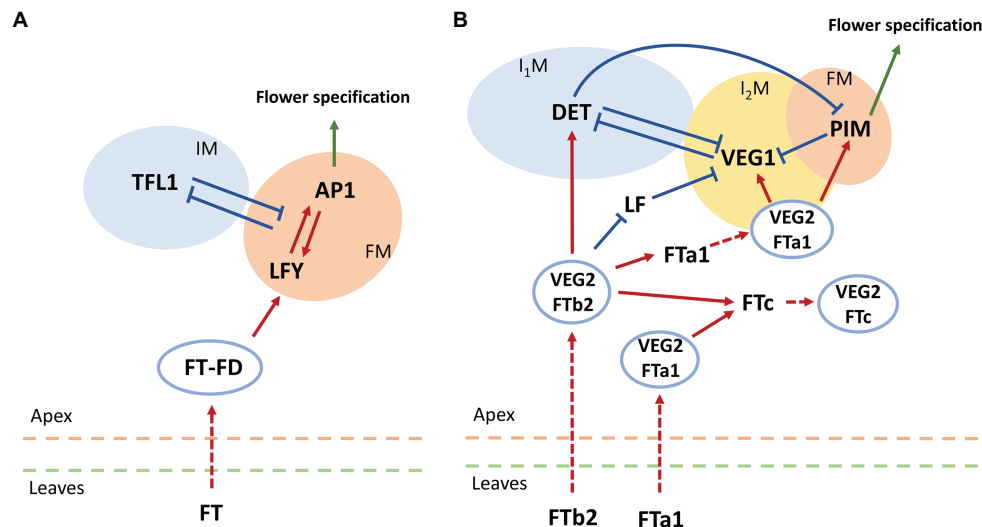


FIGURE 1 | Schemes for genetic control of floral initiation in **(A)** *Arabidopsis* and **(B)** pea (*Pisum sativum*). **(A)** The core gene network controlling floral transition in *Arabidopsis thaliana* (Jaeger et al., 2013). IM, inflorescence meristem; FM, flower meristem. **(B)** The core gene network controlling floral transition in pea. FTb2 is expressed in plant leaves under LD conditions; it then moves to the apex, where it interacts with VEGETATIVE2 (VEG2). The resulting complex VEG2-FTb2 stimulates the formation of the primary inflorescence meristem (I₁M) by activating the meristem identity gene *DETERMINATE* (DET). Within the apex, FTb2 also activates *FTa1* and *FTc* and downregulates *LF*. *FTa1* is expressed both in the leaves and in the apex. By means of the complex with VEG2, *FTa1* probably stimulates *FTc* expression and activates *VEGETATIVE1* (VEG1). The latter activation leads to higher expression of VEG1, enhanced by the reduced repression from LF, and this expression initiates the secondary inflorescence meristem formation (I₂M). *FTa1*-VEG2 also activates the floral meristem (FM) identity gene *PROLIFERATING INFLORESCENCE MERISTEM* (PIM). The mutual repression between the three meristem identity genes (DET, VEG1, and PIM) ensures a spatial separation of the corresponding developmental compartments, maintaining the indeterminate inflorescence development. Dashed arrow-headed lines indicate movement of proteins from leaves to apex and protein complex formation within the apex. Red solid arrow-headed lines correspond to transcriptional activation, and blue solid T-like lines indicate transcriptional repression. The regulation scheme is based on a figure by Sussmilch et al., 2015.

A dynamical model of the flowering gene network was developed and used for testing various hypotheses on how the *FT*- and *TFL1*-like genes combine in regulating the flower meristem identity genes in the ICCV 96029 chickpea cultivar (Gursky et al., 2018). The same model was not successful for CDC Frontier, which is another chickpea cultivar. A machine learning-based modeling approach was developed and applied for this cultivar, predicting that SD and LD growing conditions may be associated with different architectures of the flowering gene network (Podolny et al., 2020). Extending a classical qualitative model for the control of flowering initiation, Wenden et al. (2009) elaborated a quantitative model of flowering in pea (Wenden and Rameau, 2009). This model was used to formulate new hypotheses about the signals controlling flowering. More sophisticated modeling and software platforms were proposed taking into account mechanical processes during flower development and, more generally, morphogenesis in plants, and using advanced data quantification methods (Barbier de Reuille et al., 2015; Boudon et al., 2015).

We extend the previous modeling attempts to floral transition in pea. We construct several dynamical models and apply them to the previously published data on the photoperiodic pathway of flowering initiation in pea (Hecht et al., 2011; Sussmilch et al., 2015). We specifically investigate the compatibility of the network from Figure 1B to the data at the quantitative level.

RESULTS

We calibrated our models on the previously published dynamic expression data of genes responsible for flowering initiation in pea (cultivar NGB5839; Hecht et al., 2011; Sussmilch et al., 2015). We extracted the expression data for three *FT*-like genes (*FTa1*, *FTb2*, and *FTc*), two homologues of the *TFL1* gene (*DET* and *LF*), one homologue of the *FD* gene (*VEG2*), a homologue of the flower meristem identity gene *AP1* (*PIM*), and the *VEG1* gene responsible for secondary meristem formation. For all genes except *VEG1*, data were available for the SD and LD growth conditions in the wild type; *VEG1* expression data were available only for LD. In addition, expression data for the same genes were extracted for three mutant genotypes: *late1-2*, *dne-1*, and *gigas-2*. *late1-2* is a mutant for gene *LATE1*, which has delayed flowering under LD. *dne-1* represents a mutant for gene *DNE1*, which starts flowering under SD at the same time as a wild-type plant under LD. *gigas-2* is the *FTa1* null mutant.

Dynamical Models Based on the Proposed Regulation Scheme

We developed a dynamical model describing gene expression according to the regulation scheme shown in Figure 1B. We formulated ODEs implementing the Michaelis-Menten kinetics for the expression of each gene under the influence

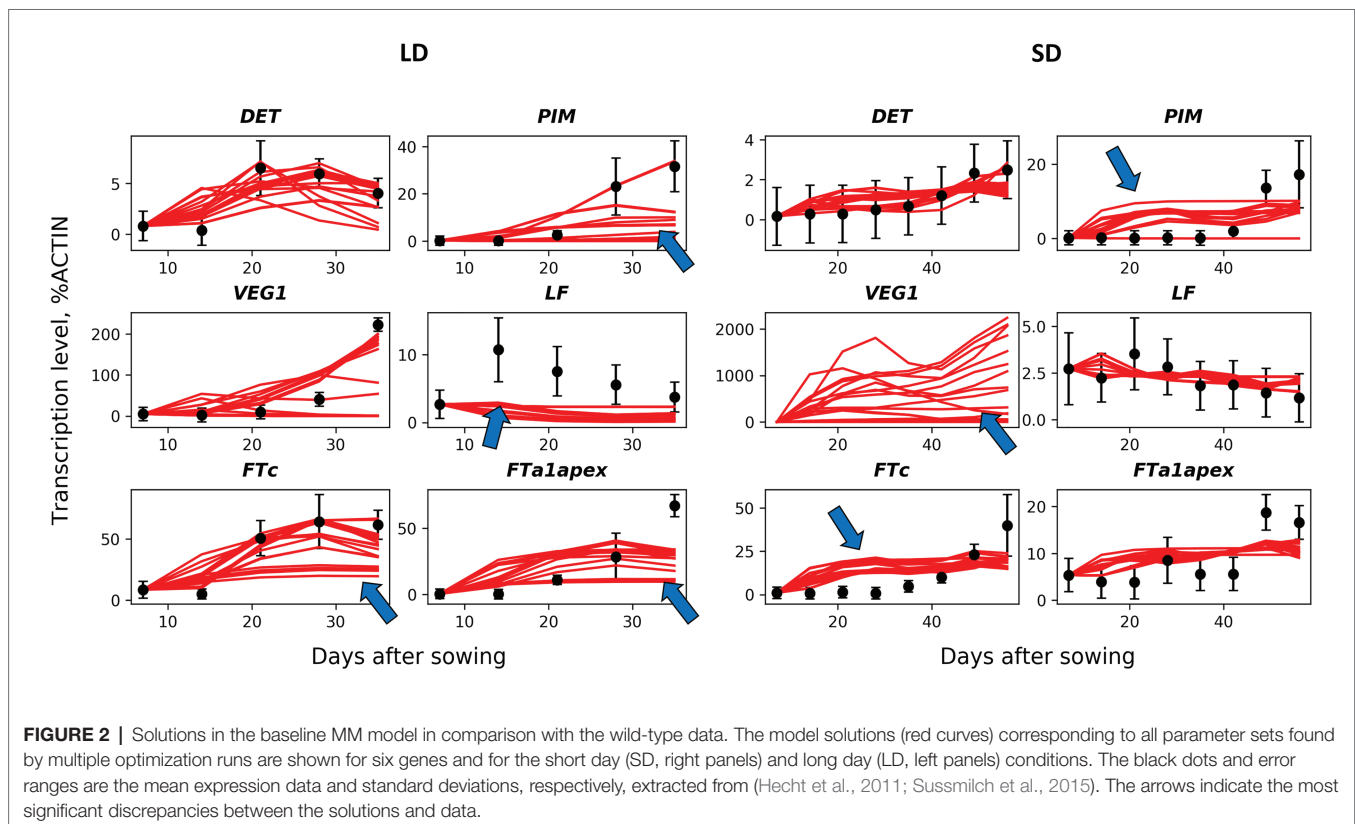
of its regulators and fitted this model to the expression data, in order to understand how the proposed regulation scheme matches the data at the quantitative level. We first investigated a baseline model [the MM model; equations (1)–(11) in Materials and Methods] which includes only regulations shown in **Figure 1B** and, in particular, considers the competitive binding of VEG2 by FTa1, FTb2, and FTc. We found values of free parameters by fitting this model to the wild-type expression data. In order to reduce the probability of overfitting, we analyzed all solutions resulted from a series of the numerical optimization runs (**Figure 2**). These solutions qualitatively match the data dynamics but have several quantitative discrepancies. In SD, insufficient repression at early times and insufficient activation at later times of *PIM* and *FTc* are observed. As data for *VEG1* were absent in SD, the solution for this protein was not fitted to data. As a consequence, most of the *VEG1* solutions have unrealistically high expression values in SD. The defects in LD include deficient activation at later times in most solutions for *PIM*, *FTc*, and apical FTa1, and deficient activation of *LF* at early times. Testing the model on the data from mutants also showed a qualitative correspondence between the model and the data, but with quantitative defects (**Supplementary Figures 1, 2**).

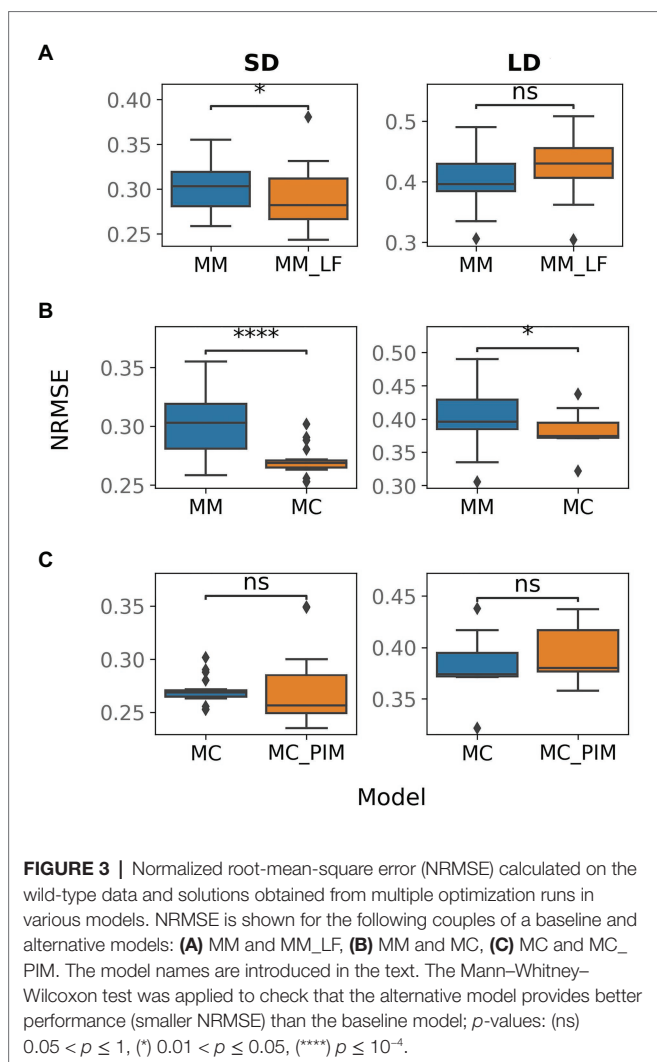
Testing Alternative Hypotheses About Gene Regulations

To improve the baseline model, we tested several alternative hypotheses about additional interactions in the gene network. TFL1 inhibits floral initiation in *Arabidopsis* by repressing

expression of *API*. Among two pea homologues of the *TFL1* gene (*DET* and *LF*), only *DET* was suggested as a repressor of *PIM*, which is the pea homologue of *API* (**Figure 1B**). We assumed that *LF* also represses *PIM* and that this repression would reduce overexpressed *PIM* at early times in SD. To test this hypothesis, we formulated the MM_LF model by adding the new regulation into equations of the baseline MM model [see equation (12) in Materials and Methods] and fitted the new model to the wild-type data. The MM_LF model showed a slightly better performance as compared to the MM model in SD, but the performance became worse in LD (**Figure 3A**). Taking into account that the early dynamics of *PIM* is not improved essentially (**Supplementary Figure 3**) and MM_LF is much worse than MM on data from the *gigas-2* mutant (**Supplementary Figure 4**), we can reject the hypothesis about *PIM* repression by LF.

On the next step, we tested whether the competitive binding of VEG2 by FTa1, FTb2, and FTc is essential. We adjusted the MM model by assuming that the binding is uncompetitive, so that the concentrations of the complexes were taken equal to the product of the concentrations of the corresponding proteins [equations (13)–(15) in Materials and Methods]. This new alternative model (MC model) demonstrated a better performance on the wild-type data as compared to the MM model, both in SD and in LD (**Figure 3B**), and was also better than the MM model on the data from mutant genotypes (**Supplementary Figure 4; Supplementary Table 3**). The wild-type solutions in the MC model are less variable than in the





MM model and show improvement in expression dynamics of *PIM* and *FTc* in LD (Supplementary Figure 5). VEG1 in the MC model also has a more reasonable expression dynamics range in SD (Supplementary Figure 5). These results suggest that the binding of VEG2 by the FT proteins is uncompetitive.

In SD, the MM and MC models both have solutions with an overstated early and understated late expression of *PIM*. One possible solution to this problem is to make the dynamical curve of *PIM* respond in a more nonlinear way to the monotonically increasing expression curve of the complex VEG2-FTa1, which is the only activator of *PIM*. This nonlinearity can be achieved by adding a cooperativity parameter into the model, responsible for the putative cooperative binding of VEG2-FTa1 to the *PIM* promoter. It was shown that homologues of the FD (VEG2) and FT proteins form a complex consisting of several subunits in rice (*Oryza sativa*), thus sustaining the hypothesis about cooperative regulation by the VEG2-FT complexes (Taoka et al., 2011; Tsuji et al., 2013). We implemented the cooperativity hypothesis into the MC model by assuming that the cooperativity parameter n ($n > 1$) in the term responsible for the regulation of *PIM* by VEG2-FTa1 is an additional free parameter (MC_PIM model).

However, the new model neither improved the total performance as compared to the MC model (Figure 3C) nor fixed the *PIM* expression dynamics in SD (Supplementary Figure 6), thus suggesting that the regulation of *PIM* is noncooperative.

Models Trained on Full Data

The models described above were trained on the wild-type data and tested on the mutant data. In these computational experiments, the MC model outperformed other alternative models on both the wild-type and mutant data (Figure 3; Supplementary Figure 4). However, most of the defects shown for the baseline MM model persisted in the MC model. In order to increase the amount of data used to optimize parameter values, we used the same equations as in the MC model and fitted them to gene expression data for all genotypes (wild type, *dne-1*, *late1-2*, and *gigas-2*). Since we used all the available data to fit the model, we aimed to investigate the maximal possible performance of the model in this computational experiment. Later, we will split both the wild-type and mutant data into training and testing subsets when modeling with neural networks.

We refer to this model trained on the complete data set as MC_Cdata. In terms of the normalized error, the MC_Cdata model expectedly outperforms MC on the mutant data but has a bit worse performance on the wild-type data (Supplementary Figure 7). However, the wild-type expression dynamics is qualitatively similar in the two models (Supplementary Figures 5, 8).

One of the defects observed in all models is a low FTa1 concentration at later times in the wild type (Figure 2; Supplementary Figures 5, 8). According to the proposed regulation scheme, FTa1 in the apical meristem is activated only by the VEG2-FTb2 complex. In order to add activation to the FTa1 expression, we suggested that FTa1 activates its own production in the apex. We tested this hypothesis by inserting an additional term into the equation for FTa1 that characterized FTa1 activation by the VEG2-FTa1 complex [equation (16) in Materials and Methods] and fitting the resulted model to the complete data set (MC_Cdata_FTa1 model). The new model did not show improved performance as compared to the MC_Cdata model (Figure 4), thus rejecting the hypothesis.

It was shown for soybean (*Glycine max*); another representative of legumes, that activation of flowering initiation under LD conditions involves different regulatory blocks than under SD conditions (Wu et al., 2019). We investigated whether the model performance can be improved if we use the same proposed regulatory scheme for pea but fit the model to the SD and LD data separately (MC_SDdata and MC_LDdata models, respectively). The SD data comprise the SD part from the wild type and data from the *dne-1* mutant, and the LD data include the LD portion of the wild-type data and data from the *late1-2* and *gigas-2* mutants. MC_SDdata and MC_LDdata showed better performance than the MC_Cdata model for the SD and LD growing conditions, respectively (Figure 4). It should be noted that the comparison between these models is not a rigorous test, because the MC_SDdata and MC_LDdata models were fitted to fewer data points than MC_Cdata for the same number of parameters.

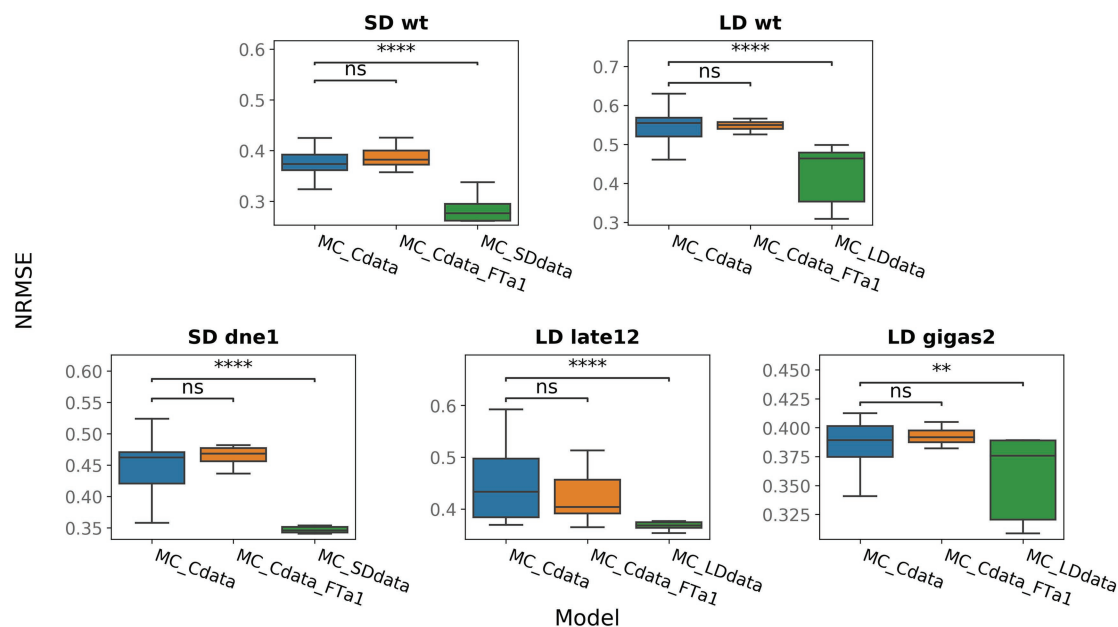


FIGURE 4 | NRMSE in models trained on the full data (MC_Cdata and MC_Cdata_FTa1) or on the SD and LD portions of the full data (MC_SDdata and MC_LDdata, respectively). The Mann–Whitney–Wilcoxon test was applied to check that the alternative models provide better performance (smaller NRMSE) than the MC_Cdata model; *p*-values: (ns) $0.05 < p \leq 1$, (**) $0.001 < p \leq 0.01$, (****) $p \leq 10^{-4}$.

However, this computational experiment shows that it is possible to reduce the modeling error by narrowing the model to either SD or LD. This suggests that it may be not feasible to use uniform regulatory assumptions under the two growing conditions. The analysis of individual genes reveals that the MC_LDdata model most significantly improves the expression dynamics of *PIM* and *FTc* (both in the wild type, LD), while MC_SDdata improves the expression dynamics of *DET* (wild type, SD, and *dne-1*) and *FTa1* in the apex (*dne-1*; **Supplementary Figures 9–13**).

Dynamical Models Based on Neural Networks

The previously described dynamical models were based on the suggested regulation scheme underlying floral initiation in pea (**Figure 1B**), so studying these models was aimed at answering the question about the quantitative correspondence between this scheme and the expression data. In the next stage of the study, we developed models without prescribing a specific topology of the gene regulatory network, thus trying to answer the question of what regulations can be inferred from the expression data ab initio. Along with changing the question, we also changed the formalism of ODEs to the neural network method to formulate new models, so as not to be dependent on only one modeling method and, thus, increase the robustness of conclusions.

We developed three models (NN, NN_SDdata, and NN_LDdata) based on neural networks, all of which were constructed on the same principles and differed from each other only by the data used for their training. The models were formulated as dynamical regression models in which the apical expression of all genes on the current day was determined by the apical

expression of the same genes and the expression of the *FT*-genes from the leaves from the previous day (Podolny et al., 2020). As VEG1 data were present only in LD, we excluded VEG1 from the model for simplicity; we also considered VEG2 as an independent variable. The NN model was trained on the full data (wild-type, *dne-1*, *late1-2*, and *gigas-2*), NN_SDdata on the SD portion of the full data (wild-type, SD, and *dne-1*), and NN_LDdata on the LD portion of the full data (wild-type, LD, *late1-2*, and *gigas-2*). For the NN and NN_SDdata models, we separated data from several days for each condition (daylight and genotype) as the testing dataset, and all data from the *late1-2* mutant were used as the testing dataset for the NN_LDdata model.

The solutions in the NN_SDdata and NN_LDdata models show better correspondence to the wild-type data compared to the models based on the proposed regulation scheme (**Figures 5–6**). There are improvements in expression dynamics of *DET*, *LF*, and *FTa1* in the apex under the LD conditions and of *PIM*, *FTc*, and *FTa1* in the apex under the SD conditions (arrows in **Figure 5**). The solutions in the NN model is close to NN_SDdata and NN_LDdata but have defects for *FTa1* and *DET* in LD and for *LF* and *FTc* in SD. In contrast to the wild-type data, the neural network models do not show a definite difference with the ODE-based models on the mutant data (**Figure 6**). NN and NN_LDdata are better for *gigas-2*, while the comparison is in favor of the ODE-based models for the other two mutants. A worse performance of NN_LDdata for *late1-2* can be explained by the fact that the whole data from this mutant were used as a testing set in this model.

In order to understand what interactions were restored in the neural network models, we simulated gene knockouts in

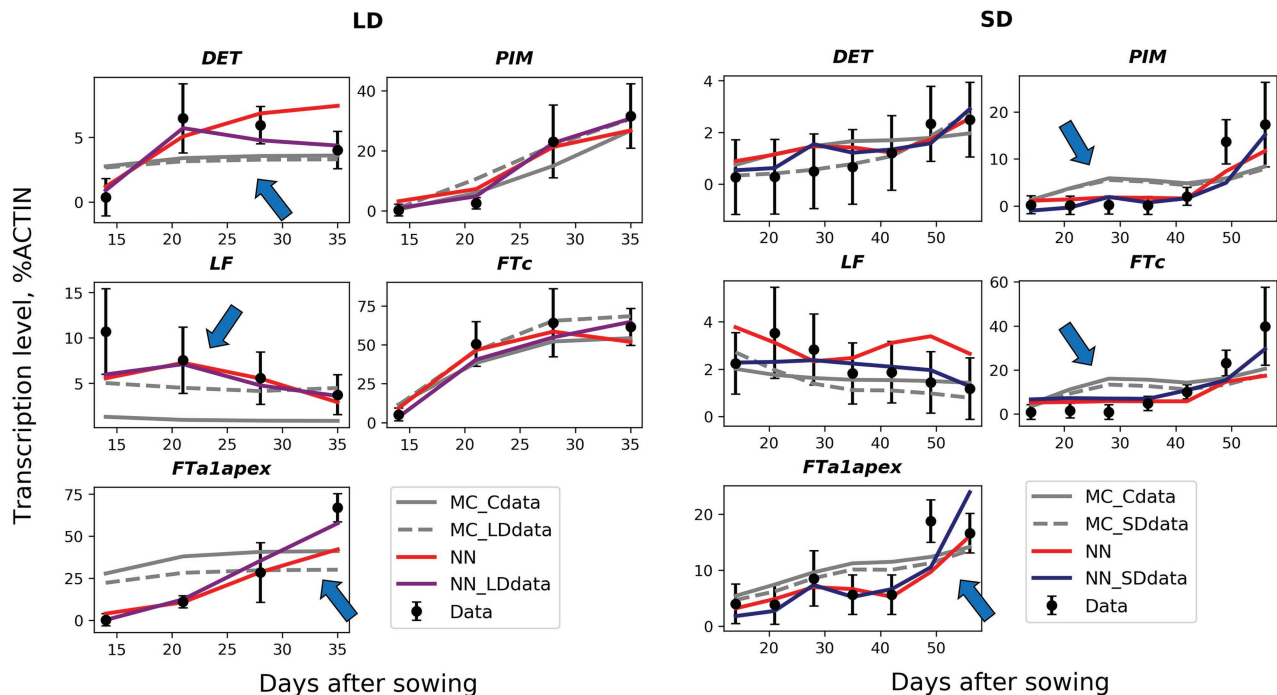


FIGURE 5 | Best solutions in the neural network models (NN, NN_SDdata, and NN_LDdata) in comparison with the models based on the proposed regulation scheme and ordinary differential equations (ODEs; MC_Cdata, MC_SDdata, and MC_LDdata), for the wild type and two growing conditions. The black dots and error ranges are the mean expression and standard deviation, respectively, in the data. The arrows indicate the improvements in expression dynamics achieved in the neural network models.

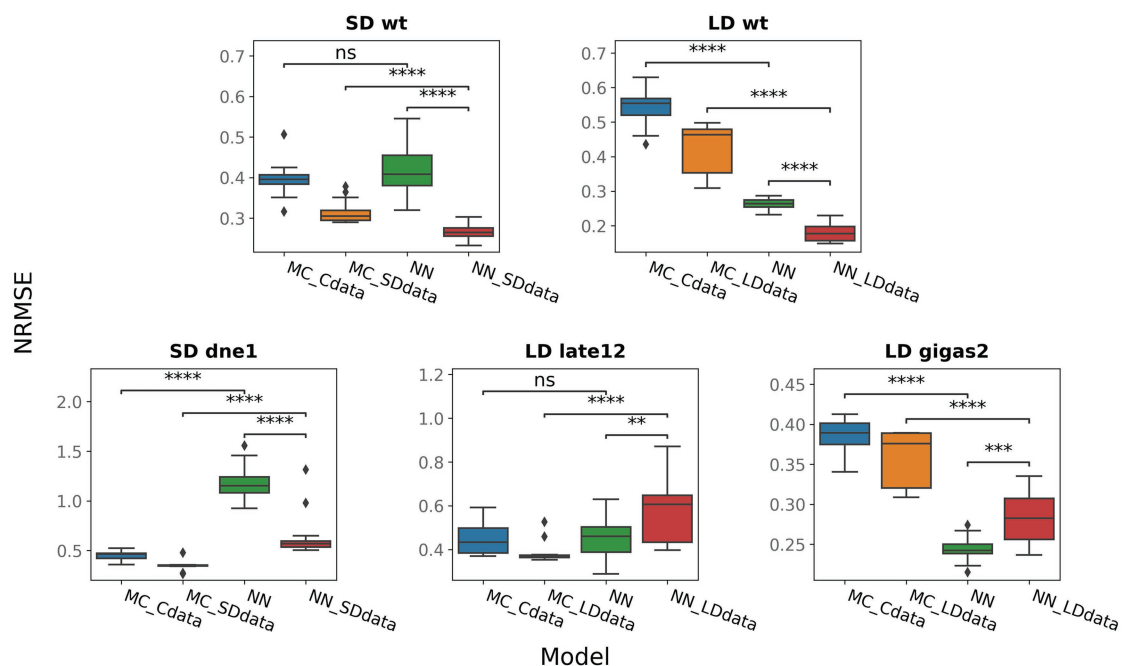
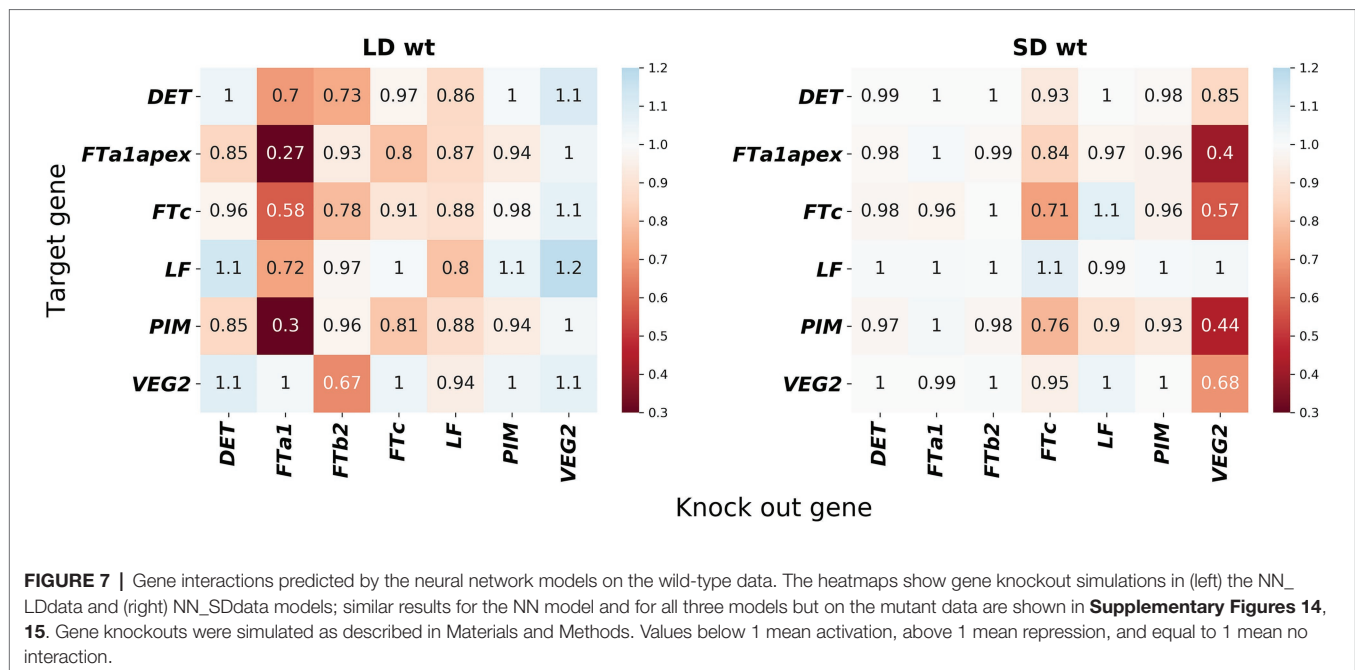


FIGURE 6 | NRMSE for neural network models (NN, NN_SDdata, and NN_LDdata) and models based on ODEs (MC_Cdata, MC_SDdata, and MC_LDdata), for different genotypes and growing conditions. NRMSE was calculated for five genes shown in Figure 5. For *gigas-2*, *PIM* was excluded from the NRMSE calculation in NN and NN_LDdata as *PIM* was also excluded from the ODE-based models by construction. The two-tailed Mann-Whitney-Wilcoxon test was applied to check the performance difference between the indicated models; *p*-values: (ns) $0.05 < p \leq 1$, (**) $0.001 < p \leq 0.01$, (***) $10^{-4} < p \leq 0.001$, (****) $p \leq 10^{-4}$.



the models. In these knockouts, we set a potential regulator protein concentration to zero in the model and calculated how the area under the dynamic expression curve changed for each potential target as the result of such perturbation. We kept the concentrations of all other proteins fixed at their values from the data during this simulation in order to estimate the direct influence of the regulator on the target, excluding possible feedbacks from other genes whose dynamics may also be altered by the perturbation. The resulted gene network topology exhibits some deviations from the proposed regulation scheme from **Figure 1B** and is qualitatively different in SD and LD (**Figure 7**). In the proposed regulation scheme, FTa1 is the only activator of the floral meristem identity gene *PIM*, while the neural network models predict FTc as an additional activator both in SD and LD. Other noticeable differences include strong *FTa1* self-activation in LD and *FTc* self-activation in SD.

Concerning differences between LD and SD, the regulatory topology exhibits more activation on the whole in LD compared to SD (**Figure 7**). Interestingly, VEG2 and FTa1 are predicted to be independent activators. VEG2 is the main activator in SD, with FTa1 almost not influencing other genes. FTa1 serves as the main activator in LD, while VEG2 is either non-active or even shows some repressive potential under this growing condition. Overall, these results show that the improvement in the solution quality demonstrated by the neural network models comes at the price of perturbations to the regulation scheme from **Figure 1B**.

DISCUSSION

The classical approach to elucidating functional regulations in a gene network consists in obtaining and qualitatively analyzing the expression patterns of genes involved in the network in

various genetic backgrounds. As more data are collected on the genes controlling floral initiation in legumes (Hecht et al., 2011; Sussmilch et al., 2015; Ridge et al., 2017; Cheng et al., 2018), more quantitative approaches are required to infer the interactions in the gene regulatory network underlying this process (Jaeger et al., 2013; Leal Valentim et al., 2015; Gursky et al., 2018). Just as it has successfully been done for *Arabidopsis*, modeling gene networks responsible for the transition to flowering in legumes can be used for testing various hypotheses about the network structure and other properties of the process, in order to better understand the mechanism or to find possible flaws in the current understanding. In this study, we elaborated several models of the core gene network involved in flowering initiation in pea and applied them to the previously obtained expression data in the wild type and in mutants. In order to make our results more robust, we used two different methods to construct models. We showed that both formalisms, ODEs and neural networks, can be utilized to formulate dynamical models suited for the gene expression data used in the study.

Our modeling results indicate that the regulation scheme that was previously proposed by analyzing the expression data qualitatively does not fully correspond to these expression data at the quantitative level. There are two types of evidence in our results for this conclusion. Firstly, the best models implementing the proposed gene regulations (the MC and MC_Cdata models) consistently generated solutions with defects in the expression dynamics of several genes. These defects comprise wrong expression dynamics of *PIM* and *FTc* in SD, *LF* in LD, and inconsistent apical expression of *FTa1*. Moreover, we showed that this picture cannot be fixed by targeted and fine-tuned modifications of the regulation scheme. The rejected alternative interactions include repression of *PIM* by *LF*, cooperative activation of *PIM* by the VEG2-FTa1 complex, and *FTa1* self-activation. As a floral meristem identity gene, *PIM* is of a special interest. One of the strongest constrains for introducing new potential activators

of *PIM* for testing in the model is in the fact that *PIM* expression is almost zero in *FTa1* mutant *gigas-2* (Hecht et al., 2011). Therefore, more complicated regulatory modules have to be devised to provide an additional activation to *PIM*, so that they can be deactivated in the absence of *FTa1*.

Secondly, gene interactions reconstructed from the data by the neural network-based dynamical models contain new regulations compared to the proposed scheme. It is interesting that one of these new regulations was *FTa1* self-activation, which was rejected at the stage of fine-tuning the proposed scheme with the help of the ODEs-based modeling. This is an example of a hypothesis about a new regulation that does not work when implemented alone but fits in when the regulation acts in concert with other modifications. Another such new regulation is *PIM* activation by *FTc*. The solution for *PIM* in the neural network model with this regulation is not zero on the *gigas-2* data but is small enough to stay within error ranges (Supplementary Figure 16), i.e., the activation by *FTc* is compensated by all other *PIM*'s regulators in the absence of *FTa1*.

Our modeling results also support the possibility that different regulatory modules are active in SD and LD. The models based on the proposed regulation scheme show the best performance when fitted to the SD and LD data separately. The use of these data in the neural network models lead to qualitatively different regulatory topologies. In SD, *VEG2* acts as the main activator, while *FTa1* does not play a significant role, and the opposite situation is observed in LD. This possible activating role of *VEG2* is in accordance with a previously obtained result showing that the model of floral initiation in *Arabidopsis* is effective under the assumption that *FD* (*VEG2*) can activate *API* (*PIM*) as a monomer (Leal Valentim et al., 2015). However, it is also possible that this *VEG2* and *FTa1* decoupling somehow reflects the activating role of the *VEG2-FTa1* complex captured by the model differently for different daylight conditions. Another finding about *VEG2* concerns cooperative binding in the formation of complexes between *VEG2* and *FT* proteins, which appears to be less favorable than the assumption about binding without constraints. This result can indicate that *FTa1*, *FTb2*, and *FTc* bind *VEG2* without essential competition.

Not all regulations predicted by the neural network approach should be considered as real, so that conclusions about those regulations should be made with caution. The inconsistencies observed in the models based on the prescribed regulation scheme most probably mean that some important regulators are missing. A nonlinear response of the gene network to the unknown dynamic expression of these unknown regulators can be encoded in spurious interactions between the genes in the current version of the network. The defects in the model solutions highlight possible genes involved in missing regulations and, thus, can be used to plan further experimental searches.

MATERIALS AND METHODS

Flowering Gene Expression Data

For model calibration, we used previously published dynamic expression data of genes responsible for flowering initiation

in pea (cultivar NGB5839; Hecht et al., 2011; Sussmilch et al., 2015). The expression data in the wild type and in mutants were extracted from the published sources using the web-based tool WebPlotDigitizer (Rohatgi, 2018). The data represent the means and SDs of the expression levels of the following genes: *FTa1*, *FTb2*, *FTc*, *DET*, *LF*, *VEG1*, *VEG2*, and *PIM*. The wild type data comprise the expression dynamics from 7 to 35th days after sowing under LD conditions and from 7 to 56th days under SD. Only LD data in the wild type were available for *VEG1*. The mutant data contain the gene expression dynamics from the mutants *dne-1* (mutation in the *DNE1* gene; 7–35 days after sowing under SD), *late1-2* (mutation in *LATE1*; 14–56 days under LD), and *gigas-2* (mutation in *FTa1*; 7–56 days under LD).

Dynamical Model Based on Differential Equations

We use the same methodology to construct the model as in Gursky et al. (2018). We model the expression of *DET*, *PIM*, *VEG1*, *LF*, *FTc*, and *FTa1* in the apex with the following set of ODEs:

$$\frac{du_{DET}}{dt} = v_1 \left(\frac{u_{VEG2} FTb2}{K_1 + u_{VEG2} FTb2} \right) - \left(\frac{K_2}{K_2 + u_{VEG1}} \right) - \lambda_1 u_{DET} \quad (1)$$

$$\frac{du_{PIM}}{dt} = v_2 \left(\frac{u_{VEG2} FTa1^n}{K_3^n + u_{VEG2} FTa1^n} \right) - \left(\frac{K_4}{K_4 + u_{DET}} \right) - \lambda_2 u_{PIM} \quad (2)$$

$$\frac{du_{VEG1}}{dt} = v_3 \left(\frac{u_{VEG2} FTa1}{K_5 + u_{VEG2} FTa1} \right) \left(\frac{K_6}{K_6 + u_{DET}} \right) - \left(\frac{K_7}{K_7 + u_{PIM}} \right) \left(\frac{K_8}{K_8 + u_{LF}} \right) - \lambda_3 u_{VEG1} \quad (3)$$

$$\frac{du_{LF}}{dt} = v_4 \frac{K_9}{K_9 + u_{VEG2} FTb2} - \lambda_4 u_{LF} \quad (4)$$

$$\frac{du_{FTc}}{dt} = v_5 \left(\frac{u_{VEG2} FTa1}{K_{10} + u_{VEG2} FTa1} + \frac{u_{VEG2} FTb2}{K_{11} + u_{VEG2} FTb2} \right) - \lambda_5 u_{FTc} \quad (5)$$

$$\frac{du_{FTa1\ apex}}{dt} = v_6 \frac{u_{VEG2} FTb2}{K_{12} + u_{VEG2} FTb2} - \lambda_6 u_{FTa1\ apex} \quad (6)$$

where u_i describes protein concentrations. The concentrations of complexes of *VEG2* with the *FT* proteins are denoted as $u_{VEG2\ FTa1}$ in the case of *FTa1*, and similarly for other *FTs*. As *FTa1* is expressed both in the leaves and in the apex, the concentration of apically expressed proteins is written as $u_{FTa1\ apex}$. The parameters v_i are the maximal

protein synthesis rates, and K_i are the Michaelis–Menten constants, which can be interpreted as equilibrium dissociation constants for regulator-promoter binding in the case of direct transcriptional regulation. The Hill constant n is used to account for the potential cooperative binding effect in *PIM* regulation by the VEG2-FTa1 complex; $n = 1$ in all versions of the model except the MM_PIM model, in which n was a free parameter. The parameters λ_i are protein degradation constants. The translation process is not explicitly considered in these equations; we assume that protein concentrations are proportional to the concentrations of corresponding mRNAs.

FTb2 in the apex comprises the protein transported from the leaves, while FTa1 in the apex additionally include the apically expressed fraction. Considering a time delay τ for the transport process, we write the total apical concentrations u_{FTa1} and u_{FTb2} of FTa1 and FTb2, respectively, as follows:

$$u_{FTa1}(t) = u_{FTa1\text{ apex}}(t) + u_{FTa1\text{ leaf}}(t - \tau) \quad (7)$$

$$u_{FTb2}(t) = u_{FTb2\text{ leaf}}(t - \tau) \quad (8)$$

where $u_{FTa1\text{ leaf}}$ and $u_{FTb2\text{ leaf}}$ are the concentrations of corresponding proteins expressed in the leaves.

The baseline model considers competitive binding between VEG2 and FTa1, FTb2, and FTc. Under equilibrium competitive binding conditions, the concentrations of VEG2 complexes with the corresponding FT proteins are as follows:

$$u_{VEG2\text{ FTa1}} = \frac{K_{13}u_{FTa1}u_{VEG2}}{1 + K_{13}u_{FTa1} + K_{14}u_{FTb2} + K_{15}u_{FTc}} \quad (9)$$

$$u_{VEG2\text{ FTb2}} = \frac{K_{14}u_{FTb2}u_{VEG2}}{1 + K_{13}u_{FTa1} + K_{14}u_{FTb2} + K_{15}u_{FTc}} \quad (10)$$

$$u_{VEG2\text{ FTc}} = \frac{K_{15}u_{FTc}u_{VEG2}}{1 + K_{13}u_{FTa1} + K_{14}u_{FTb2} + K_{15}u_{FTc}} \quad (11)$$

Therefore, the baseline model MM consists of the equations (1)–(11).

Model Modifications to Test Alternative Hypotheses

The MM_LF model is equivalent to MM but with an additional repression of *PIM* by LF, introduced by adding a repressive term into equation (2) as follows:

$$\frac{du_{PIM}}{dt} = v_2 \left(\frac{u_{VEG2\text{ FTa1}}^n}{K_3^n + u_{VEG2\text{ FTa1}}^n} \right) \left(\frac{K_4}{K_4 + u_{DET}} \right) \left(\frac{K_{16}}{K_{16} + u_{LF}} \right) - \lambda_2 u_{PIM} \quad (12)$$

The MC model is equivalent to MM but with the binding between VEG2 and FT proteins assumed to be noncompetitive. Under this assumption, the concentrations of complexes are written as follows:

$$u_{VEG2\text{ FTa1}} = u_{FTa1}u_{VEG2} \quad (13)$$

$$u_{VEG2\text{ FTb2}} = u_{FTb2}u_{VEG2} \quad (14)$$

$$u_{VEG2\text{ FTc}} = u_{FTc}u_{VEG2} \quad (15)$$

It is not necessary to add free constants of proportionality into (13)–(15), since they can be effectively scaled into free K_i already present in equations (1)–(6).

The MC_PIM model is equivalent to MC but leaves the Hill parameter n free in equation (2). This value, together with values of all other parameters, is found by parameter optimization. A value n larger than one would suggest the cooperative binding of the VEG2-FTa1 complex to the promoter of *PIM*.

The models described above were fitted to the wild type data. The MC_Cdata model is the model MC in which the values of free parameters were found by fitting to combined data, which join the wild-type data and data from *dne-1*, *late1-2*, and *gigas-2* mutants. The MC_Cdata_FTa1 model is equivalent to MC_Cdata but with added *FTa1* self-activation in the apex, which was introduced by changing equation (6) to the following one:

$$\frac{du_{FTa1\text{ apex}}}{dt} = v_6 \left(\frac{\frac{u_{VEG2\text{ FTb2}}}{K_{12} + u_{VEG2\text{ FTb2}}} + \frac{u_{VEG2\text{ FTa1}}}{K_{17} + u_{VEG2\text{ FTa1}}} \right) - \lambda_6 u_{FTa1\text{ apex}} \quad (16)$$

The MC_SDdata model is the MC model in which free parameters were found by fitting to the combined SD data, consisting of the SD part of the wild type data and *dne-1* mutant data. Similarly, MC_LDdata is the MC model fitted to the combined LD data, consisting of the LD part of the wild type data together with *late1-2* and *gigas-2* mutant data. **Supplementary Table 1** summarizes all the models investigated in the study with their main characteristics.

Numerical solutions of the model equations were obtained using the *ode23s* solver in Octave. The concentrations of all regulators on the right-hand side of the equations were replaced by data interpolated in time. The initial conditions for all proteins were set to the data values at the first day.

Parameter Optimization

The parameter values were found by minimizing the following weighted residual sum of squares (wRSS):

$$wRSS = \sum_{g=1}^N \sum_{k=1}^T \frac{(u_g(t_k) - u_g^{dat}(t_k))^2}{(u_{g,\max} - u_{g,\min})^2} \quad (17)$$

where the difference between the model solution u_g and the data u_g^{dat} for gene g is summed over all genes and times for which the data is available; $u_{g,\max}$ and $u_{g,\min}$ are the maximum and minimum concentrations in the data for gene g . Since VEG1 data was absent in SD, the numerical solution for this protein was calculated in the model but did not participate in the cost function (17). The data portion (wild type, mutant, SD, and LD) used in equation (17) depended on a model, as described above. This cost function was minimized using the DEEP software, which implements an entirely parallelized version of the differential evolution optimization method (Kozlov et al., 2016).

To reduce the number of free parameters in the models, we set $\lambda_i = 0.199$ for all proteins based on an experimental estimate of 3.49 days for the protein half-life in *Arabidopsis* grown at 20°C (Ishihara et al., 2015). To further reduce the possibility for

overfitting, we applied an ensemble approach (Samee et al., 2015; Gursky et al., 2018). The optimization for each model was performed 20 times, and the judgment about the model performance was made by analyzing the resulted distribution of the wRSS values.

We compared the models using the normalized root-mean-square deviation:

$$NRMSE = \sqrt{\frac{1}{NT} \sum_{g=1}^N \sum_{k=1}^T \frac{(u_g(t_k) - u_g^{dat}(t_k))^2}{(u_{g,max} - u_{g,min})^2}} \quad (18)$$

The Mann–Whitney–Wilcoxon test was used to compare the normalized root-mean-square error (NRMSE) distributions resulted from the parameter optimization in the models.

Neural Network Models

The neural network models were constructed as described in full details elsewhere (Podolny et al., 2020). The data set was expanded to 1,000 gene expression values per time point by sampling from normal distributions with the mean and variance taken from the initial data. The expanded data set was used for training and testing the models. The models were constructed as dynamical regression models in which the apical expression levels of six target genes (*DET*, *PIM*, *FTc*, *FTa1*, *LF*, and *VEG2*) on the current day was determined by the apical expression levels of seven genes (*DET*, *PIM*, *FTc*, *FTa1*, *LF*, *VEG2*, and *FTb2*) taken from the expanded data on the previous day.

The models were trained using the multilayer-perceptron regressor “MLPRegressor” of the Scikit-learn package (Pedregosa et al., 2011), with $f(x) = (1 + \exp(-x))^{-1}$ as the activation function and the Adam stochastic method as the parameter optimization method (Kingma and Ba, 2015). The network architecture was chosen by training the models with different topologies and picking up the best one. Each model was trained 20 times using 5-fold cross-validation, and the ensemble approach was applied for the performance analysis, as described above.

The NN model is the neural network model trained on the combined data (wild type and all mutant conditions). The NN_SDdata model was trained on the SD data (SD wild type data and mutant *dne-1* data). For these two models, the testing sample was constructed by taking data values from the last day of each separate condition. The NN_LDdata model was trained on the portion of the LD data that included the LD wild-type data and the *gigas-2* mutant data, while data from the *late1-2* mutant served as a testing sample for this model. **Supplementary Table 2** summarizes the neural models with their main characteristics.

Simulating Knockouts in Neural Network Models

In order to find out which interactions between genes are restored in the neural network models, a gene knockout analysis was

performed. The models were tested on the wild type data in which the expression of one regulator gene was set to zero. Then the AUC of a target gene expression dynamics was calculated in this simulation ($S_{\text{knock_out}}$) and in the non-perturbed case (S_{wt}). The ratio of these quantities provides information on the influence type that the regulator directly exerts on the target, as follows:

$$\frac{S_{\text{knock_out}}}{S_{\text{wt}}} > 1 \rightarrow \text{repressor}, \quad \frac{S_{\text{knock_out}}}{S_{\text{wt}}} < 1 \rightarrow \text{activator}$$

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: the source code of the models used in the study and analyzed expression data extracted from the published sources can be found in the Zenodo repository (doi: 10.5281/zenodo.4059688; <https://zenodo.org/record/4059688#.X8-ho2QzZpQ>). The source code of the DEEP software used for parameter optimization in our study can be found in the Gitlab repository (<https://gitlab.com/mackoel/deepmethod/-/tree/master>).

AUTHOR CONTRIBUTIONS

MS and VG conceived and coordinated the study. PP conducted the computational experiments and wrote the first draft of the manuscript. PP, MS, and VG analyzed the results. All authors contributed to the article and approved the submitted version.

FUNDING

The research was funded by the Ministry of Science and Higher Education of the Russian Federation as part of World-class Research Center program: Advanced Digital Technologies (contract No. 075-15-2020-934 dated 17.11.2020).

ACKNOWLEDGMENTS

We are grateful to Konstantin Kozlov for his help in setting up the optimization program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.614711/full#supplementary-material>

REFERENCES

- Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., et al. (2005). FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. *Science* 309, 1052–1056. doi: 10.1126/science.1115983
- Andrés, F., and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nat. Rev. Genet.* 13, 627–639. doi: 10.1038/nrg3291

- Barbier de Reuille, P., Routier-Kierzkowska, A., Kierzkowski, D., Bassel, G., Schüpbach, T., Tauriello, G., et al. (2015). MorphoGraphX: a platform for quantifying morphogenesis in 4D. *elife* 4:e05864. doi: 10.7554/eLife.05864
- Benlloch, R., Berbel, A., Ali, L., Gohari, G., Millán, T., and Madueño, F. (2015). Genetic control of inflorescence architecture in legumes. *Front. Plant Sci.* 6:543. doi: 10.3389/fpls.2015.00543
- Berbel, A., Ferrándiz, C., Hecht, V., Dalmais, M., Lund, O. S., Sussmilch, F. C., et al. (2012). VEGETATIVE1 is essential for development of the compound inflorescence in pea. *Nat. Commun.* 3:797. doi: 10.1038/ncomms1801
- Boudon, F., Chopard, J., Ali, O., Gilles, B., Hamant, O., Boudaoud, A., et al. (2015). A computational framework for 3D mechanical modeling of plant morphogenesis with cellular resolution. *PLoS Comp. Biol.* 11:e1003950. doi: 10.1371/journal.pcbi.1003950
- Chai, L. E., Loh, S. K., Low, S. T., Mohamad, M. S., Deris, S., and Zakaria, Z. (2014). A review on the computational approaches for gene regulatory network construction. *Comput. Biol. Med.* 48, 55–65. doi: 10.1016/j.combiomed.2014.02.011
- Cheng, X., Li, G., Tang, Y., and Wen, J. (2018). Dissection of genetic regulation of compound inflorescence development in *Medicago truncatula*. *Development* 145:dev.158766. doi: 10.1242/dev.158766
- Foucher, E., Morin, J., Courtiade, J., Cadioux, S., Ellis, N., Banfield, M. J., et al. (2003). Determinate and late flowering are two terminal flower/centroradialis homologs that control two distinct phases of flowering initiation and development in pea. *Plant Cell* 15, 2742–2754. doi: 10.1105/tpc.015701
- Goretti, D., Silvestre, M., Collani, S., Langenecker, T., Méndez, C., Madueno, F., et al. (2020). TERMINAL FLOWER 1 functions as a mobile transcriptional cofactor in the shoot apical meristem. *Plant Physiol.* 182, 2081–2095. doi: 10.1104/pp.19.00867
- Gursky, V. V., Kozlov, K. N., Nuzhdin, S. V., and Samsonova, M. G. (2018). Dynamical modeling of the core gene network controlling flowering suggests cumulative activation from the FLOWERING LOCUS T gene homologs in chickpea. *Front. Genet.* 9:547. doi: 10.3389/fgene.2018.00547
- Haque, S., Ahmad, J. S., Clark, N. M., Williams, C. M., and Sozzani, R. (2019). Computational prediction of gene regulatory networks in plant growth and development. *Curr. Opin. Plant Biol.* 47, 96–105. doi: 10.1016/j.pbi.2018.10.005
- Hecht, V., Laurie, R., Schoor, J., Ridge, S., Knowles, C., Liew, L. C., et al. (2011). The pea GIGAS gene is a FLOWERING LOCUS T homolog necessary for graft-transmissible specification of flowering but not for responsiveness to photoperiod. *Plant Cell* 23, 147–161. doi: 10.1105/tpc.110.081042
- Hofer, J., Turner, L., Hellens, R., Ambrose, M., Matthews, P., Michael, A., et al. (1997). UNIFOLIATA regulates leaf and flower morphogenesis in pea. *Curr. Biol.* 7, 581–587. doi: 10.1016/S0960-9822(06)00257-0
- Ishihara, H., Obata, T., Sulpice, R., Fernie, A. R., and Stitt, M. (2015). Quantifying protein synthesis and degradation in Arabidopsis by dynamic ¹³CO₂ labeling and analysis of enrichment in individual amino acids in their free pools and in protein. *Plant Physiol.* 168, 74–93. doi: 10.1104/pp.15.00209
- Jaeger, K. E., Pullen, N., Lamzin, S., Morris, R. J., and Wigge, P. A. (2013). Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. *Plant Cell* 25, 820–833. doi: 10.1105/tpc.113.109355
- Kardailsky, I., Shukla, V. K., Ahn, J. H., Dagenais, N., Christensen, S. K., Nguyen, J. T., et al. (1999). Activation tagging of the floral inducer FT. *Science* 286, 1962–1965. doi: 10.1126/science.286.5446.1962
- Kaufmann, K., Wellmer, F., Muiño, J. M., Ferrier, T., Wuest, S. E., Kumar, V., et al. (2010). Orchestration of floral initiation by APETALA1. *Science* 328, 85–89. doi: 10.1126/science.1185244
- Khan, M., Ai, X., and Zhang, J. (2014). Genetic regulation of flowering time in annual and perennial plants. *Wiley Interdiscip. Rev. RNA* 5, 347–359. doi: 10.1002/wrna.1215
- Kingma, D. P., and Ba, J. L. (2015). “Adam: A method for stochastic optimization” in *ICLR 2015: Proceedings of 3rd International Conference for Learning Representations*; May 7–9, 2015; 1–15.
- Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M., and Araki, T. (1999). A pair of related genes with antagonistic roles in mediating flowering signals. *Science* 286, 1960–1962. doi: 10.1126/science.286.5446.1960
- Kozlov, K., Samsonov, A. M., and Samsonova, M. (2016). A software for parameter optimization with differential evolution entirely parallel method. *PeerJ Comput. Sci.* 2016, 1–20. doi: 10.7717/peerj-cs.74
- Leal Valentim, F., Mourik, S. V., Posé, D., Kim, M. C., Schmid, M., van Ham, R. C. H. J., et al. (2015). A quantitative and dynamic model of the Arabidopsis flowering time gene regulatory network. *PLoS One* 10:e0116973. doi: 10.1371/journal.pone.0116973
- Lee, J., Oh, M., Park, H., and Lee, I. (2008). SOC1 translocated to the nucleus by interaction with AGL24 directly regulates LEAFY. *Plant J.* 55, 832–843. doi: 10.1111/j.1365-3113X.2008.03552.x
- Le Novère, N. (2015). Quantitative and logic modelling of molecular and gene networks. *Nat. Rev. Genet.* 16, 146–158. doi: 10.1038/nrg3885
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.
- Podolny, B., Gursky, V., and Samsonova, M. (2020). A machine-learning analysis of flowering gene expression in the CDC frontier chickpea cultivar. *Biophysics* 65, 225–236. doi: 10.1134/S0006350920020189
- Pullen, N., Jaeger, K. E., Wigge, P. A., and Morris, R. J. (2013). Simple network motifs can capture key characteristics of the floral transition in Arabidopsis. *Plant Signal. Behav.* 8:e26149. doi: 10.4161/psb.26149
- Ridge, S., Deokar, A., Lee, R., Daba, K., Macknight, R. C., Weller, J. L., et al. (2017). The chickpea early flowering 1 (Efl1) locus is an ortholog of arabidopsis ELF3. *Plant Physiol.* 175, 802–815. doi: 10.1104/pp.17.00082
- Rohatgi, A. (2018). *WebPlotDigitizer [Internet]. Version 4.1. Austin, Texas (USA).* Available at: <https://automeris.io/WebPlotDigitizer> (Accessed September 01, 2019).
- Samee, M. A. H., Lim, B., Samper, N., Lu, H., Rushlow, C. A., Jiménez, G., et al. (2015). A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. *Cell Syst.* 1, 396–407. doi: 10.1016/j.cels.2015.12.002
- Srikanth, A., and Schmid, M. (2011). Regulation of flowering time: all roads lead to Rome. *Cell. Mol. Life Sci.* 68, 2013–2037. doi: 10.1007/s00018-011-0673-y
- Sussmilch, F. C., Berbel, A., Hecht, V., Schoor, K. V., Ferrándiz, C., Madueno, F., et al. (2015). Pea VEGETATIVE2 is an FD homolog that is essential for flowering and compound inflorescence. *Development* 142, 1046–1060. doi: 10.1105/tpc.115.136150
- Taoka, K. I., Ohki, I., Tsuji, H., Furuita, K., Hayashi, K., Yanase, T., et al. (2011). 14-3-3 proteins act as intracellular receptors for rice Hd3a florigen. *Nature* 476, 332–335. doi: 10.1038/nature10272
- Taylor, S. A., Hofer, J. M., Murfet, I. C., Sollinger, J. D., Singer, S. R., Knox, M. R., et al. (2002). PROLIFERATING INFLORESCENCE MERISTEM, a MADS-box gene that regulates floral meristem identity in pea. *Plant Physiol.* 129, 1150–1159. doi: 10.1104/pp.001677
- Tsuji, H., Nakamura, H., Taoka, K. I., and Shimamoto, K. (2013). Functional diversification of FD transcription factors in rice, components of florigen activation complexes. *Plant Cell Physiol.* 54, 385–397. doi: 10.1093/pcp/pct005
- Wang, C., Chang, P., Ng, K., Chang, C., Sheu, P., and Tsai, J. (2014). A model comparison study of the flowering time regulatory network in Arabidopsis. *BMC Syst. Biol.* 8:15. doi: 10.1186/1752-0509-8-15
- Welch, S. M., Roe, J. L., and Dong, Z. (2003). A genetic neural network model of flowering time control in *Arabidopsis thaliana*. *Agron. J.* 95, 71–81. doi: 10.2134/agronj2003.0071
- Wenden, B., Dun, E. A., Hanan, J., Andrieu, B., Weller, J. L., Beveridge, C. A., et al. (2009). Computational analysis of flowering in pea (*Pisum sativum*). *New Phytol.* 184, 153–167. doi: 10.1111/j.1469-8137.2009.02952.x
- Wenden, B., and Rameau, C. (2009). Systems biology for plant breeding: the example of flowering time in pea. *C. R. Biol.* 332, 998–1006. doi: 10.1016/j.crv.2009.09.011
- Wigge, P. A., Kim, M. C., Jaeger, K. E., Busch, W., Schmid, M., Lohmann, J. U., et al. (2005). Integration of spatial and temporal information during floral induction in Arabidopsis. *Science* 309, 1056–1059. doi: 10.1126/science.1114358
- Wu, F., Kang, X., Wang, M., Haider, W., Price, W. B., Hajek, B., et al. (2019). Transcriptome-enabled network inference revealed the GmCOL1 feed-forward loop and its roles in photoperiodic flowering of soybean. *Front. Plant Sci.* 10:1221. doi: 10.3389/fpls.2019.01221
- Yoo, S. K., Chung, K. S., Kim, J., Lee, J. H., Hong, S. M., Yoo, S. J., et al. (2016). CONSTANS activates SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 through FLOWERING LOCUS T to promote flowering in Arabidopsis. *Plant Physiol.* 139, 770–778. doi: 10.1104/pp.105.066928

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Pavlinova, Samsonova and Gursky. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



The Links of Ghrelin to Incretins, Insulin, Glucagon, and Leptin After Bariatric Surgery

Daria Skuratovskaia^{1*}, Maria Vulf¹, Nataliya Chasovskikh², Aleksandra Komar¹, Elena Kirienkova¹, Egor Shunkin¹, Pavel Zatolokin¹ and Larisa Litvinova¹

¹ Center for Immunology and Cellular Biotechnology, Immanuel Kant Baltic Federal University, Kaliningrad, Russia,

² Department of Medical and Biological Cybernetics, Siberian State Medical University, Tomsk, Russia

OPEN ACCESS

Edited by:

Yuriy L. Orlov,
I.M. Sechenov First Moscow State
Medical University, Russia

Reviewed by:

Dmitriy N. Atochin,
Harvard Medical School,
United States
Anastasia Efimenko,
Lomonosov Moscow State University,
Russia

*Correspondence:

Daria Skuratovskaia
DariaSK@list.ru

Specialty section:

This article was submitted to
Computational Genomics,
a section of the journal
Frontiers in Genetics

Received: 30 September 2020

Accepted: 15 March 2021

Published: 20 April 2021

Citation:

Skuratovskaia D, Vulf M,
Chasovskikh N, Komar A,
Kirienkova E, Shunkin E, Zatolokin P
and Litvinova L (2021) The Links
of Ghrelin to Incretins, Insulin,
Glucagon, and Leptin After Bariatric
Surgery. *Front. Genet.* 12:612501.
doi: 10.3389/fgene.2021.612501

Type 2 diabetes mellitus (T2DM) is one of the most prominent and socially significant problems. The present study aimed to identify the mechanisms of interaction of critical regulators of carbohydrate metabolism using bioinformatics and experimental methods and to assess their influence on the development of T2DM. We conducted an *in silico* search for the relationship of hormones and adipokines and performed functional annotation of the receptors for ghrelin and incretins. Hormones and adipokines were assessed in the plasma of obese patients with and without T2DM as well as after laparoscopic sleeve gastrectomy (LSG) and Roux-en-Y gastric bypass (RYGB) surgeries. Incretin- and ghrelin-associated functions and metabolic processes were discovered. Low ghrelin levels were observed in obese patients without T2DM compared with healthy volunteers and the other groups. The highest ghrelin levels were observed in obese patients with T2DM. This defense mechanism against insulin resistance could be realized through the receptors G-protein-coupled receptor (GPCR), growth hormone secretagogue receptor (GHSR), and growth hormone-releasing hormone receptor (GHRHR). These receptors are associated with proliferative, inflammatory, and neurohumoral signaling pathways and regulate responses to nutrient intake. Signaling through the GPCR class unites ghrelin, glucagon, glucose-dependent insulinotropic polypeptide (GIP), and glucagon-like peptide (GLP)-1. Ghrelin impairs carbohydrate and lipid metabolism in obese patients. Ghrelin is associated with elevated plasma levels of insulin, glucagon, and leptin. Specific activation of receptors and modulation by posttranslational modifications of ghrelin can control IR's development in obesity, which is a promising area for research.

Keywords: gene annotation, type 2 diabetes mellitus, bariatric surgery, GHSR, GIPR, GLP-1r, adipokines, obesity

INTRODUCTION

Type 2 diabetes mellitus (T2DM) is one of the most prominent and socially significant problems (Dedov et al., 2017; Statistics About Diabetes | ADA, 2020). T2DM and associated diseases, in particular, abdominal obesity, occupy a leading position among the causes of mortality in the population (WHO | Raised fasting blood glucose, 2020). T2DM is characterized by the impaired metabolic response of insulin-dependent tissue [e.g., muscle, adipose tissue (AT), liver] to insulin,

which leads to an increase in its concentration in human plasma (Finan et al., 2013). It is known that gastroduodenal zone hormones and mediators of AT, namely, adipokines, regulate carbohydrate metabolism components and play an important role in the pathogenesis of IR in T2DM (Finan et al., 2013; Vejrazkova et al., 2017). Different parts of the intestine secrete many hormones, and the key role belongs to incretin glucose-dependent insulinotropic polypeptide (GIP) and glucagon-like peptide-1 (GLP-1) (Skow et al., 2016; Brandt et al., 2018).

The main function of incretins is to stimulate insulin secretion by the pancreatic β -cells of the islets of Langerhans (Yabe and Seino, 2011; Campbell et al., 2016). GLP-1 promotes the normalization of carbohydrate metabolism and decreases body mass index (BMI) (Hong et al., 2016), while GIP has multidirectional effects on carbohydrate metabolism. The role of GIP in incretins is still controversial (Finan et al., 2013). It has been demonstrated that incretins are interrelated with the regulation of leptin and ghrelin production, the main modulators of carbohydrate metabolism (Kim and Egan, 2008; Karim et al., 2015; Ronveaux et al., 2015). Ghrelin is an orexigenic hormone that increases ingestion by activating agouti-related peptide (AgRP)/neuropeptide Y (NPY) neurons (Yanagi et al., 2018). Fasting plasma ghrelin level increases, then ghrelin activates gluconeogenesis in the liver, suppresses insulin production, and maintains glucose levels within control parameters. Postprandial hormone production decreases when there is a glucose-stimulated increase in insulin secretion.

In addition to metabolic functions, leptin is a pleiotropic inflammatory mediator and modulates glucose homeostasis and insulin release by reducing glucagon secretion (Hong et al., 2016; Vejrazkova et al., 2017). Glucagon stimulates glucose production in the liver, preventing hypoglycemia under normal physiological conditions; hyperglucagonemia is an indicator of T2DM (Vejrazkova et al., 2017; Brandt et al., 2018). It has been established that resistin also promotes the development of IR (Stofkova, 2010; Vejrazkova et al., 2017). Visfatin has pro-inflammatory and immunomodulatory properties and has insulin-sensitizing and insulin-mimetic effects. Thus, visfatin is of interest as a possible target for modulating blood glucose (Stofkova, 2010; Vejrazkova et al., 2017).

Incretin-stimulated insulin secretion accounts for approximately 50% of the total insulin production (Kim and Egan, 2008). Insulin biosynthesis and secretion are closely related to incretin receptors (Kim and Egan, 2008). In patients with T2DM, the sensitivity of cells to insulin is reduced, and glucose-dependent secretion of insulin is impaired (Nielsen et al., 2015). It has been established that in patients with T2DM, the absence and decrease in response to incretin therapy may be associated with dysregulation of expression or defects in incretin receptors (Yabe and Seino, 2013). Therefore, incretins play an important role in the regulation of insulin production. Consequently, studies of the causes of impaired secretion of incretins and decreased insulin-dependent receptor sensitivity in T2DM are relevant.

The present study aimed to identify the mechanisms of interaction of critical regulators of carbohydrate metabolism using bioinformatics and experimental methods, and

to assess their influence on the development of T2DM complicated by obesity.

MATERIALS AND METHODS

Experimental Research Methods

The study included 225 obese patients. Of these, 113 obese patients had T2DM (45.18 ± 8.29 years; 45.69 ± 10.51 kg/m²; 46 men and 67 women), and 115 obese patients did not have carbohydrate metabolism disorders (46.41 ± 9.3 years; 46.31 ± 7.56 kg/m²; 41 men and 74 women). The presence of arterial hypertension was noted in 43% of patients. Obese patients with T2DM underwent surgery with two types of surgical treatment: laparoscopic sleeve gastrectomy (LSG) (48.53 ± 6.13 years; 40.59 ± 6.55 kg/m²) and Roux-en-Y gastric bypass (RYGB) (46.08 ± 10.63 years; 33.52 ± 6.08 kg/m²). The results of these patients were recorded 6 months after surgery. The control group of healthy volunteers included 102 apparently healthy donors with normal anthropometric and biochemical parameters (39.3 ± 6.51 years; 22.82 ± 2.18 kg/m²; 61 men and 41 women). Venous blood was obtained before and 60 min after the test breakfast. In the test breakfast, the protein content was 9.1 g, the carbohydrate content was 88.1 g, and the fat content was 10.6 g. In 45% of the patients, essential hypertension was diagnosed according to the classification of arterial hypertension by the arterial pressure level (EHS/ESC 2003–2013).

All the study participants provided informed consent to participate in the research study. The study was carried out according to the World Medical Association (WMA) Declaration of Helsinki (2000) and the Protocol to the Convention on Human Rights and Biomedicine (1999). The Local Ethics Committee approved the study protocol of the Innovation Park of the Immanuel Kant Baltic Federal University (protocol no. 4 from October 23, 2013).

The analysis of biochemical parameters in the blood serum was carried out on a Furuno CA-180 automatic biochemical analyzer (Furuno Electric Company, Japan) using DiaSys test systems (DiaSys Diagnostic Systems, Germany). Plasma hormone levels were assessed by flow fluorimetry (Bio-Plex Protein Assay System, Bio-Rad, United States) using commercial test systems (Bio-Plex Pro Human Diabetes 10-Plex Assay, Bio-Rad, United States). In obese patients and healthy donors, the concentrations of mediators (hormones: ghrelin, GIP, GLP-1, insulin, C-peptide, and glucagon) and the mediator SERPINE1 (PAI-1) and adipokines (resistin, leptin, and visfatin) were assessed in plasma.

A block diagram of the research is presented in **Figure 1**.

Bioinformatic Research Methods

Interactions of the studied proteins (nodes: proteins; edges: cooperations between them) were evaluated using the Cytoscape version 3.2.1 (United States) network.

Initial data on protein–protein interactions were obtained from the Human Protein Reference Database (HPRD). Protein–protein interactions from the HPRDs were used to construct a network of interactions for the proteins under study

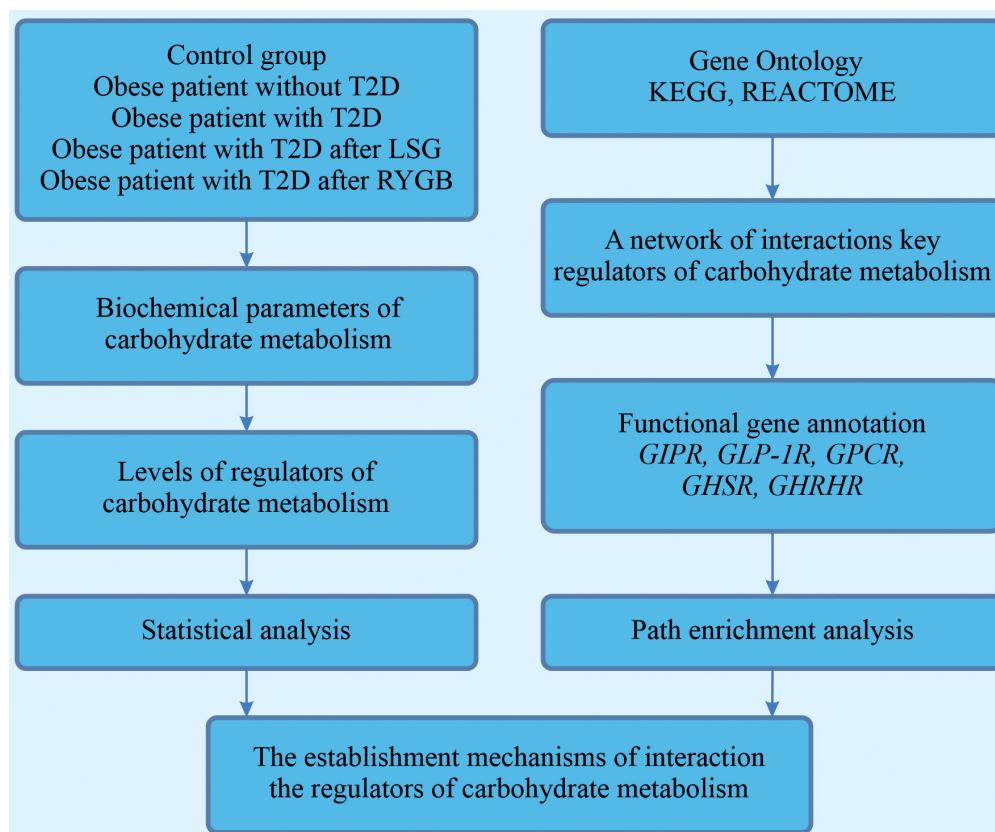


FIGURE 1 | A block-diagram of the research.

(Keshava Prasad et al., 2009). In the resulting network, proteins from the HPRDs interacting with these proteins and common bonds were identified. All studied proteins were applied to the protein–protein interaction HPRD network and then extracted together with the associated proteins.

The functional annotation of the gastric inhibitory polypeptide receptor (GIPR) and glucagon-like peptide-1 receptor (GLP-1R) genes and the analysis of their representation in the signaling and metabolic pathways were carried out using the algorithm implemented in the ClueGO Cytoscape version 3.2.1 plugin (Bindea et al., 2009) based on the use of the hypergeometric test ($p < 0.05$). The value of the kappa statistic, reflecting the functional relationships between genes, was set at 0.4. The functional characterization of genes was carried out based on the terminology of Gene Ontology within the categories of “biological process” and “molecular function” (Ashburner et al., 2000; Rivals et al., 2007). In this study, the functions were described from the third to the eighth levels of the hierarchy. The representation of genes in signaling and metabolic pathways was determined using the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway and Reactome pathway analyses.

Statistical Analysis of Experimental Data

The normal distribution of quantitative indicators was checked using the Kolmogorov–Smirnov test. If the normal law of

distributing a feature in the studied samples was consistent, the hypothesis about the average sample values’ equality was tested using Student’s t test. If the data distribution did not obey the normal distribution law, further assessment of the sample differences was calculated using the non-parametric Mann–Whitney test for pairwise comparisons. According to the Spearman method, a relationship between the studied parameters was carried out using correlation analysis. Differences were considered significant at a significance level of $p < 0.05$. Statistical processing of the obtained results was carried out using the R statistical software (version 3.3.1).

RESULTS

Ghrelin Is Related to Insulin

Body mass index and biochemical parameters of carbohydrate metabolism (glucose and insulin) and lipid metabolism (cholesterol, triglycerides, HDL, and LDL) were measured in obese patients. Predictably, we observed a disturbance in carbohydrate and lipid metabolism parameters in obese patients with T2DM (**Supplementary Table 1**). We investigated a comparison group of obese patients without T2DM to look for the involvement of mediators in maintaining normal glucose levels. These obese patients had a high BMI, but carbohydrate

and lipid metabolism were within normal limits, and hormone levels were comparable or lower than healthy volunteers. We did not find any sex differences in parameters between groups.

Disorders of carbohydrate and lipid metabolism can inhibit various compensatory mechanisms. A likely compensation mechanism could be an abnormal decrease in the total ghrelin level in obese patients without T2DM relative to healthy volunteers and obese patients with T2DM (**Figure 2**).

We analyzed patients after LSG surgery to determine the importance of lowering ghrelin levels in maintaining normal carbohydrate metabolism in obese patients. This surgery removes the fundus of the stomach (the main area of food addiction) that produces ghrelin. Basal ghrelin level was lower in patients operated on LSG than that in healthy donors and in patients before surgery. Thus, a low plasma total ghrelin level is a compensatory mechanism for IR in obese patients.

After eating, ghrelin levels increased in patients after LSG surgery. This may seem contradictory if we do not consider the correlation results and data from Gene Ontology and compare them in patients with T2DM before and after LSG surgery. The ghrelin level was positively correlated with insulin before surgery ($r = 0.420$) (**Figure 3C**) and negatively correlated in patients after LSG surgery ($r = -0.392$) ($p < 0.05$) (**Figure 3D**). We showed *in silico* that ghrelin was associated with insulin, leptin, glucagon, and CRP (**Figure 4**). In particular, we found *in silico* connections between ghrelin and insulin through several pathways: GHRL-MLNR-GPRASP1-LRP2-INS, GHRL-HK3-LEP-A2M-CTSE-INS, GHRL-HK3-LEP-CLU-CPE-INS, and GHRL-HK3-LEP-LEPR-CLU-CPE-INS (**Figure 4**). These pathways can be activated depending on the microenvironment. The high levels of insulin after breakfast and the associated mediators stimulated ghrelin production in an endocrine manner.

Ghrelin level negatively correlated with BMI and glucagon in patients with T2DM (**Figure 3C**) and with insulin after LSG (**Figure 3D**). Stimulation of ghrelin can occur not only by irritating the stomach's fundus but also at the endocrine level.

Ghrelin Is Associated With Incretins

Laparoscopic sleeve gastrectomy is an effective weight loss surgery with the removal of the central ghrelin production gastric zone. Ghrelin levels were negatively correlated with BMI and decreased in patients after surgery. Ghrelin levels were lower in patients after LSG than in patients after RYGB. The reason for the differences is the peculiarities of the surgery themselves. In patients after surgery, BMI decreased, and lipid metabolism indicators returned to normal but not glucose and CRP levels. Despite this, the patients, after LSG and RYGB, are characterized by low insulin levels.

The association of ghrelin and incretins confirms the revealed positive correlations between ghrelin and GIP in obese patients with T2DM (**Figure 3C**); in patients after LSG, positive correlations were noted between ghrelin and GLP-1 (**Figure 3D**). *In silico*, ghrelin has been shown to bind GLP-1 (**Figure 4**) via insulin or glucagon. In this regard, ghrelin is associated with GLP-1 and may affect GLP-1 levels after LSG in obese patients ($r = 0.401$) (**Figure 3**).

Functional Annotation of Ghrelin and Its Receptors

The effects of ghrelin are multidirectional and depend on its isoforms and the activation of the corresponding receptors. There are several receptors for ghrelin. With the functional annotation of these receptor names, only one was found: G protein-coupled receptor (GPCR). This gene is designated FZD4 in the search results. The following functions were revealed in GO terms (**Supplementary Table 4**). The participation of FZD4 in the regulation of the morphogenesis of the organs of vision and the brain and the regulation of the secretion of steroid hormones, including progesterone, has been shown.

Enrichment analysis of the pathways for this receptor showed that GPCR (FZD4) belongs to the following signaling pathways:

R-HSA: 4641263 Regulation of FZD by ubiquitination;
R-HSA: 5099900 WNT5A-dependent internalizations of FZD4;
R-HSA: 5340588 RNF mutants show enhanced WNT signaling and proliferation.

Figure 4 shows that ghrelin is associated with the growth hormone secretagogue receptor (GHSR). The functional annotation of the GHSR gene is provided in the **Supplementary Material (Supplementary Table 5)**. GHSR is associated with food responses and negative appetite regulation. In the gastrointestinal tract, the GHSR receptor is associated with food transit through the intestine and intestinal musculature contraction. GHSR is associated with the synthesis of neurotransmitters, norepinephrine, and catecholamines to synthesize growth hormones. GHSR negatively regulates the apoptosis of macrophages and myeloid cells and is associated with blocking inflammation: it inhibits tumor necrosis factor- α (TNF- α) and interleukin-6 (IL-6) functions.

Ghrelin is also associated with the growth hormone-releasing hormone receptor (GHRHR) (**Figure 4**). Functional annotation of GHRHR revealed the functions (signaling pathways were also not identified) and are listed in **Supplementary Table 8**. GHRHR was associated with the activation of the growth hormone-releasing receptor, regulating the production of growth hormone, and insulin-like growth factor. Additionally, receptor activation has been associated with the regulation of the sleep/wake cycle.

Thus, activation of GHSR has protective effects on carbohydrate and lipid metabolism.

Functional Annotation of the Gastric Inhibitory Polypeptide Receptor and Glucagon-Like Peptide-1 Receptor Incretin Receptors

Incretins mediate their effects through the receptors GIPR and GLP-1R. According to our earlier data (Skuratovskaia, 2018), a special role in the disturbance of carbohydrate metabolism in obesity is played by the occurrence of polymorphisms

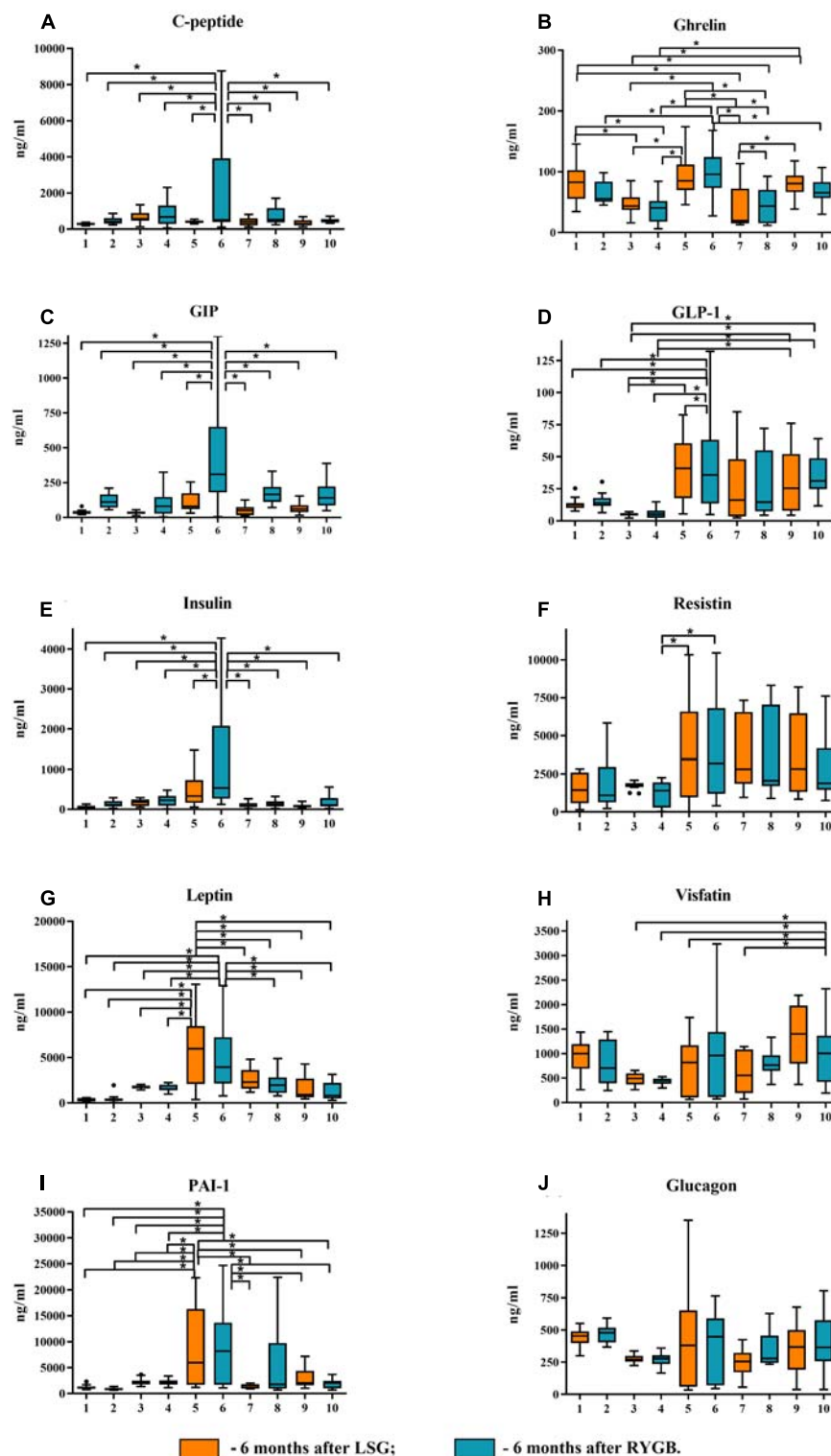


FIGURE 2 | Plasma levels of the studied mediators in obese patients with type 2 diabetes mellitus (T2DM) before and after laparoscopic sleeve gastrectomy (LSG) and Roux-en-Y gastric bypass (RYGB). **(A)** Plasma level of C-peptide. **(B)** Plasma level of ghrelin. **(C)** Plasma level of GIP. **(D)** Plasma level of GLP-1. **(E)** Plasma level of insulin. **(F)** Plasma level of resistin. **(G)** Plasma level of leptin. **(H)** Plasma level of visfatin. **(I)** Plasma level of PAI-1. **(J)** Plasma level of glucagon. * $p < 0.05$; differences in significance level were determined using one-way ANOVA. 1 – healthy volunteers before breakfast; 2 – healthy volunteers after breakfast; 3 – obese patient without T2DM before breakfast; 4 – obese patient without T2DM after breakfast; 5 – obese patient with T2DM before breakfast; 6 – obese patient with T2DM after breakfast; 7 – 6 months after LSG before breakfast; 8 – 6 months after LSG after breakfast; 9 – 6 months after RYGB before breakfast; 10 – 6 months after RYGB after breakfast.

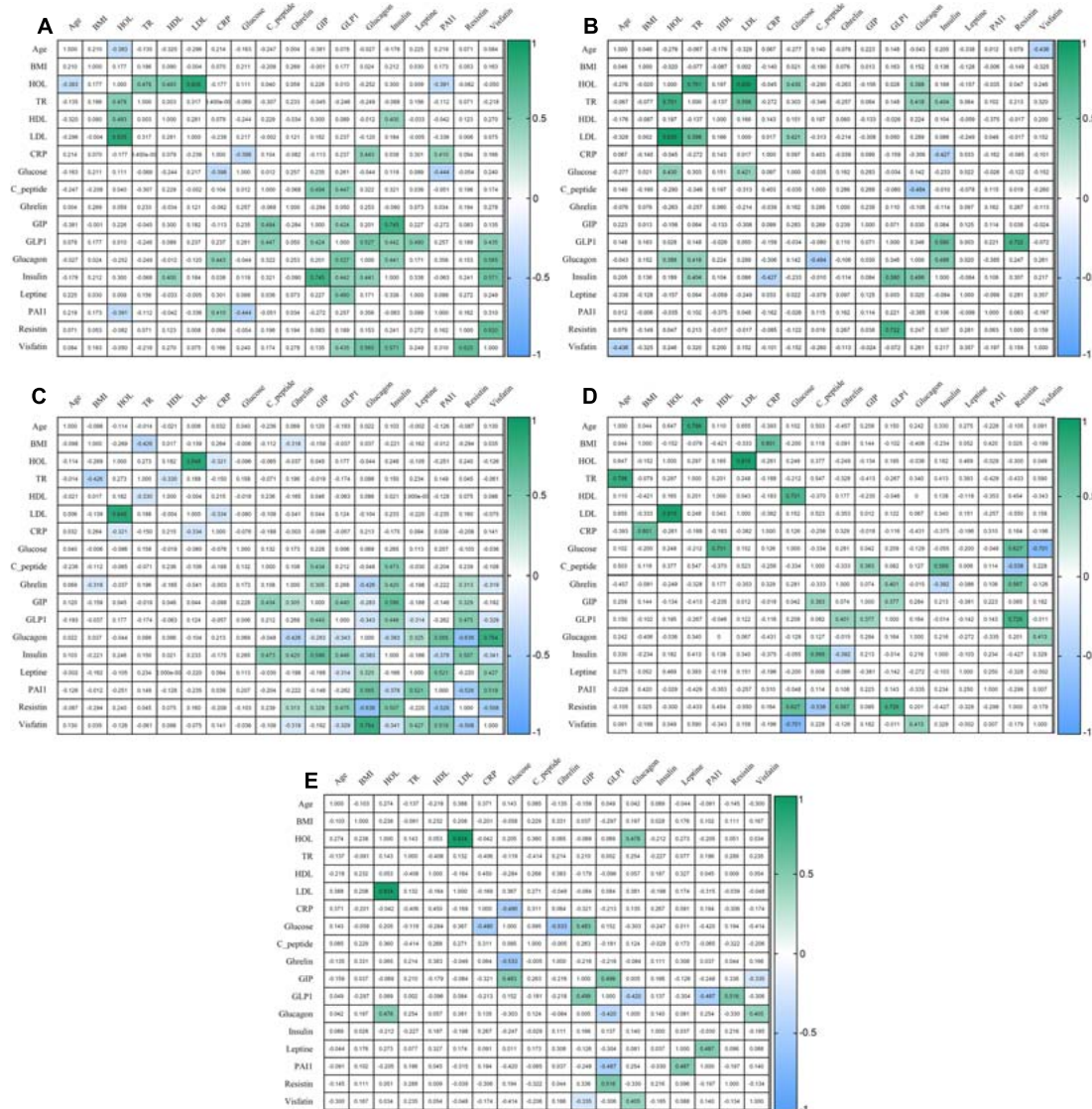


FIGURE 3 | Correlation of mediators. (A) Correlation in the healthy volunteers; (B) Correlation in an obese patient without T2DM; (C) Correlation in an obese patient with T2DM; (D) Correlation in patients 6 months after LSG; (E) Correlation in patients 6 months after RYGB. The analysis was performed using Pearson's test correlation.

in the GIPR and GLP-1R genes. To identify possible mechanisms of these genes' participation in metabolic processes, functional annotation and analysis of their involvement in metabolic and signaling pathways were carried out (Supplementary Table 9).

Based on the functional annotation results, the GIPR gene was assigned to the following terms in the Gene Ontology category "biological process": realization of mechanisms of development of many diseases, including those associated with digestion processes. GIPR has been associated with the development of a response to fatty acids, as well as with the processes responsible for the development of the pancreas from its formation to a mature structure, including islet cells that produce insulin, glucagon, and somatostatin.

Molecular functions corresponding to these biological processes were also identified: selective non-covalent interaction with a protein with hormonal activity (GO: 0017046, peptide hormone binding) and activity of the GIP receptor, that is, binding to GIP and signal transduction across the membrane to activate the G-protein (GO: 0016519, gastric inhibitory peptide receptor activity) (Usdin et al., 1993).

The functional annotation of GLP-1R was performed, which showed the participation of the GLP-1R gene in the regulation of adenylate cyclase activity (category "biological process") (Supplementary Table 6).

In this regard, the effects of GLP-1R are not only implemented through glucose-dependent stimulation of insulin but are also

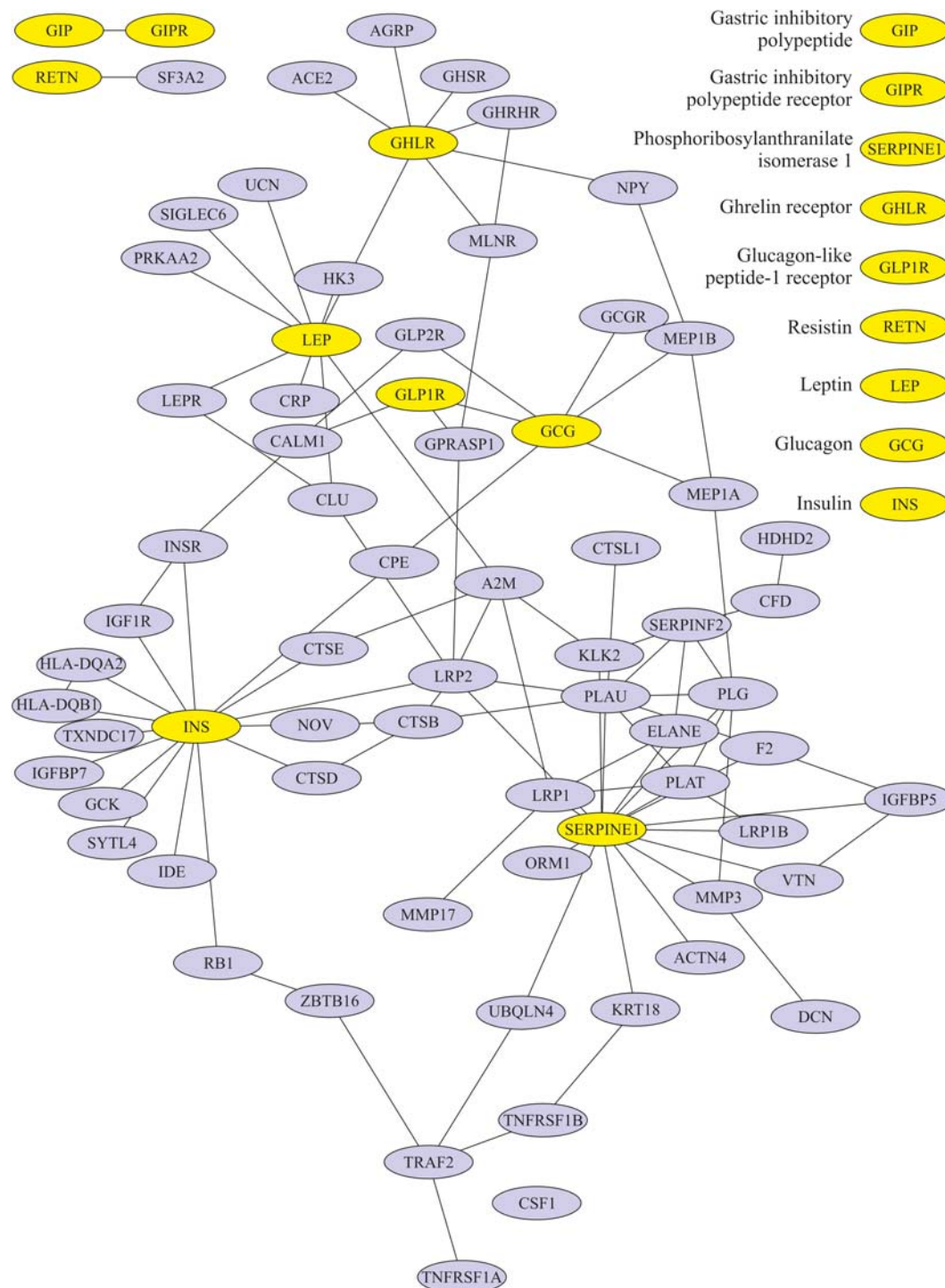


FIGURE 4 | Subnetwork of the proteins insulin, ghrelin, glucagon-like peptide-1 (GLP-1), leptin, and their nearest neighbors. Light nodes represent the proteins under study; dark nodes are their nearest neighbors.

associated with many other metabolic processes. Thus, GLP-1R activation can participate in the regulation of adenylate cyclase activity and cAMP-mediated signaling. Additionally, the relationship between GLP-1R and glucagon was shown, which

was reflected in experimental studies: the level of glucagon positively correlated with GLP-1 in patients in the healthy donors ($r = 0.527$, $p < 0.05$) and negatively correlated with GLP-1 in patients with T2DM ($r = -0.343$, $p < 0.05$).

Contribution of C-Peptide, Insulin, Leptin, PAI, Glucose-Dependent Insulinotropic Polypeptide, Glucagon-Like Peptide-1, Ghrelin, Resistin, and Visfatin to the Formation of Insulin Resistance

The data obtained indicate a change in carbohydrate and lipid metabolism, as the content of cholesterol, triglycerides, and LDL increased in both groups of patients (**Supplementary Table 1**).

The levels of hormones of the gastroduodenal zone, adipokines, and PAI-1 on an empty stomach (odd numbers of groups) and after a test breakfast (even numbers of groups) were studied in the following groups of patients: 1 and 2 – healthy volunteers, 3 and 4 – a comparison group of obese patients without T2DM, 5 and 6 – obese patients with T2DM, 7 and 8 – obese patients with T2DM 6 months after LSG, and 9 and 10 – obese patients with T2DM 6 months after RYGB (**Figure 2**). We measured mediator levels on an empty stomach and 60 min after the test breakfast, as the intensity of their production depends on the amount of nutrient intake.

Postprandial dynamics of mediators were observed only in patients with T2DM (groups 5 and 6). Significant differences between hormones (ghrelin, C-peptide, GIP, GLP-1, insulin, leptin, and resistin) and PAI-1 from other study groups were found in patients with T2DM. Thus, in patients with T2DM on an empty stomach and/or after a test breakfast, C-peptide, insulin, GIP, GLP-1, glucagon, ghrelin, leptin, PAI-1, and resistin were higher than those in obese patients without T2DM. The same changes in C-peptide levels, insulin, leptin, PAI, and GIP were revealed; the maximum value was found in patients with T2DM after a test breakfast compared with the other groups.

Simultaneously, changes in the concentration of these mediators (C-peptide, insulin, GIP, GLP-1, glucagon, ghrelin, leptin, PAI-1, resistin, and visfatin) before and after the test breakfast did not depend on food intake. The GLP-1 and ghrelin levels reached maximum values in patients with T2DM relative to the other groups, but their change was associated with food intake.

In patients without T2DM, significant differences from the healthy volunteers were revealed only with respect to the ghrelin level, which was significantly lower. In general, in obese patients without T2DM, C-peptide, insulin, leptin, PAI, GIP, GLP-1, ghrelin, and resistin were significantly lower than those in obese patients with T2DM.

We investigated hormone and adipokine levels in patients with T2DM 6 months after LSG and RYGB bariatric surgeries.

In our study, all patients had hyperglycemia, normalization of lipid metabolism, and decreased body weight after 6 months.

Thus, in patients after LSG and RYGB surgeries, there was a decrease in C-peptide levels, insulin, leptin, PAI, GIP, GLP-1, ghrelin, resistin, and visfatin compared with obese patients with T2DM before surgery (**Figure 2**). After LSG, it was found that the level of ghrelin on an empty stomach and after breakfast was significantly lower than that in the other groups. The ghrelin level after the test breakfast was higher than the fasting values in LSG patients.

After RYGB, it was found that the level of GLP-1 did not change relative to the obese patients with T2DM before surgery, and the level of GIP after the test breakfast was lower compared with obese patients with T2DM (**Figure 2**). However, after RYGB, fasting, and the test breakfast, GLP-1 levels were higher than those in obese patients without T2DM (**Figure 2**).

The ghrelin level in patients after RYGB was higher than that in patients after LSG but lower than that in patients with T2DM before surgery (**Figure 2**). This finding indicates the regulatory role of these hormones on plasma glucose.

Visfatin levels changed only after RYGB. The concentration of visfatin in patients after RYGB was higher than obese patients without T2DM and compared with obese patients with T2DM (**Figure 2**).

Thus, C-peptide, insulin, leptin, PAI, GIP, GLP-1, ghrelin, resistin, and visfatin are closely related to carbohydrate and lipid metabolism and BMI. It contributes to the formation of IR in obese patients. However, the nature of these relationships is unclear.

Enrichment Assay for GIPR, GLP-1R, Insulin, Ghrelin, GIP, GLP-1, Leptin, and Resistin

Identifying the functions of the studied genes correlates with the data on participation in the corresponding processes of signal transduction in cells. The following pathways were identified in which the *GIPR* and *GLP-1R* genes are involved: the cAMP signaling pathway, GPCR ligand binding, G alpha(s) signaling events, and glucagon-type ligand receptors (**Supplementary Table 9**).

Thus, activation of GIPR and GLP-1R is associated with the following signaling pathways: cAMP, neuroactive interaction, secretin family receptor class B/2, glucagon-type ligand receptors, G alpha(s) signaling events, and GPCR ligand binding. It was shown that GIP was negatively correlated with glucagon in patients with T2DM ($r = -0.283$, $r = -0.343$, $p < 0.05$) (**Figure 3C**). In patients with T2DM after RYGB, glucagon levels negatively correlated with GLP-1 ($r = -0.420$, $p < 0.05$) (**Figure 3E**).

Interestingly, the GPCR receptor can also interact with ghrelin. In obese patients with T2DM, ghrelin positively correlated with GIP ($r = 0.305$), insulin ($r = 0.420$), and resistin ($r = 0.313$) and negatively correlated with BMI ($r = -0.318$), glucagon ($r = -0.426$), and visfatin ($r = -0.319$) ($p < 0.05$) (**Figure 3C**).

Pathway enrichment analysis for the remaining proteins under study was performed to assess the possible mechanisms of the involvement of key regulators of carbohydrate metabolism in metabolism and intracellular signaling processes. These were the identified pathways: FOXO-mediated transcription of oxidative stress, metabolism, and neuronal genes; synthesis, secretion, and deacylation of ghrelin.

The results obtained demonstrate the joint participation of the studied regulators of carbohydrate metabolism in incretin-mediated and ghrelin-mediated functions and metabolic processes. In this case, these factors' mutual influence is implemented through a network of direct and indirect interactions.

DISCUSSION

C-peptide, ghrelin, GIP, GLP-1, insulin glucagon, PAI-1, resistin, leptin, and visfatin have been studied for assessing insulin sensitivity after bariatric surgery. It has been shown that these hormones have glucose-dependent secretion, have close relationships with each other, and have indicators of carbohydrate and lipid metabolism (Goktas et al., 2013; Pérez-Pevida et al., 2019). However, the results are multidirectional, and some relationships have not yet been deciphered. In our study, we combined an *in silico* and *in vivo* analysis unit.

Analyzing the data obtained, in obese patients with T2DM, the function of β -cells and the mechanism of substrate regulation works were not impaired, i.e., the higher the glucose level was, the more insulin the β -cells of the pancreas produced.

Interestingly, changes in the concentration of ghrelin were observed in various pathological conditions. Ghrelin is metabolically active with a negative metabolic balance (Al Qarni et al., 2017). At the excess intake of nutrients in obesity, the pathways that control energy balance became dysfunctional (Pöykkö et al., 2003; Al Qarni et al., 2017).

We showed *in silico* that ghrelin was associated with insulin, leptin, glucagon, and CRP. In particular, we found *in silico* connections between ghrelin and insulin through several pathways: GHRL-MLNR-GPRASP1-LRP2-INS, GHRL-HK3-LEP-A2M-CTSE-INS, GHRL-HK3-LEP-CLU-CPE-INS, and GHRL-HK3-LEP-LEPR-CLU-CPE-INS.

The level of ghrelin in obese patients with T2DM significantly increased relative to the other groups. The high levels of ghrelin may be due to the influence of high insulin levels in obese patients with T2DM, and its level depends on the level of other hormones.

The levels of C-peptide, insulin, GIP, GLP-1, leptin, and PAI-1 increased only in patients with T2DM. We have shown several links of ghrelin with hormones *in silico*: leptin, glucagon, insulin, GLP-1R, and PAI-1. Overall, this was consistent with our findings in obese patients.

The levels of total ghrelin and acyl-ghrelin are reduced in obese patients. The diacyl-ghrelin/acyl-ghrelin balance may change in obesity. The presence of one form or another can bind to different receptors and contribute to a change in ghrelin's effects. Interestingly, ghrelin levels were lower in obese patients without T2DM than in obese patients with T2DM and healthy volunteers. We emphasize this as a compensatory mechanism for maintaining normal carbohydrate and lipid metabolism in obesity. In patients after LSG and RYGB, against the background of a decrease in BMI, the blood plasma content of ghrelin was less than that in obese patients with T2DM before surgery.

Other authors have shown that ghrelin deficiency does not prevent diet-induced obesity (McFarlane et al., 2014). This proves that all patients showed hyperglycemia after 6 months, in contrast to our previous results (Skuratovskaia et al., 2019), where normoglycemia was noted after 12 months. Patients 6 months after LSG and RYGB are in an adaptive rehabilitation period, as evidenced by high glucose and C-reactive protein levels, indicating an acute phase of inflammation.

We believe that the compensation mechanism consists of the activation of specific receptors.

To elucidate the interaction pathways, we performed an *in silico* analysis of ghrelin receptors. It is known that preproghrelin undergoes posttranslational processing to form at least five products. Acyl-ghrelin is biologically active and interacts with the GHSR, and deacyl-ghrelin acts independently of GHSR1 (Gray et al., 2019). Acyl-ghrelin stimulates food intake, gastrointestinal motility, lipogenesis, and glycemia and reduces energy expenditure and insulin secretion/sensitivity (Gray et al., 2019). Deacyl-ghrelin inhibits food intake, gastrointestinal motility, and glycemia and stimulates insulin secretion through an as yet unknown receptor (Cui et al., 2017). Deacyl-ghrelin prevents the development of obesity and positively affects insulin sensitivity (Cui et al., 2017).

According to the latest data, the ghrelin receptors are GPCRs, GHSR1a, GHSR1b, ghrelin receptor-like receptor, and seven- and five-transmembrane GPCR to motilin (Sanger and Furness, 2016).

According to the functional annotation of ghrelin receptors, it is known that it activates three receptors: GPCR, GHSR, and GHRHR (Sanger and Furness, 2016). GPCRs have been associated with FZD4 modifications and the proliferative Wnt signaling pathway. GHSR has been associated with food responses, gut muscle contraction, and neurotransmitter and growth hormone synthesis. The role of GHSR in the immune response is interesting. GHSR blocks TNF- α and IL-6 functions and apoptosis of macrophages and myeloid cells (Cabral et al., 2017). GHRHR is associated with the production of growth factors, somatotropin, and insulin-like growth factor. Additionally, receptor activation is associated with the regulation of the sleep/wake cycle (Sanger and Furness, 2016).

It has been shown that ghrelin stimulates appetite by activating the hypothalamus and activates lipogenesis, leading to obesity (Qader et al., 2005). However, in our study, there was no relationship with lipid metabolism indicators.

It has been suggested that ghrelin and its receptor expressed in α -cells of the pancreas affect glucose metabolism not only by directly inhibiting the stimulation of insulin secretion by glucose (Qader et al., 2005) but also by stimulating glucagon secretion by α -cells (Chuang et al., 2011). In patients with T2DM, ghrelin positively correlated with insulin ($r = 0.402$, $p < 0.05$) and negatively correlated with glucagon ($r = -0.426$, $p < 0.05$), which indicates a violation of the relationship between hormones.

Thus, the predominance of one form or another of ghrelin and the activation of its receptors determine the fate of many links in the pathogenesis of IR in obesity. Our study measured total ghrelin; however, all its posttranslational modifications and effector pathways are of interest.

Association of Ghrelin With Incretins

In healthy people, the effects of ghrelin and GLP-1 on glucose metabolism are oppositely directed, but ghrelin can modulate postprandial GLP-1 secretion (Gagnon et al., 2015; Tong et al., 2016, 1; Gray et al., 2019) *in silico*. It was found that ghrelin interacts with GLP-1 and glucagon: GHRL-HK3-LEP-CRP-GLP2R-GCG, GHRL-HK3-LEP-LEPR-CLU-CPE-GCG, and GHRL-NPY-MEP1B-GCG. The incretin receptors GIPR and

GLP-1R have been implicated in the regulation of the cAMP signaling pathway.

We obtained earlier data on the important role of defects in incretin receptors in IR development. Thus, it was shown that the genotypes associated with an increased risk of developing T2DM, CC *rs1042044*, and AA *rs6923761* of the *GLP-1R* gene polymorphism, are characterized by an increase in the plasma level of incretin in the group of obese patients with T2DM and serum glucose levels in the group of obese patients without T2DM (Skuratovskaia, 2018). Impaired activation of GIPR and GLP-1R contributes to a decrease in the secretion of insulin and ghrelin and, on the contrary, to an increase in leptin and glucagon in the circulation, which contributes to the formation of insulin resistance in obese patients with T2DM.

Glucagon-like peptide-1 receptor is involved in the glucagon-type ligand-receptor family signaling pathway, which regulates the activity of GPCRs from the class II/B secretin receptor subfamily (Yabe and Seino, 2011). GLP-1R is synthesized in intestinal L cells in response to the presence of glucose and fatty acids (Tong et al., 2016). Most GLP-1 is in the GLP-1 (7–36) amidated form; some are the GLP-1–GLP-1 (7–37) form (Farr et al., 2016). GLP-1 circulates to the pancreas, where it binds to GLP-1R (Tong et al., 2016). GLP-1R is a transmembrane protein and a member of the B family of GPCRs with an N-terminal extracellular domain (Gagnon et al., 2015).

We performed functional annotation and analysis of the involvement of the *GLP-1R* gene in the signaling pathways, indicating its participation in the initiation of cAMP signaling, neuroactive interaction, and the provision of cell response to a stimulus in the form of glucagon, which was reflected in experimental studies: in patients of the healthy volunteers, the glucagon level positively correlated with GLP-1 ($r = 0.527$, $p < 0.05$). In obese patients with and without T2DM, as well as in healthy volunteers, GLP-1 and insulin levels were positively correlated. However, after RYGB surgery, no such relationship was found. GLP-1 and glucagon levels were negatively correlated.

Although GLP-1 and glucagon are formed from a common precursor (Hayashi, 2019), glucagon levels did not change among the study groups.

We have shown the dependence of the total ghrelin content on other mediators, such as insulin, leptin, GIP, GLP-1, and PAI-1. Our data suggest the cooperative interaction of mediators associated with ghrelin, which realizes their obesity effects, depending on the presence or absence of IR. However, the effects of ghrelin may depend on how its effects are realized.

We used models of different surgeries to study in detail the mutual regulation of hormones. LSG removes most of the ghrelin-producing zone in the stomach. Unusually, the ghrelin level after the breakfast test did not decrease but was significantly higher than the fasting values in patients after LSG. This fact may indicate that other additional stimulatory signals exist for the secretion of ghrelin, acting differently from the mechanical stimulation of the cells of the fundus of the stomach and depending on remote regulation by the intake of nutrients. Moreover, it is believed that the return of insulin sensitivity is facilitated by a decrease in the inhibitory effect of ghrelin on insulin (McFarlane et al., 2014), which is consistent with our

results on the construction of a network of interactions of the studied proteins. We found an indirect interaction of ghrelin and insulin (through three nodes: MLNR, GPRASP1, and LRP2). Of interest is the study of some links of this pathway: MLNR is a motilin receptor involved in hormone binding, GPRASP1 promotes the degradation of G protein-coupled receptors in lysosomes (Whistler et al., 2002), and LRP2 is required for insulin-dependent internalization of IR (Seo et al., 2020).

According to the results of the study, two groups of pathways were identified in which the studied molecules are involved: FOXO-mediated transcription of genes for oxidative stress, metabolism, and neurons (through the participation of *INS* and *RETN*) and the synthesis, secretion, and diacylation of ghrelin (with the involvement of *GCG*, *GHRL*, *INS*, *LEP*, and *GIP*). The results obtained demonstrate the joint participation of the studied regulators of carbohydrate metabolism in incretin-mediated and ghrelin-mediated functions and metabolic processes. Simultaneously, the mutual influence of these factors is realized through a network of direct and mediated interactions.

Interestingly, the GPCR can also interact with ghrelin. Moreover, this same receptor can be activated by GLP-1 and GIP (Abdullah et al., 2016). In obese patients with T2DM, ghrelin positively correlated with GIP ($r = 0.305$), insulin ($r = 0.420$), and resistin ($r = 0.313$) and negatively correlated with BMI ($r = -0.318$), glucagon ($r = -0.426$), and visfatin ($r = -0.319$) ($p < 0.05$). Thus, C-peptide, insulin, leptin, PAI, GIP, GLP-1, ghrelin, resistin, and visfatin are closely associated with carbohydrate and lipid metabolism and BMI in obese patients, which contributes to the formation of IR in obesity.

The use of ghrelin receptor antagonists is effective in correcting carbohydrate metabolism. There is one clinical trial in humans that promoted glucose-dependent insulin secretion by blocking the GHSR1a receptor (Schalla and Stengel, 2019). Despite the significant effect, ghrelin did not have the desired effect in treating metabolic disorders, and the drug is currently used to treat insomnia (Schalla and Stengel, 2019).

The animal models have recently shown that GHS-R1b antagonists lead to decreased food intake (Schalla and Stengel, 2019). The development of GHSR antagonists leads to unpredictable results and requires a detailed study of the mechanisms of action. The predominance of one form or another of ghrelin and the activation of its receptors determine the fate of many links in the pathogenesis of IR in obesity. The search for new receptors for ghrelin and the study of posttranslational modifications of ghrelin will make it possible to study the regulation of metabolic disorders more closely. Further studies can help identify the mechanisms of participation of ghrelin in metabolic disorders in patients.

CONCLUSION

A decreased level of total ghrelin before and after breakfast is typical only for obese patients without insulin resistance. Its increased level is typical for obese patients with T2DM. Ghrelin exerts its effects through the receptors GPCR, GHSR, and GHRHR, which are associated with proliferative, inflammatory,

and neurohumoral signaling pathways and regulate responses to nutrient intake. The signaling pathway for realizing the effects of ghrelin, GIP, GLP-1, and GCG lies through the class of GPCRs. This demonstrates common regulatory mechanisms and cross-talk between ghrelin and incretins. This will help draw attention to posttranslational modifications and the associated ghrelin and incretin receptors for use in targeted therapy of insulin resistance.

The results obtained demonstrate the joint participation of the studied regulators of carbohydrate metabolism in incretin-mediated and ghrelin-mediated functions and metabolic processes. Simultaneously, the mutual influence of these factors is realized through a network of direct and mediated interactions. A cooperative exchange of ghrelin and mediators associated with it, such as insulin, leptin, GIP, GLP-1, glucagon, and PAI-1, was revealed, which realizes their effects on obesity. Insulin, leptin, GIP, GLP-1, glucagon, and PAI-1 are linked directly or indirectly (see the “Results” section for more details). Thus, ghrelin and incretins can modulate insulin, leptin, glucagon, and PAI-1.

This does not exclude the influence of insulin and glucose on the joint change in mediators due to other factors not included in the study. The influence of the microenvironment on the production of these mediators should be taken into account.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by all the study participants provided informed consent to participate in a research study. The study was carried out in accordance with the World Medical Association

REFERENCES

- Abdullah, N., Beg, M., Soares, D., Dittman, J. S., and McGraw, T. E. (2016). Downregulation of a GPCR by β -Arrestin2-Mediated Switch from an Endosomal to a TGN Recycling Pathway. *Cell Rep.* 17, 2966–2978. doi: 10.1016/j.celrep.2016.11.050
- Al Qarni, A. A., Joatar, F. E., Das, N., Awad, M., Eltayeb, M., Al-Zubair, A. G., et al. (2017). Association of Plasma Ghrelin Levels with Insulin Resistance in Type 2 Diabetes Mellitus among Saudi Subjects. *Endocrinol. Metab.* 32, 230–240. doi: 10.3803/EnM.2017.32.2.230
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556
- Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., et al. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091–1093. doi: 10.1093/bioinformatics/btp101
- Brandt, S. J., Götz, A., Tschöp, M. H., and Müller, T. D. (2018). Gut hormone polyagonists for the treatment of type 2 diabetes. *Peptides* 100, 190–201. doi: 10.1016/j.peptides.2017.12.021
- Cabral, A., López Soto, E. J., Epelbaum, J., and Perelló, M. (2017). Is Ghrelin Synthesized in the Central Nervous System? *Int. J. Mol. Sci.* 18:18030638. doi: 10.3390/ijms18030638

(WMA) Declaration of Helsinki (2000) and the Protocol to the Convention on Human Rights and Biomedicine (1999). The Local Ethical Committee approved the study protocol of the Innovation Park of the Immanuel Kant Baltic Federal University (Protocol No. 4 from October 23, 2013). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DS and MV conceptualized the study. LL and EK performed the validation. AK did the formal analysis. ES and AK wrote and prepared the original draft and performed the visualization. DS, MV, and LL reviewed and edited the article. NC did the bioinformatics analysis. AK and DS performed the experimental work. All authors contributed to the article and approved the submitted version.

FUNDING

The research was supported by the Russian Foundation for Basic Research (No. 18-015-00084-a to LL), by the Russian Foundation for Basic Research and Kaliningrad Region (No. 19-415-393004-r_mol_a to DS and No. 19-44-390005-r_a to DS), the state assignment (No. FZWM-2020-0010 to LL), and the state support of leading scientific schools of the Russian Federation (No. 2495.2020.7 to LL).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.612501/full#supplementary-material>

- Campbell, J. E., Ussher, J. R., Mulvihill, E. E., Kolic, J., Baggio, L. L., Cao, X., et al. (2016). TCF1 links GIPR signaling to the control of beta cell function and survival. *Nat. Med.* 22, 84–90. doi: 10.1038/nm.3997
- Chuang, J.-C., Sakata, I., Kohno, D., Perello, M., Osborne-Lawrence, S., Repa, J. J., et al. (2011). Ghrelin Directly Stimulates Glucagon Secretion from Pancreatic α -Cells. *Mol. Endocrinol.* 25, 1600–1611. doi: 10.1210/me.2011-1001
- Cui, H., López, M., and Rahmouni, K. (2017). The cellular and molecular bases of leptin and ghrelin resistance in obesity. *Nat. Rev. Endocrinol.* 13, 338–351. doi: 10.1038/nrendo.2016.222
- Dedov, I. I., Shestakova, M. V., Peterkova, V. A., Ёїїñòàíòëñàíà, ĂĬ, Vikulova, O. K., ZHeleznyakova, A. V., et al. (2017). *Saharnyj diabet u detej i podrostkov po dannym Federal'nogo registra Rossijskoj Federacii: dinamika osnovnyh epidemiologicheskij harakteristik za 2013-2016 gg.* *Saharnyj Diabet* 20. Available online at: <https://cyberleninka.ru/article/n/saharnny-diabet-u-detey-i-podrostkov-po-dannym-federalnogo-registra-rossijskoj-federatsii-dinamika-osnovnyh-epidemiologicheskij> (accessed September 16, 2020).
- Farr, O. M., Tsoukas, M. A., Triantafyllou, G., Dincer, F., Filippaios, A., Ko, B.-J., et al. (2016). Short-term administration of the GLP-1 analog liraglutide decreases circulating leptin and increases GIP levels and these changes are associated with alterations in CNS responses to food cues: A randomized, placebo-controlled, crossover study. *Metabolism* 65, 945–953. doi: 10.1016/j.metabol.2016.03.009

- Finan, B., Ma, T., Ottaway, N., Müller, T. D., Habegger, K. M., Heppner, K. M., et al. (2013). Unimolecular dual incretins maximize metabolic benefits in rodents, monkeys, and humans. *Sci. Transl. Med.* 5:209ra151. doi: 10.1126/scitranslmed.3007218
- Gagnon, J., Baggio, L. L., Drucker, D. J., and Brubaker, P. L. (2015). Ghrelin Is a Novel Regulator of GLP-1 Secretion. *Diabetes* 64, 1513–1521. doi: 10.2337/db14-1176
- Goktas, Z., Moustaid-Moussa, N., Shen, C.-L., Boylan, M., Mo, H., and Wang, S. (2013). Effects of Bariatric Surgery on Adipokine-Induced Inflammation and Insulin Resistance. *Front. Endocrinol.* 4:69. doi: 10.3389/fendo.2013.00069
- Gray, S. M., Page, L. C., and Tong, J. (2019). Ghrelin Regulation of Glucose Metabolism. *J. Neuroendocrinol.* 31:e12705. doi: 10.1111/jne.12705
- Hayashi, Y. (2019). Glutaminostatin: Another facet of glucagon as a regulator of plasma amino acid concentrations. *J. Diabetes Investig.* 10, 1391–1393. doi: 10.1111/jdi.13110
- Hong, X., Zhang, H., Liang, H., Li, D., Huang, J., Li, Z., et al. (2016). Exendin-4 decreases ghrelin levels through mTOR signaling. *Mol. Cell. Endocrinol.* 437, 201–212. doi: 10.1016/j.mce.2016.08.039
- Karim, R., Stanczyk, F. Z., Brinton, R. D., Rettberg, J., Hodis, H. N., and Mack, W. J. (2015). Association of Endogenous Sex Hormones with Adipokines and Ghrelin in Postmenopausal Women. *J. Clin. Endocrinol. Metab.* 100, 508–515. doi: 10.1210/jc.2014-1839
- Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., et al. (2009). Human protein reference database 2009 update. *Nucleic Acids Res.* D767–772. doi: 10.1093/nar/gkn892
- Kim, W., and Egan, J. M. (2008). The Role of Incretins in Glucose Homeostasis and Diabetes Treatment. *Pharmacol. Rev.* 60, 470–512. doi: 10.1124/pr.108.000604
- McFarlane, M. R., Brown, M. S., Goldstein, J. L., and Zhao, T.-J. (2014). Induced Ablation of Ghrelin Cells in Adult Mice Does Not Decrease Food Intake, Body Weight, or Response to High-Fat Diet. *Cell Metab.* 20, 54–60. doi: 10.1016/j.cmet.2014.04.007
- Nielsen, S. T., Janum, S., Krogh-Madsen, R., Solomon, T. P., and Møller, K. (2015). The incretin effect in critically ill patients: a case-control study. *Crit. Care Lond. Engl.* 19:402. doi: 10.1186/s13054-015-1118-z
- Pérez-Pevida, B., Escalada, J., Miras, A. D., and Frühbeck, G. (2019). Mechanisms Underlying Type 2 Diabetes Remission After Metabolic Surgery. *Front. Endocrinol.* 10:641. doi: 10.3389/fendo.2019.00641
- Pöykkö, S. M., Kellokoski, E., Hökkö, S., Kauma, H., Kesäniemi, Y. A., and Ukkola, O. (2003). Low Plasma Ghrelin Is Associated With Insulin Resistance, Hypertension, and the Prevalence of Type 2 Diabetes. *Diabetes* 52, 2546–2553. doi: 10.2337/diabetes.52.10.2546
- Qader, S. S., Lundquist, I., Ekelund, M., Håkanson, R., and Salehi, A. (2005). Ghrelin activates neuronal constitutive nitric oxide synthase in pancreatic islet cells while inhibiting insulin release and stimulating glucagon release. *Regul. Pept.* 128, 51–56. doi: 10.1016/j.regpep.2004.12.018
- Rivals, I., Personnaz, L., Taing, L., and Potier, M.-C. (2007). Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 23, 401–407. doi: 10.1093/bioinformatics/btl633
- Ronveaux, C. C., Tomé, D., and Raybould, H. E. (2015). Glucagon-like peptide 1 interacts with ghrelin and leptin to regulate glucose metabolism and food intake through vagal afferent neuron signaling. *J. Nutr.* 145, 672–680. doi: 10.3945/jn.114.206029
- Sanger, G. J., and Furness, J. B. (2016). Ghrelin and motilin receptors as drug targets for gastrointestinal disorders. *Nat. Rev. Gastroenterol. Hepatol.* 13, 38–48. doi: 10.1038/nrgastro.2015.163
- Schalla, M. A., and Stengel, A. (2019). Pharmacological Modulation of Ghrelin to Induce Weight Loss: Successes and Challenges. *Curr. Diab. Rep.* 19:102. doi: 10.1007/s11892-019-1211-9
- Seo, J. A., Kang, M.-C., Yang, W.-M., Hwang, W. M., Kim, S. S., Hong, S. H., et al. (2020). Apolipoprotein J is a hepatokine regulating muscle glucose metabolism and insulin sensitivity. *Nat. Commun.* 11:2024. doi: 10.1038/s41467-020-15963-w
- Skow, M. A., Bergmann, N. C., and Knop, F. K. (2016). Diabetes and obesity treatment based on dual incretin receptor activation: “twincretins.”. *Diabetes Obes. Metab.* 18, 847–854. doi: 10.1111/dom.12685
- Skuratovskaia, D. A. (2018). *Rol' receptorov k inkretinam v patogeneze insulinorezistentnosti pri ozhireнии avtoferat dis. na soisk. uch. step. kandidata biologicheskikh nauk special'nost' < Patologicheskaya fiziologiya >.* Available online at: https://rusneb.ru/catalog/000200_000018_RU_NLR_BIBL_A_012055329/ (accessed September 23, 2020).
- Skuratovskaia, D., Litvinova, L., Vulf, M., Zatolokin, P., Popadin, K., and Mazunin, I. (2019). From Normal to Obesity and Back: The Associations between Mitochondrial DNA Copy Number, Gender, and Body Mass Index. *Cells* 8:430.
- Statistics About Diabetes | ADA, (2020). *Statistics About Diabetes*. Available online at: <https://www.diabetes.org/resources/statistics/statistics-about-diabetes> (accessed September, 2020).
- Stofkova, A. (2010). Resistin and visfatin: regulators of insulin sensitivity, inflammation and immunity. *Endocr. Regul.* 44, 25–36. doi: 10.4149/endo_2010_01_25
- Tong, J., Davis, H. W., Gastaldelli, A., and D'Alessio, D. (2016). Ghrelin Impairs Prandial Glucose Tolerance and Insulin Secretion in Healthy Humans Despite Increasing GLP-1. *J. Clin. Endocrinol. Metab.* 101, 2405–2414. doi: 10.1210/jc.2015-4154
- Udén, T. B., Mezey, E., Button, D. C., Brownstein, M. J., and Bonner, T. I. (1993). Gastric inhibitory polypeptide receptor, a member of the secretin-vasoactive intestinal peptide receptor family, is widely distributed in peripheral organs and the brain. *Endocrinology* 133, 2861–2870. doi: 10.1210/endo.133.6.8243312
- Vejrazkova, D., Lischkova, O., Vankova, M., Stanicka, S., Vrbikova, J., Lukasova, P., et al. (2017). Distinct response of fat and gastrointestinal tissue to glucose in gestational diabetes mellitus and polycystic ovary syndrome. *Physiol. Res.* 66, 283–292. doi: 10.33549/physiolres.933366
- Whistler, J. L., Enquist, J., Marley, A., Fong, J., Gladher, F., Tsuruda, P., et al. (2002). Modulation of postendocytic sorting of G protein-coupled receptors. *Science* 297, 615–620. doi: 10.1126/science.1073308
- WHO | Raised fasting blood glucose, (2020). *Raised fasting blood glucose*. Geneva: WHO.
- Yabe, D., and Seino, Y. (2011). Two incretin hormones GLP-1 and GIP: comparison of their actions in insulin secretion and β cell preservation. *Prog. Biophys. Mol. Biol.* 107, 248–256. doi: 10.1016/j.pbiomolbio.2011.07.010
- Yabe, D., and Seino, Y. (2013). Incretin actions beyond the pancreas: lessons from knockout mice. *Curr. Opin. Pharmacol.* 13, 946–953. doi: 10.1016/j.coph.2013.09.013
- Yanagi, S., Sato, T., Kangawa, K., and Nakazato, M. (2018). The Homeostatic Force of Ghrelin. *Cell Metab.* 27, 786–804. doi: 10.1016/j.cmet.2018.02.008

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Skuratovskaia, Vulf, Chasovskikh, Komar, Kirienkova, Shunkin, Zatolokin and Litvinova. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership