BIOINFORMATICS OF GENOME REGULATION, VOLUME II

EDITED BY: Yuriy L. Orlov, Ancha Baranova, Tatiana V. Tatarinova and Anastasia A. Anashkina

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-177-9
DOI 10.3389/978-2-88974-177-9

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 II

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
Anastasia A. Anashkina, Engelhardt Institute of Molecular Biology (RAS), Russia

Citation: Orlov, Y. L., Baranova, A., Tatarinova, T. V., Anashkina, A. A., eds. (2022). Bioinformatics of Genome Regulation, Volume II. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-177-9

Table of Contents

frontiers in Genetics

EDITORIAL
published: 08 November 2021
doi: 10.3389/fgene.2021.795257

Editorial: Bioinformatics of Genome Regulation, Volume II

Yuriy L. Orlov[1,2]*, *Anastasia A. Anashkina*[1,3], *Tatiana V. Tatarinova*[4] and *Ancha V. Baranova*[5,6]

[1]The Digital Health Institute, I.M.Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia, [2]Agrobiotechnology Department, Agrarian and Technological Institute, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia, [3]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, [4]Natural Science Division, La Verne University, La Verne, CA, United States, [5]School of Systems Biology, George Mason University, Fairfax, VA, United States, [6]Research Centre for Medical Genetics, Moscow, Russia

Keywords: bioinformatics, transcriptomics, plant science, gene networks, gene expression regulation, genetics, computational genomics

Editorial on the Research Topic

Bioinformatics of Genome Regulation, Volume II

This Research Topic Volume II, "Bioinformatics of Genome Regulation," continues the studies in the field of bioinformatics of gene expression presented initially at *Frontiers in Genetics* journal (https://www.frontiersin.org/research-topics/8383/bioinformatics-of-genome-regulation-and-systems-biology) and then in Volume I (https://www.frontiersin.org/research-topics/14266/bioinformatics-of-genome-regulation-volume-i). The materials presented here were discussed in a BGRS\SB (Bioinformatics of Genome Regulation and Structure Systems Biology) conference series in Novosibirsk, Russia (https://bgrssb.icgbio.ru/2020/). The BGRS is the central event in the computational genetics field, held in Novosibirsk, Russia, every other year since 1998 (Orlov et al., 2015). The publications were later completed by new studies on computational methods of gene expression analysis regulation. Starting in 2018, materials of the conference materials in genetics and genomics were presented in *Frontiers in Genetics*, and due to popular demand in 2021 it was extended as the Volume II. The BGRS conference series have been presented at special journal issues earlier (Orlov et al., 2016; Tatarinova et al., 2019; Orlov et al., 2015; Orlov et al., 2019a; Orlov et al., 2019b; Baranova et al., 2019) and recently (Tatarinova et al., 2020; Orlov and Baranova, 2020; Orlov et al., 2020; Orlov et al., 2021a; Orlov et al., 2021b). We have to acknowledge "Bioinformatics of Gene Regulations and Structure" special issue at *MDPI IJMS* (https://www.mdpi.com/journal/ijms/special_issues/Bioinformatics_Genomics), as well as *PeerJ* journal BGRS-2020 collection (https://peerj.com/collections/72-bgrs-sb-2020).

This research topic presents seventeen papers on medical genomics applications, new bioinformatics tools and applications to laboratory animal models, and plant sciences.

Biomedical papers start from applications to cancer studies. Xiaorong Yang et al. discussed the interplay between human diseases at multiple biological levels. The authors show the role of schizophrenia in the pathological development of myocardial infarction, suggesting its role in promoting the development and progression of myocardial infarction at different levels, including genes, small molecules, and complex molecules. Pathway analysis revealed nine genes connecting these diseases.

Maxim Sorokin et al. proposed an algorithm that identifies functional roles of the pathway components and applied it to annotate 3,044 human molecular pathways extracted from the Biocarta, Reactome, KEGG, Qiagen Pathway Central, NCI, and HumanCYC databases. The resulting knowledgebase may be used to calculate the levels of activation for individual pathways and to establish large-scale profiles of the signaling, metabolic, and DNA repair regulation using high throughput gene expression data, as was presented recently (Wang et al., 2021).

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: *14 October 2021*
Accepted: *25 October 2021*
Published: *08 November 2021*

Citation:
Orlov YL, Anashkina AA, Tatarinova TV and Baranova AV (2021) Editorial: Bioinformatics of Genome Regulation, Volume II. Front. Genet. 12:795257. doi: 10.3389/fgene.2021.795257

Anna E. Letiagina et al. presented a new bioinformatics application on massively parallel reporter assays (MPRAs). The assays are based on the construction of reporter plasmid libraries. The authors present a pipeline for processing raw MPRA data obtained by NGS for reporter construct libraries with a priori unknown sequences of a region of interest and a barcode in the plasmid, located outside and within the transcription unit. The pipeline robustly identifies unambiguous barcodes, calculates the normalized expression level, and provides a graphical visualization of the processed data.

Maria M. Litvinova et al. have studied genes polymorphisms associations with calcium urolithiasis development. Authors found statistically significant associations between calcium urolithiasis and the polymorphisms in *CASR* rs1042636, *CALCR* rs1801197, and *ORAI1* rs6486795.

Yi He et al. explored the regulatory network among circRNA-miRNA-mRNA in osteosarcoma. The authors interrogated NCBI GEO osteosarcoma datasets and extracted circRNA, microRNA, and mRNA expression profiles. Downstream network analysis was performed with CircInteractome (Circular RNA Interactome) and mirDIP (microRNA Data Integration Portal) databases. Authors show that quinacridine, thalidomide, and zonisamide may be potential drugs for the treatment of osteosarcoma.

Stepan Nersisyan et al. investigated the effects of hypoxia on pathophysiological processes, including cancer progression and metastasis formation. This work continued the earlier studies on miRNA regulatory interactions in cancer (Shkurnikov et al., 2019). The landscape of hypoxia-induced miRNA and mRNA expression alterations was studied in human colorectal cancer cell lines (HT-29 and Caco-2) to show that miR-148a downregulation contributes to poor survival due to overexpression of the *ITGA5* and *PRNP* genes.

Sergey Nikulin et al.) have studied gene expression in breast cancer cells. The authors showed that malignant breast tumors with reduced expression of the *ELOVL5* and *IGFBP6* genes could metastasize at a higher probability due to a more efficient invasion of tumor cells. In addition, a set of novel computational techniques was developed for deciphering gene expression regulation.

Igor B. Rogozin et al. discussed fundamental biochemical mechanisms mutations in cancer driver genes. A large fraction of these mutations arose due to the off-target activity of DNA/RNA editing cytosine deaminases, followed by the replication/repair of edited sites by DNA polymerases. Using methylation data from malignant lymphomas, the authors showed that driver genes are subject to different (de)methylation processes than non-driver genes.

Natalya V. Klimova et al., using previously published web-tool SNP_TATA_Comparator (Ponomarenko et al., 2017), conducted a genome-wide study of single-nucleotide polymorphisms (SNPs) within core promoters of 68 human rheumatoid arthritis-related genes. They show that the disruptive natural selection of human immunostimulatory and immunosuppressive genes concurrently elevates and reduces the risk of rheumatoid arthritis. The authors hypothesize that rheumatoid arthritis in humans could be a self-domestication syndrome referring to evolution patterns in domestic animals (Chadaeva et al., 2021).

James Sweet-Jones et al. applied genetic tools to livestock breeding on the example of the Welsh mountain sheep breeds. Genotyping data from 317 individuals representing 15 Welsh sheep breeds were used alongside the whole-genome resequencing data of 14 species from the same set to scan for the signatures of selection and candidate genetic variants using haplotype SNP-based approaches. The authors found new variants in genes with potential functional consequences to the adaptation of local sheep to their environments in Wales. The study continues in new markers search in animal adaptation to cold climate (Igoshin et al., 2021).

Lilit Nersisyan et al. investigated telomere maintenance mechanisms for studying cancers and designing therapies.

In Brief Research Report, Elena V. Ignatieva et al. presented a new database - a catalog of human genes associated with pathologies of the sperm. It contains data genes related to male fertility and their functional annotation based on the literature data and clinical trials (Kolmykov et al., 2021). Functional annotation of genes from the catalog showed that spermatogenic failure could be associated with mutations in genes that control biological processes essential for spermiogenesis (such as DNA metabolism, cell division). Azoospermia can be caused by mutations in genes that control cellular responses to unfavorable conditions (stress factors, including oxidative stress and exposure to toxins).

Ekaterina Ilgisonis et al. considered a problem of transcriptome annotation by sequencing. They compared the results obtained from different transcriptome analysis platforms (quantitative polymerase chain reaction, Illumina RNASeq, and Oxford Nanopore Technologies MinION) for the transcriptome encoded by human chromosome 18 using the same sample types. The combination of Illumina RNASeq and MinION nanopore technologies reduced the probability of false-positive detection of low-copy transcripts due to the simultaneous confirmation of the presence of a transcript by the two fundamentally different technologies: short reads essential for reliable detection and long-read sequencing data.

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 (Computer plant biology Session of BGRS conference, and the Plantgen conference series, https://conf.icgbio.ru/plantgen2021/) (Orlov, 2019; Orlov et al., 2019b).

Nann Miky Moh Moh et al. studied miRNAs and lncRNAs from the mango (*Mangifera indica* L.). Although mango is a popular food having pharmacological potential, its non-coding RNA data were limited. For the first time, a large-scale study identified nearly a hundred miRNAs and over 7,000 temperature-responsive lncRNAs. Characterization of target genes for these ncRNAs was performed.

Ekaterina M. Dvorianinova et al. presented the application of sequencing technologies to study plant pathogens of flax (*Linum usitatissimum* L.). Genome Assembly of the pathogen *Fusarium oxysporum* f. sp. *lini* was presented (Krasnov et al., 2020). Due to *F. oxysporum* f. sp. *lini* includes many genotypes, it is of high significance to study the origins of pathogenicity at the molecular level. This work mainly focused on genome sequencing of strains

of the flax pathogen *F. oxysporum* f. sp. *lini*, possessing diverse pathogenicity degrees, on two platforms—Oxford Nanopore Technologies and Illumina. Sequencing using these two platforms proved to effectively achieve high-quality genome assemblies for complex plant genomes (Melnikova et al., 2021).

Victoria A. Scobeyeva et al. studied patterns of evolution in plant genome on the example of *Allium* species. *Alliums* are widespread and diversified; they are adapted to various habitats, from shady forests to open steppes. The genes present in chloroplast genomes (plastomes) play fundamental roles for photosynthetic plants. Plastome traits could thus be associated with geophysical abiotic characteristics of habitats. The authors compared their data with previously published plastomes and provided our interpretation of *Allium* plastome genes' annotations. They can hypothesize that adaptive evolution in genes, coding subunits of NADH-plastoquinone oxidoreductase could be driven by abiotic factors of alpine habitats, especially by intensive light and UV radiation.

Elena N. Pushkova et al. studied Allele-Specific Expression in plants. Transcriptome sequencing of plant tissues from the male and female trees of *Populus* × *sibirica* and genome sequencing of the same plants were performed first. Targeted sequencing of sex-determining region (SDR) genes such as *CLC* (Chloride channel protein CLC-c on a representative set of trees confirmed the sex-associated allele-specific expression in generative and vegetative tissues of *P.* × *sibirica*.

Overall, we are proud of the continuing Research Topic at Frontiers in Genetics we collated. Biomedical applications for gene expression studies in chronic diseases are presented in ongoing 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). We hope 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/).

AUTHOR CONTRIBUTIONS

YO, AA, TT, and AB organized the Research Topic as guest editors, supervised the reviewing of the manuscripts. All the authors wrote this Editorial paper. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The guest editors are grateful to the authors contributing to this special issue papers collection and thank all the reviewers who helped improve the manuscripts. The BGRS-2020 conference organization was supported by Novosibirsk State University and the Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia. The publication has been prepared with the support of the RUDN University Strategic Academic Leadership Program (recipient YO).

REFERENCES

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

Chadaeva, I., Ponomarenko, P., Kozhemyakina, R., Suslov, V., Bogomolov, A., Klimova, N., et al. (2021). Domestication Explains Two-Thirds of Differential-Gene-Expression Variance between Domestic and Wild Animals; the Remaining One-Third Reflects Intraspecific and Interspecific Variation. *Animals.* 11 (9), 2667. doi:10.3390/ani11092667

Igoshin, A., Yudin, N., Aitnazarov, R., Yurchenko, A. A., and Larkin, D. M. (2021). Whole-Genome Resequencing Points to Candidate DNA Loci Affecting Body Temperature Under Cold Stress in Siberian Cattle Populations. *Life.* 11 (9), 959. doi:10.3390/life11090959

Kolmykov, S., Vasiliev, G., Osadchuk, L., Kleschev, M., and Osadchuk, A. (2021). Whole-Exome Sequencing Analysis of Human Semen Quality in Russian Multiethnic Population. *Front. Genet.* 12, 662846. doi:10.3389/fgene.2021.662846

Krasnov, G. S., Pushkova, E. N., Novakovskiy, R. O., Kudryavtseva, L. P., Rozhmina, T. A., Dvorianinova, E. M., et al. (2020). High-Quality Genome Assembly of *Fusarium Oxysporum* F. Sp. *Lini*. *Front. Genet.* 11, 959. doi:10.3389/fgene.2020.00959

Melnikova, N. V., Pushkova, E. N., Dvorianinova, E. M., Beniaminov, A. D., Novakovskiy, R. O., Povkhova, L. V., et al. (2021). Genome Assembly and Sex-Determining Region of Male and Female Populus × Sibirica. *Front. Plant Sci.* 12, 625416. doi:10.3389/fpls.2021.625416

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., 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. doi:10.18097/PBMC20216703222

Orlov, Y. L., and Baranova, A. V. (2020). Editorial: Bioinformatics of Genome Regulation and Systems Biology. *Front. Genet.* 11, 625. doi:10.3389/fgene.2020.00625

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., Baranova, A. V., and Tatarinova, T. V. (2020). Bioinformatics Methods in Medical Genetics and Genomics. *Int. J. Mol. Sci.* 21 (17), 6224. doi:10.3390/ijms21176224

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

Ponomarenko, P., Chadaeva, I., Rasskazov, D. A., Sharypova, E., Kashina, E. V., Drachkova, I., et al. (2017). Candidate SNP Markers of Familial and Sporadic Alzheimer's Diseases Are Predicted by a Significant Change in the Affinity of TATA-Binding Protein for Human Gene Promoters. *Front. Aging Neurosci.* 9, 231. doi:10.3389/fnagi.2017.00231

Shkurnikov, M., Nikulin, S., Nersisyan, S., Poloznikov, A., Zaidi, S., Baranova, A., et al. (2019). LAMA4-Regulating miR-4274 and its Host Gene SORCS2 Play a Role in IGFBP6-Dependent Effects on Phenotype of Basal-Like Breast Cancer. *Front. Mol. Biosci.* 6, 122. doi:10.3389/fmolb.2019.00122

Tatarinova, T. V., Baranova, A. V., Anashkina, A. A., and Orlov, Y. L. (2020). Genomics and Systems Biology at the "Century of Human Population Genetics" Conference. *BMC Genomics.* 21 (Suppl. 7), 592. doi:10.1186/s12864-020-06993-1

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

Wang, Y., Tong, Z., Zhang, W., Zhang, W., Buzdin, A., Mu, X., et al. (2021). FDA-Approved and Emerging Next Generation Predictive Biomarkers for Immune Checkpoint Inhibitors in Cancer Patients. *Front. Oncol.* 11, 683419. doi:10.3389/fonc.2021.683419

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 February 2021
doi: 10.3389/fgene.2021.617059

Algorithmic Annotation of Functional Roles for Components of 3,044 Human Molecular Pathways

*Maxim Sorokin[1,2,3], Nicolas Borisov[1,3], Denis Kuzmin[3], Alexander Gudkov[2], Marianna Zolotovskaia[3], Andrew Garazha[1] and Anton Buzdin[1,3,4,5]**

[1]Omicsway Corp., Walnut, CA, United States, [2]Laboratory of Clinical Genomic Bioinformatics, I.M. Sechenov First Moscow State Medical University, Moscow, Russia, [3]Laboratory for Translational Bioinformatics, Moscow Institute of Physics and Technology, Moscow, Russia, [4]Laboratory of Systems Biology, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia, [5]World-Class Research Center "Digital Biodesign and Personalized Healthcare", Sechenov First Moscow State Medical University, Moscow, Russia

Current methods of high-throughput molecular and genomic analyses enabled to reconstruct thousands of human molecular pathways. Knowledge of molecular pathways structure and architecture taken along with the gene expression data can help interrogating the pathway activation levels (PALs) using different bioinformatic algorithms. In turn, the pathway activation profiles can characterize molecular processes, which are differentially regulated and give numeric characteristics of the extent of their activation or inhibition. However, different pathway nodes may have different functions toward overall pathway regulation, and calculation of PAL requires knowledge of molecular function of every node in the pathway in terms of its activator or inhibitory role. Thus, high-throughput annotation of functional roles of pathway nodes is required for the comprehensive analysis of the pathway activation profiles. We proposed an algorithm that identifies functional roles of the pathway components and applied it to annotate 3,044 human molecular pathways extracted from the Biocarta, Reactome, KEGG, Qiagen Pathway Central, NCI, and HumanCYC databases and including 9,022 gene products. The resulting knowledgebase can be applied for the direct calculation of the PALs and establishing large scale profiles of the signaling, metabolic, and DNA repair pathway regulation using high throughput gene expression data. We also provide a bioinformatic tool for PAL data calculations using the current pathway knowledgebase.

Keywords: functional algorithmic annotation, signaling pathways, DNA repair pathways, metabolic pathways, transcriptomics, proteomics, human molecular pathway regulation

OPEN ACCESS

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

Reviewed by:
Peter D'Eustachio, New York University, United States
Pavel Kopnin, Russian Cancer Research Center NN Blokhin, Russia
Teresa Bernadette Steinbichler, Innsbruck Medical University, Austria

***Correspondence:**
Anton Buzdin
buzdin@oncobox.com;
bu3din@mail.ru

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

Received: *13 October 2020*
Accepted: *20 January 2021*
Published: *09 February 2021*

Citation:
Sorokin M, Borisov N, Kuzmin D, Gudkov A, Zolotovskaia M, Garazha A and Buzdin A (2021) Algorithmic Annotation of Functional Roles for Components of 3,044 Human Molecular Pathways. Front. Genet. 12:617059. doi: 10.3389/fgene.2021.617059

INTRODUCTION

Intracellular molecular pathways are specific networks of interacting molecules that are involved in certain molecular functions (Junaid et al., 2020; Ma and Liao, 2020; Zheng et al., 2020). Knowledge of molecular pathways regulation is important for understanding intracellular processes related to all major events, including cell survival, growth, differentiation, motility, proliferation, senescence, malignization, and death (Buzdin et al., 2018). Molecular pathways are affected during organism growth and development, aging and disease progression (Parkhitko et al., 2020). Current methods of large-scale molecular and genomic analyses enabled to catalogue thousands of human molecular pathways (Wishart et al., 2020). In turn, high-throughput gene expression analyses like RNA sequencing (Sorokin et al., 2020a), expression

microarrays (Schulze and Downward, 2001; Shih et al., 2005; Willier et al., 2013), or modern proteomic techniques (Buzdin et al., 2019) can provide adequate amounts of data to enable interactome-wide assessment of pathway activation.

Several popular algorithms and software like gene ontology (GO) analysis tools (Huang et al., 2009a,b), Metacore (Ekins et al., 2007) and Pathway Studio (Thomas and Bonchev, 2010) can analyze gene expression data to identify pathways significantly enriched by differentially regulated genes (Dubovenko et al., 2017). However, those techniques cannot identify the enhanced or inhibited status of a pathway regulation, because pathways may have numerous negative feedback loops or negative regulatory nodes (Khatri et al., 2012) and, therefore, the pathway nodes may involve both genes with its activating and genes with inhibitory functions (Borisov et al., 2020). Thus, upregulation of an inhibitory gene means pathway downregulation, and vice versa (Buzdin et al., 2018).

On the other hand, knowledge of the individual gene product roles within a pathway can make it readable in terms of finding its activation profiles. Indeed, several techniques had been proposed, e.g., Oncofinder (Buzdin et al., 2014b), iPANDA (Ozerov et al., 2016), and Oncobox (Borisov et al., 2020) that utilize transcriptome-wide or even proteome-wide (Borisov et al., 2017) data to calculate pathway activation levels (PALs). Those are the numeric characteristics that can be used in all types of comparisons including biomarker investigations. Overall, PALs were found to be superior cancer biomarkers compared to individual gene expression levels (Borisov et al., 2014; Lezhnina et al., 2014). A number of PALs were found to be characteristic for cancer drug response (Zhu et al., 2015) and sensitivity to X-ray irradiation (Sorokin et al., 2018), asthma (Alexandrova et al., 2016), Hutchinson-Gilford progeria (Aliper et al., 2015), macular degeneration (Makarev et al., 2014), fibrosis (Makarev et al., 2016), viral infection (Buzdin et al., 2016), and aging (Aliper et al., 2016). Algorithms were developed to convert pathway activation data into the optimized selection of cancer drugs (Artemov et al., 2015; Tkachev et al., 2020) that had several recent clinical applications (Poddubskaya et al., 2018, 2019a,b; Sorokin et al., 2020b). However, those studies used manually curated/annotated pathways and were, therefore, limited by the overall number (~10 or ~100) of pathways under analysis. Thus, it is important to annotate more pathways in a universal way to obtain a large-scale overview of the human interactome.

We proposed an algorithm that identifies functional roles of the pathway components based on the pathway topology and applied it here to annotate 3,044 human molecular pathways extracted from the Biocarta, Reactome, KEGG, NCI, and HumanCYC databases, collectively covering 9,022 gene products. The resulting knowledgebase can be applied for the direct calculation of the PALs and establishing large scale profiles of the signaling, metabolic, and DNA repair pathway regulation using high throughput gene expression data.

RESULTS AND METHODS

Extraction of Molecular Pathway Data

We extracted structures of molecular pathways from the National Cancer Institute (NCI; Schaefer et al., 2009), Biocarta (Nishimura, 2001), Qiagen Pathway Central,[1] HumanCyC (Romero et al., 2004), Reactome (Croft et al., 2014), and Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa et al., 2010) databases (**Table 1**). For all the databases but Qiagen Pathway Central, the data on the pathway architecture, nodes and pairwise activation/inhibition interactions were extracted in *biopax* format. In the case of Qiagen Pathway Central database, no machine-readable format of data was available, and we manually curated data from the available graphical pathway representations (**Table 1**).

In addition to the extracted full-size pathways, we also generated a number of subsequent "micropathways" that were derivatives of the complete pathways (**Figures 1A,B**) Micropathway is a sub-graph, which contains "molecular function" node and nodes from all possible paths of length 3 including terminal "molecular function" node. Many full-size pathways have two or more terminal branches that may have different functional impact(s). We, therefore, introduced micropathways to characterize molecular processes in more detail by separately analyzing different terminal branches of the pathways. Totally, we processed 3,044 pathways including 2018 full-size, or "core" pathways, and 1,026 micropathways that covered collectively products of 9,022 human genes (**Table 1**).

[1]https://www.qiagen.com/gb/resources/resourcedetail?id=5869e38a-5033-4ccb-a281-d869893acf4e&lang=en

TABLE 1 | Statistics of the curated pathway databases.

Database	References	Number of			Data curation format
		core pathways	all pathways	unique genes	
Biocarta	Nishimura, 2001	198	337	1,082	Automated
Reactome	Croft et al., 2014	945	945	6,105	Automated
KEGG	Kanehisa et al., 2010	288	288	5,593	Automated
Qiagen	Pathway Map Reference Guide–QIAGEN, 2014	57	380	2,493	Manual
NCI	Schaefer et al., 2009	211	775	2,214	Automated
HumanCYC	Romero et al., 2004	319	319	1,038	Automated
Total number		2,018	3,044	9,022	

Number of all pathways includes core pathways and micropathways. Number of unique gene products covered by pathways from the respective database. For total number, the amount of unique gene products for all pathways is shown.

Note that number of pathway nodes was smaller than the number of genes involved in a pathway because one node could correspond to several gene products.

For several pathway components alternative gene names were used and we then converted all gene names according to the Human Genome Organization HGNC nomenclature (Povey et al., 2001).

Algorithmic Annotation of Molecular Pathways

For most of the published PAL applications, maximum five types of functional roles for gene products were comprised. These roles, described by an Activator/Repressor Role (ARR) parameter can be formulated as follows: pathway activator (ARR = 1), rather activator (ARR = 0.5), repressor (ARR = −1), rather repressor (ARR = −0.5), and gene product with uncertain or inconsistent role (ARR = 0). In the previous studies, ARR values were obtained by manually curating pathway graphs. This is however not feasible for annotating thousands of molecular pathways. We developed an original algorithm that automatically assigns ARR score values to gene products that participate in a molecular pathway.

The ARR annotation algorithm is based on the machine reading of gene product interaction graph within each pathway. Nodes correspond to gene products, and the ribs between every pair of nodes represent molecular interaction between the corresponding gene products. Each rib on the graph has a direction and is characterized by an activator or inhibitory nature of the molecular interaction it represents. For the correct calculation of ARR values, the pathway graph must be connected, wherein a weak connectivity is acceptable.

If the pathway molecular interaction graph meets these criteria, then ARR coefficients can be algorithmically assigned to the participating gene products. For the biochemical pathways, we put enzyme gene names on the pathway nodes, and the interaction ribs represented directions of the catalyzed reactions.

The algorithm used consisted of the following major steps.

i. Initialization. At this stage, a major node is algorithmically identified to be the "central" node of the pathway graph (**Figure 1C**). The major node will be used as the standard of pathway function with ARR = 1. To identify the central node, for every pathway node (V) two parameters N and M are calculated where N is the number of other nodes, which can be reached when moving from the node V, and M is the number of other nodes from which the node V can be reached. N+M, therefore, is the number of other nodes that are directly connected with the node V. The central node will be the node V_{max} for which N+M reaches the maximum value. The central node identified is then assigned with ARR = 1 value. It serves as the starting point for further recursive assignment of ARR values to the other nodes. If multiple nodes have the same maximal N+M, then V-node for a pathway is selected randomly among those "maximal" nodes. Therefore, the algorithm is suitable also for circular-organized pathways, where all nodes will have equal N+M.

ii. Recursion. For every node V, all connected nodes P_i under ARR annotation may have ribs either directed toward V (Pi → V) or outward V (Pi ← V) on the graph. During recursion, each rib can be considered only once in order to prevent endless recursion in case of cyclic interactions on the graph. If the rib has an "activator" characteristic, temporary ARRtemp = 1 is assigned to the node P_i. In contrast, if the rib has an "inhibitor" characteristic, P_i is assigned with $ARR_{temp} = -1$. Conversely, all the gene products included in the node P_i receive the same ARR_{temp} characteristics.

Let gene product GP_i belongs to node P_i. If GP_i was never previously considered in the graph traversal, $ARR = ARR_{temp(Pi)}$ for the node P_i would be assigned for GP_i. In the case when GP_i was previously considered in the graph traversal and the previously assigned ARR of it node is equal to the current $ARR_{temp(Pi)}$ then $ARR = ARR_{temp}$ would be assigned to the node P_i. If GP_i was previously considered in the graph traversal but its previously assigned ARR is not equal to $ARR_{temp(Pi)}$, then ARR is assigned to the gene product GP_i according to the following conflict resolution rule.

If a gene product GP_i with previously specified ARR or ARRs is currently considered in the graph traversal but its previously assigned ARR(s) contradict(s) with the $ARR_{temp(Pi)}$, then the conflict(s) should be resolved as follows:

1. If the signs of the previous ARR coefficient(s) and $ARR_{temp(Pi)}$ are different, then the resulting $ARR_{final(Pi)} = 0$;
2. If the difference between $ARR_{temp(Pi)}$ and any of the previous $ARRs_{(GPi)}$ does not exceed 0.5 and at least one of the ARRs is positive, the resulting $ARR_{final(Pi)} = 0.5$;
3. If the difference between $ARR_{temp(Pi)}$ and any of the previous $ARRs_{(GPi)}$ does not exceed 0.5 and at least one of the ARRs is negative, the resulting $ARR_{final(Pi)} = -0.5$.

Then the recursion R is initiated for every node P_i all of its gene products starting from the nodes proximate to the central node V. As a result, the algorithm will assign ARR values to all the connected the graph nodes and the enclosed gene products.

After the recursion finalization pathway activators will have ARR = 1, rather activators – ARR = 0.5, inhibitors – ARR = −1, rather inhibitors – ARR = −0.5, and genes with inconsistent role – ARR = 0. The recursion is stopped when a vertex with 0, 0.5, or −0.5 ARR is encountered during the traversal of the graph. This rule is needed because otherwise all vertices will have ARR 0, 0.5, or −0.5 in case of the only one ARR inconsistency found. However, this rule also may lead to exclusion of some genes described in the original source.

Therefore, the gene products included in the molecular pathway database will have the assigned ARR values representing their functional significances in the given molecular pathway. These values can be used for further calculations of the PALs according to any algorithm of PAL calculation.

Annotated Pathways Knowledgebase

We report here an ARR-curated database of 3,044 molecular pathways including 2,018 core pathways and 1,026 micropathways

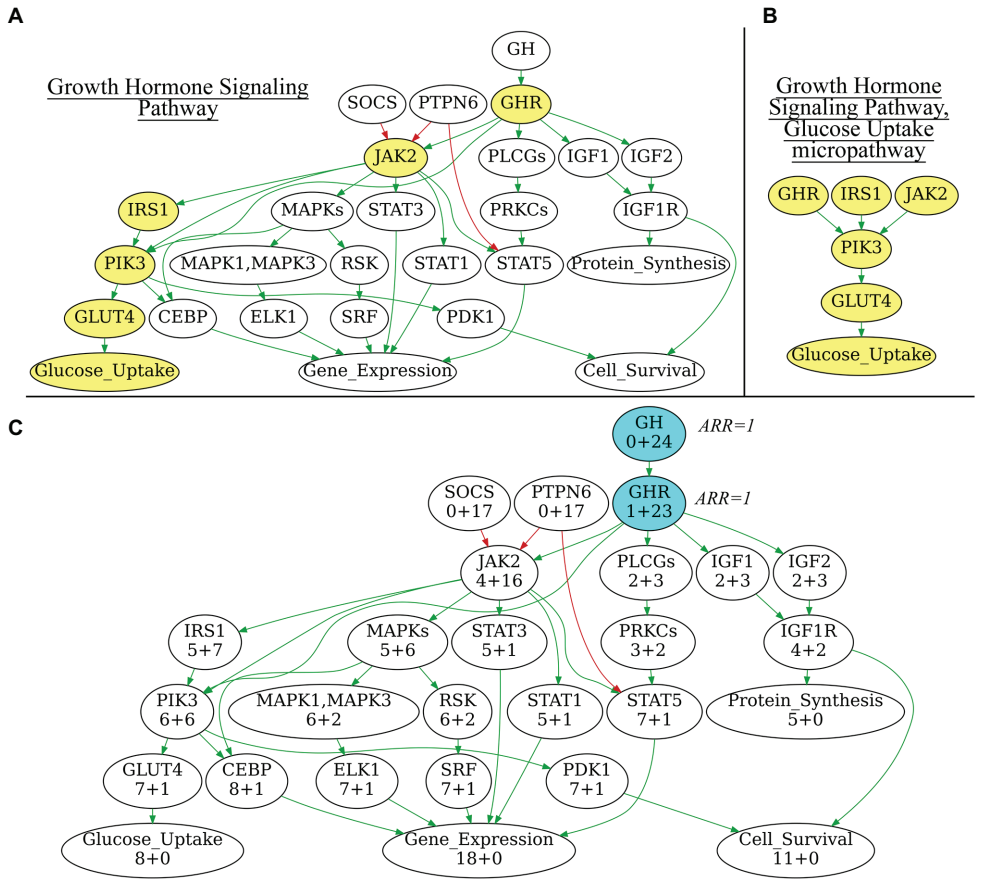


FIGURE 1 | (A) Growth Hormone Signaling Pathway with highlighted Glucose Uptake micropathway. **(B)** Glucose Uptake micropathway obtained from Growth Hormone Signaling Pathway. **(C)** N+M values for all vertices of Growth Hormone Signaling Pathway graph. The vertices with maximal N+M values are highlighted in blue, these vertices are equal *major node* candidates and get Activator/Repressor Role (ARR) = 1. Different edge colors indicate edge attribute: green is for "activation," red is for "inhibition." Structure of the Growth Hormone Signaling Pathway is derived from Qiagen Pathway Central. Yellow vertices on panel 1A indicate micropathway Glucose Uptake within Growth Hormone Signaling Pathway.

(**Supplementary Dataset 1**). The current pathway name reflects its source database and its name in the source database. For every pathway, there is a separate .csv file including the following three worksheets: (i) genes, (ii) edges, and (iii) nodes. The worksheet (i) *genes* include gene names according to HUGO Gene Nomenclature Committee (HGNC) nomenclature and the corresponding ARRs for the gene products participating in the pathway under consideration. The worksheet (ii) *edges* include information about molecular interactions between every pair of the interacting pathway nodes. Every node is defined by the names of gene products or physiological outcome(s) that form this node. The interaction type is specified as "activation," "inhibition," or "undefined," where appropriate. The worksheet (iii) *nodes* include node names and gene names corresponding to every node on the pathway graph.

It should be noted that annotation of similar pathways may be different between the source databases. For example, EGFR signaling pathway is presented in Qiagen database as "EGF_Pathway," in Reactome as "reactome_Signaling_by_EGFR_Main_Pathway" and in Biocarta as "biocarta_egf_signaling_Main_Pathway." Yet conceptually similar, all three pathways have different gene and edge compositions. In this study, we did not aim to identify inconsistences between different source databases and annotated all the pathways under their original names.

We made freely accessible software for PAL calculation using the annotated pathway database accessible following the link: https://pypi.org/project/oncoboxlib/. Algorithm is implemented as a Python library. It takes normalized (by DESeq2, quantile normalization or other) gene expression data as an input. Gene symbols should be provided in HGNC format accessible through the web-site genenames.org. At least two groups of samples are required: cases and controls, each group represented by at least one sample. Sample names should contain "Norm_" (for controls) or "Tumour_" (for cases). Output will contain PAL values for each pathway in each sample. All annotated pathway datasets mentioned in this paper alternatively can be downloaded and used for PAL calculation using the same link.[2]

[2]https://pypi.org/project/oncoboxlib/

We also provide here an example of PAL calculation for real-world data. We extracted gene expression data for gastric cancer samples (n = 16; Sorokin et al., 2020b) together with gene expression profiles of healthy stomach (n = 7) samples of patients who died in road accidents (Suntsova et al., 2019), that were sequenced using the same equipment and protocols. Cancerous samples were marked as "Tumour_" and normal samples – as "Norm_." Then we calculated PAL values (3,044 for each sample) for all molecular profiles using the above software, which produced an output file "pal.csv" (**Supplementary Dataset 1**).

DISCUSSION

We propose here the recursive algorithm for functional annotation of the molecular pathway nodes, and its application to annotation of 3,044 human molecular pathways, including signaling, metabolic, and DNA repair pathways extracted from six major pathway hubs (**Table 1**). The ARR-annotated pathways can be used for further calculations of PALs using high-throughput gene expression data, e.g., RNA sequencing or proteomic profiles (Buzdin et al., 2018; **Figure 2**). To this end, several previously published bioinformatic methods can be employed (Buzdin et al., 2014a; Ozerov et al., 2016; Borisov et al., 2020), and the PAL values returned can be applied for a variety of applications including fundamental research (Pasteuning-Vuhman et al., 2017), drug development (Aliper et al., 2017a; Ravi et al., 2018; Bakula et al., 2019), and personalized medicine (Poddubskaya et al., 2019a; Moisseev et al., 2020). Technically, PAL values can be used as the next-generation molecular biomarkers (Aliper et al., 2017b; Borisov et al., 2017; Sorokin et al., 2020c) or as the substrates for various machine learning applications (Borisov et al., 2018; Tkachev et al., 2018).

The proposed algorithm is suitable for the analysis of pathways with already established gene content and known topology of its molecular components. The algorithm can be used for agnostic

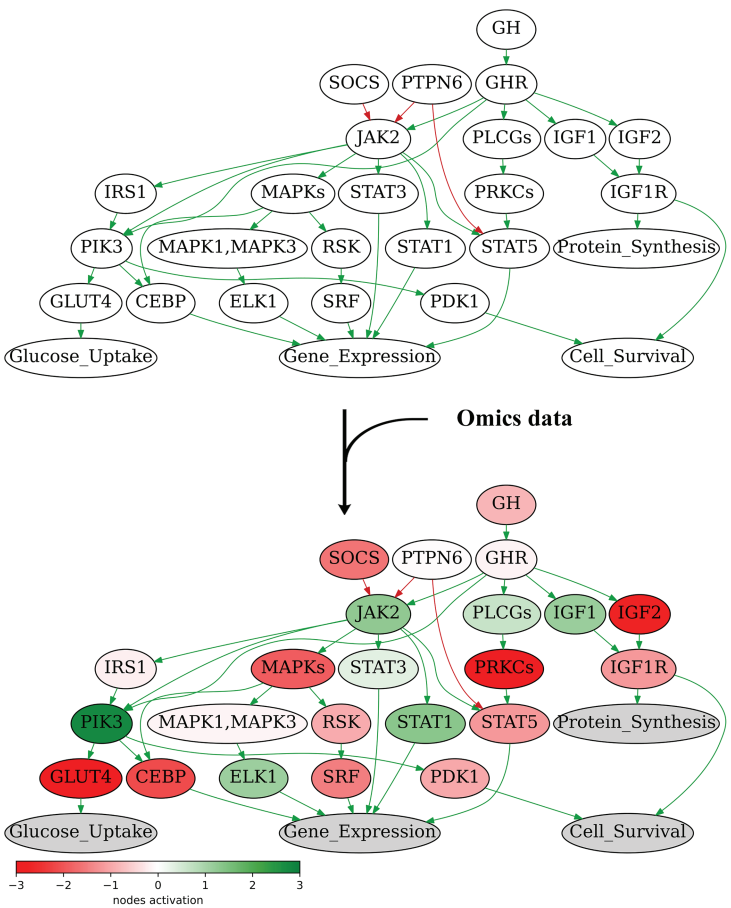


FIGURE 2 | Node activation of Growth Hormone Signaling Pathway for gastric cancer sample GC.11_S19_R1_001 from Sorokin et al. (2020b). Node activation is a sum of logarithmic case-to-norm ratio (CNR) for all genes in the node. CNR is ratio of expression levels in tumor sample and averaged normal sample. The RNA sequencing tumor profile (gastric cancer) was obtained from Sorokin et al. (2020b). The RNA sequencing profiles of normal gastric tissue were obtained from Oncobox Atlas of Normal Tissue Expression (ANTE) data (Suntsova et al., 2019). Different edge colors indicate edge attribute: green is for "activation," red is for "inhibition." Structure of the Growth Hormone Signaling Pathway is derived from Qiagen Pathway Central.

objective characterization of interacting gene networks. The underlying rationale allows reducing operator's errors and subjectivity in annotating the molecular roles of pathway components, which are inevitable in case of manual curation of the pathway graphs including hundreds of nodes. Another advantage is the pathway-centric approach during annotation, when gene product role in one pathway can be different from its role in another pathway.

The major limitations deal with the algorithm applicability only for the tasks of further calculations of pathway activation scores/ranks. Such an approach also does not address the issue of crosstalk between different molecular pathways, because all pathways are analyzed separately.

In this study, we annotated a collection of previously published human molecular pathways (**Supplementary Dataset 1**). We plan to update the current human knowledgebase annually with new releases of already included datasets and addition of new pathway collections, e.g., recently published by Wishart et al. (2020). However, the method proposed here can be used to characterize any new set of molecular pathways with the connectivity and pairwise nodes activation/inhibition information not only for the human interactome, but also for the other biological objects under investigation.

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

MS, NB, DK, and AB contributed to conception and design of the study. MS developed recursive pathway annotation algorithm. DK, AGu, MZ, and AGa manually curated the pathways and performed recursive algorithm implementations. AB, MZ, and MS wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Amazon and Microsoft Azure grants for cloud-based computational facilities. Financial support was provided by the Russian Foundation for basic research grant 19-29-01108.

ACKNOWLEDGMENTS

We thank Oncobox/OmicsWay research program in machine learning and digital oncology for software and starting pathway databases for this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.617059/full#supplementary-material

REFERENCES

Alexandrova, E., Nassa, G., Corleone, G., Buzdin, A., Aliper, A. M., Terekhanova, N., et al. (2016). Large-scale profiling of signalling pathways reveals an asthma specific signature in bronchial smooth muscle cells. *Oncotarget* 7, 25150–25161. doi: 10.18632/oncotarget.7209

Aliper, A., Belikov, A. V., Garazha, A., Jellen, L., Artemov, A., Suntsova, M., et al. (2016). In search for geroprotectors: in silico screening and in vitro validation of signalome-level mimetics of young healthy state. *Aging* 8, 2127–2152. doi: 10.18632/aging.101047

Aliper, A. M., Csoka, A. B., Buzdin, A., Jetka, T., Roumiantsev, S., Moskalev, A., et al. (2015). Signaling pathway activation drift during aging: Hutchinson-Gilford Progeria Syndrome fibroblasts are comparable to normal middle-age and old-age cells. *Aging* 7, 26–37. doi: 10.18632/aging.100717

Aliper, A., Jellen, L., Cortese, F., Artemov, A., Karpinsky-Semper, D., Moskalev, A., et al. (2017a). Towards natural mimetics of metformin and rapamycin. *Aging* 9, 2245–2268. doi: 10.18632/aging.101319

Aliper, A. M., Korzinkin, M. B., Kuzmina, N. B., Zenin, A. A., Venkova, L. S., Smirnov, P. Y., et al. (2017b). Mathematical justification of expression-based pathway activation scoring (PAS). *Methods Mol. Biol.* 1613, 31–51. doi: 10.1007/978-1-4939-7027-8_3

Artemov, A., Aliper, A., Korzinkin, M., Lezhnina, K., Jellen, L., Zhukov, N., et al. (2015). A method for predicting target drug efficiency in cancer based on the analysis of signaling pathway activation. *Oncotarget* 6, 29347–29356. doi: 10.18632/oncotarget.5119

Bakula, D., Ablasser, A., Aguzzi, A., Antebi, A., Barzilai, N., Bittner, M. I., et al. (2019). Latest advances in aging research and drug discovery. *Aging* 11, 9971–9981. doi: 10.18632/aging.102487

Borisov, N., Sorokin, M., Garazha, A., and Buzdin, A. (2020). Quantitation of molecular pathway activation using RNA sequencing data. *Methods Mol. Biol.* 2063, 189–206. doi: 10.1007/978-1-0716-0138-9_15

Borisov, N., Suntsova, M., Sorokin, M., Garazha, A., Kovalchuk, O., and Aliper, A., et al. (2017). Data aggregation at the level of molecular pathways improves stability of experimental transcriptomic and proteomic data. *Cell Cycle* 16, 1810–1823. doi: 10.1080/15384101.2017.1361068

Borisov, N., Tkachev, V., Suntsova, M., Kovalchuk, O., Zhavoronkov, A., Muchnik, I., et al. (2018). A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. *Cell Cycle* 17, 486–491. doi: 10.1080/15384101.2017.1417706

Borisov, N. M., Terekhanova, N. V., Aliper, A. M., Venkova, L. S., Smirnov, P. Y., Roumiantsev, S., et al. (2014). Signaling pathways activation profiles make better markers of cancer than expression of individual genes. *Oncotarget* 5, 10198–10205. doi: 10.18632/oncotarget.2548

Buzdin, A. A., Artcibasova, A. V., Fedorova, N. F., Suntsova, M. V., Garazha, A. V., Sorokin, M. I., et al. (2016). Early stage of cytomegalovirus infection suppresses host microRNA expression regulation in human fibroblasts. *Cell Cycle* 15, 3378–3389. doi: 10.1080/15384101.2016.1241928

Buzdin, A., Sorokin, M., Garazha, A., Glusker, A., Aleshin, A., Poddubskaya, E., et al. (2019). RNA sequencing for research and diagnostics in clinical oncology. *Semin. Cancer Biol.* 60, 311–323. doi: 10.1016/j.semcancer.2019.07.010

Buzdin, A., Sorokin, M., Garazha, A., Sekacheva, M., Kim, E., Zhukov, N., et al. (2018). Molecular pathway activation - new type of biomarkers for tumor morphology and personalized selection of target drugs. *Semin. Cancer Biol.* 53, 110–124. doi: 10.1016/j.semcancer.2018.06.003

Buzdin, A. A., Zhavoronkov, A. A., Korzinkin, M. B., Roumiantsev, S. A., Aliper, A. M., Venkova, L. S., et al. (2014a). The OncoFinder algorithm for minimizing the errors introduced by the high-throughput methods of transcriptome analysis. *Front. Mol. Biosci.* 1:8. doi: 10.3389/fmolb.2014.00008

Buzdin, A. A., Zhavoronkov, A. A., Korzinkin, M. B., Venkova, L. S., Zenin, A. A., Smirnov, P. Y., et al. (2014b). Oncofinder, a new method for the analysis of intracellular signaling pathway activation using transcriptomic data. *Front. Genet.* 5:55. doi: 10.3389/fgene.2014.00055

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Dubovenko, A., Nikolsky, Y., Rakhmatulin, E., and Nikolskaya, T. (2017). "Functional analysis of OMICs data and small molecule compounds in an integrated "knowledge-based" platform" in *Methods in molecular biology*. eds. T. V. Tatarinova and Y. Nikolsky (Humana Press Inc.), 101–124.

Ekins, S., Nikolsky, Y., Bugrim, A., Kirillov, E., and Nikolskaya, T. (2007). Pathway mapping tools for analysis of high content data. *Methods Mol. Biol.* 356, 319–350. doi: 10.1385/1-59745-217-3:319

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Junaid, M., Akter, Y., Afrose, S. S., Tania, M., and Khan, M. A. (2020). Biological role of AKT, and regulation of AKT signaling pathway by thymoquinone: perspectives in cancer therapeutics. *Mini Rev. Med. Chem.* 20. doi: 10.2174/1389557520666201005143818 [Epub ahead of print]

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). {KEGG} for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38, D355–D360. doi: 10.1093/nar/gkp896

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.* 8:e1002375. doi: 10.1371/journal.pcbi.1002375

Lezhnina, K., Kovalchuk, O., Zhavoronkov, A. A., Korzinkin, M. B., Zabolotneva, A. A., Shegay, P. V., et al. (2014). Novel robust biomarkers for human bladder cancer based on activation of intracellular signaling pathways. *Oncotarget* 5, 9022–9032. doi: 10.18632/oncotarget.2493

Ma, C. -Y., and Liao, C. -S. (2020). A review of protein-protein interaction network alignment: from pathway comparison to global alignment. *Comput. Struct. Biotechnol. J.* 18, 2647–2656. doi: 10.1016/j.csbj.2020.09.011

Makarev, E., Cantor, C., Zhavoronkov, A., Buzdin, A., Aliper, A., and Csoka, A. B. (2014). Pathway activation profiling reveals new insights into Age-related Macular Degeneration and provides avenues for therapeutic interventions. *Aging* 6, 1064–1075. doi: 10.18632/aging.100711

Makarev, E., Izumchenko, E., Aihara, F., Wysocki, P. T., Zhu, Q., Buzdin, A., et al. (2016). Common pathway signature in lung and liver fibrosis. *Cell Cycle* 15, 1667–1673. doi: 10.1080/15384101.2016.1152435

Moisseev, A., Albert, E., Lubarsky, D., Schroeder, D., and Clark, J. (2020). Transcriptomic and genomic testing to guide individualized treatment in chemoresistant gastric cancer case. *Biomedicine* 8:67. doi: 10.3390/biomedicines8030067

Nishimura, D. (2001). BioCarta. *Biotech Softw. Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Ozerov, I. V., Lezhnina, K. V., Izumchenko, E., Artemov, A. V., Medintsev, S., Vanhaelen, Q., et al. (2016). In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development. *Nat. Commun.* 7:13427. doi: 10.1038/ncomms13427

Parkhitko, A. A., Filine, E., Mohr, S. E., Moskalev, A., and Perrimon, N. (2020). Targeting metabolic pathways for extension of lifespan and healthspan across multiple species. *Ageing Res. Rev.* 64:101188. doi: 10.1016/j.arr.2020.101188

Pasteuning-Vuhman, S., Boertje-Van Der Meulen, J. W., Van Putten, M., Overzier, M., Ten Dijke, P., Kiełbasa, S. M., et al. (2017). New function of the myostatin/activin type I receptor (ALK4) as a mediator of muscle atrophy and muscle regeneration. *FASEB J.* 31, 238–255. doi: 10.1096/fj.201600675R

Pathway Map Reference Guide–QIAGEN (2014). Available at: https://www.qiagen.com/gb/resources/resourcedetail?id=5869e38a-5033-4ccb-a281-d869893acf4e&lang=en (Accessed October 12, 2020).

Poddubskaya, E. V., Baranova, M. P., Allina, D. O., Sekacheva, M. I., Makovskaia, L. A., Kamashev, D. E., et al. (2019b). Personalized prescription of imatinib in recurrent granulosa cell tumor of the ovary: case report. *Mol. Case Stud.* 5:mcs.a003434. doi: 10.1101/mcs.a003434

Poddubskaya, E. V., Baranova, M. P., Allina, D. O., Smirnov, P. Y., Albert, E. A., Kirilchev, A. P., et al. (2018). Personalized prescription of tyrosine kinase inhibitors in unresectable metastatic cholangiocarcinoma. *Exp. Hematol. Oncol.* 7:21. doi: 10.1186/s40164-018-0113-x

Poddubskaya, E., Bondarenko, A., Boroda, A., Zotova, E., Glusker, A., Sletina, S., et al. (2019a). Transcriptomics-guided personalized prescription of targeted therapeutics for metastatic ALK-positive lung cancer case following recurrence on ALK inhibitors. *Front. Oncol.* 9:1026. doi: 10.3389/fonc.2019.01026

Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., and Wain, H. (2001). The HUGO gene nomenclature committee (HGNC). *Hum. Genet.* 109, 678–680. doi: 10.1007/s00439-001-0615-0

Ravi, R., Noonan, K. A., Pham, V., Bedi, R., Zhavoronkov, A., Ozerov, I. V., et al. (2018). Bifunctional immune checkpoint-targeted antibody-ligand traps that simultaneously disable TGFβ enhance the efficacy of cancer immunotherapy. *Nat. Commun.* 9:741. doi: 10.1038/s41467-017-02696-6

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. (2004). Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* 6:R2. doi: 10.1186/gb-2004-6-1-r2

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, D674–D679. doi: 10.1093/nar/gkn653

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays - a technology review. *Nat. Cell Biol.* 3, E190–E195. doi: 10.1038/35087138

Shih, W., Chetty, R., and Tsao, M. -S. (2005). Expression profiling by microarrays in colorectal cancer (review). *Oncol. Rep.* 13, 517–524. doi: 10.3892/or.13.3.517

Sorokin, M., Ignatev, K., Barbara, V., Vladimirova, U., Muraveva, A., Suntsova, M., et al. (2020c). Molecular pathway activation markers are associated with efficacy of trastuzumab therapy in metastatic HER2-positive breast cancer better than individual gene expression levels. *Biochemistry* 85, 758–772. doi: 10.1134/S0006297920070044

Sorokin, M., Ignatev, K., Poddubskaya, E., Vladimirova, U., Gaifullin, N., Lantsov, D., et al. (2020a). RNA sequencing in comparison to immunohistochemistry for measuring cancer biomarkers in breast cancer and lung cancer specimens. *Biomedicine* 8:114. doi: 10.3390/BIOMEDICINES8050114

Sorokin, M., Kholodenko, R., Grekhova, A., Suntsova, M., Pustovalova, M., Vorobyeva, N., et al. (2018). Acquired resistance to tyrosine kinase inhibitors may be linked with the decreased sensitivity to X-ray irradiation. *Oncotarget* 9, 5111–5124. doi: 10.18632/oncotarget.23700

Sorokin, M., Poddubskaya, E., Baranova, M., Glusker, A., Kogoniya, L., Markarova, E., et al. (2020b). RNA sequencing profiles and diagnostic signatures linked with response to ramucirumab in gastric cancer. *Cold Spring Harb. Mol. Case Stud.* 6:mcs.a004945. doi: 10.1101/mcs.a004945

Suntsova, M., Gaifullin, N., Allina, D., Reshetun, A., Li, X., Mendeleeva, L., et al. (2019). Atlas of RNA sequencing profiles for normal human tissues. *Sci. Data* 6:36. doi: 10.1038/s41597-019-0043-4

Thomas, S., and Bonchev, D. (2010). A survey of current software for network analysis in molecular biology. *Hum. Genomics* 4, 353–360. doi: 10.1186/1479-7364-4-5-353

Tkachev, V., Sorokin, M., Garazha, A., Borisov, N., and Buzdin, A. (2020). "Oncobox method for scoring efficiencies of anticancer drugs based on gene expression data" in *Methods in molecular biology*. eds. K. Astakhova and S. A. Bukhari (Humana Press Inc.), 235–255.

Tkachev, V., Sorokin, M., Mescheryakov, A., Simonov, A., Garazha, A., Buzdin, A., et al. (2018). FLOating-window projective separator (FloWPS): a data trimming tool for support vector machines (SVM) to improve robustness of the classifier. *Front. Genet.* 9:717. doi: 10.3389/fgene.2018.00717

Willier, S., Butt, E., and Grunewald, T. G. P. (2013). Lysophosphatidic acid (LPA) signalling in cell migration and cancer invasion: a focussed review and analysis of LPA receptor gene expression on the basis of more than 1700 cancer microarrays. *Biol. Cell.* 105, 317–333. doi: 10.1111/boc.201300011

Wishart, D. S., Li, C., Marcu, A., Badran, H., Pon, A., Budinski, Z., et al. (2020). PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res.* 48, D470–D478. doi: 10.1093/nar/gkz861

Zheng, J., Yu, H., Zhou, A., Wu, B., Liu, J., Jia, Y., et al. (2020). It takes two to tango: coupling of Hippo pathway and redox signaling in biological process. *Cell Cycle* 19, 1–16. doi: 10.1080/15384101.2020.1824448

Zhu, Q., Izumchenko, E., Aliper, A. M., Makarev, E., Paz, K., Buzdin, A. A., et al. (2015). Pathway activation strength is a novel independent prognostic biomarker for cetuximab sensitivity in colorectal cancer patients. *Hum. Genome Var.* 2:15009. doi: 10.1038/hgv.2015.9

Conflict of Interest: MS, AGa, and AB have a financial relationship with OmicsWay Corp.

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.

The handling editor declared a shared affiliation with the authors AB, MS, and AGu at the time of review.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 April 2021
doi: 10.3389/fgene.2021.662464

Telomere Maintenance Pathway Activity Analysis Enables Tissue- and Gene-Level Inferences

Lilit Nersisyan[1,2], Arman Simonyan[1], Hans Binder[3]* and Arsen Arakelyan[1,2]*

[1] Bioinformatics Group, Institute of Molecular Biology, National Academy of Sciences, Yerevan, Armenia, [2] Pathverse, Yerevan, Armenia, [3] Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany

Telomere maintenance is one of the mechanisms ensuring indefinite divisions of cancer and stem cells. Good understanding of telomere maintenance mechanisms (TMM) is important for studying cancers and designing therapies. However, molecular factors triggering selective activation of either the telomerase dependent (TEL) or the alternative lengthening of telomeres (ALT) pathway are poorly understood. In addition, more accurate and easy-to-use methodologies are required for TMM phenotyping. In this study, we have performed literature based reconstruction of signaling pathways for the ALT and TEL TMMs. Gene expression data were used for computational assessment of TMM pathway activities and compared with experimental assays for TEL and ALT. Explicit consideration of pathway topology makes bioinformatics analysis more informative compared to computational methods based on simple summary measures of gene expression. Application to healthy human tissues showed high ALT and TEL pathway activities in testis, and identified genes and pathways that may trigger TMM activation. Our approach offers a novel option for systematic investigation of TMM activation patterns across cancers and healthy tissues for dissecting pathway-based molecular markers with diagnostic impact.

Keywords: telomere maintenance mechanisms, telomerase, alternative lengthening of telomeres, pathway signal flow, testis

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
Marco Folini,
Istituto Nazionale dei Tumori (IRCCS), Italy

****Correspondence:***
Lilit Nersisyan
l_nersisyan@mb.sci.am
Hans Binder
binder@izbi.uni-leipzig.de

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

Received: *01 February 2021*
Accepted: *16 March 2021*
Published: *07 April 2021*

Citation:
Nersisyan L, Simonyan A, Binder H and Arakelyan A (2021) Telomere Maintenance Pathway Activity Analysis Enables Tissue- and Gene-Level Inferences. Front. Genet. 12:662464. doi: 10.3389/fgene.2021.662464

INTRODUCTION

Telomeres perform protective functions at the ends of linear eukaryotic chromosomes. They constitute one of the basic molecular factors conditioning the ability of cells to divide. Excessive cell divisions lead to incomplete reconstitution of telomeres resulting in telomere shortening and loss of proper structure of telomeric ends (Blackburn, 1991; Chow et al., 2012; Martínez and Blasco, 2015). Highly proliferative cells, such as cancer and stem cells, may utilize mechanisms for preserving telomere ends despite many rounds of divisions (Greenberg, 2005). These are known as telomere maintenance mechanisms (TMMs; Hug and Lingner, 2006). The cells may trigger activation of different TMM pathways, either using the telomerase reverse transcriptase driven synthesis (TEL; Hug and Lingner, 2006; Shay, 2016), or via DNA break-induced repair (BIR) like processes, also known as alternative lengthening of telomeres (ALT; Cesare and Reddel, 2008; Neumann et al., 2013; Sobinoff and Pickett, 2017; Jia-Min Zhang et al., 2019). The TEL pathway is more commonly occurring in stem cells and the majority of cancers, while ALT is mostly activated in tumors

of mesenchymal and neuroepithelial origin (liposarcomas, osteosarcomas, and oligodendroglial gliomas), but can also be found in tumors of epithelial origin (carcinomas of the breast, lung, and kidney) (Henson et al., 2002). Some cancers such as neuroblastomas and liposarcomas, do not show evidence for activation of any of the two TMM pathways, exhibiting the ever shorter telomeres phenotype (Costa et al., 2006; Dagg et al., 2017). Finally, some indications have recently prompted that certain cancer entities (liposarcoma and other sarcomas, some tumor types) might also have both of the TMM pathways activated (Costa et al., 2006; Gocha et al., 2013).

Current experimental assays for TMM phenotyping have several shortcomings. For example, the telomeric-repeat amplification protocol (TRAP) assay for estimating telomerase activity is not very sensitive and is time and resource consuming (Fajkus, 2006). Assays to measure ALT activity are based on the assessment of chromosomal and/or cellular markers, such as C-circles, ALT-associated nuclear bodies or heterogeneous distributions of telomere length (Pickett and Reddel, 2015; Jia-Min Zhang et al., 2019), which are usually observed in ALT-type cancers. However, recent studies strongly suggest that none of these markers alone is sufficient to define the ALT status of a cell (Jia-Min Zhang et al., 2019). Attempts to use gene expression signatures for classification of TMM mechanisms have been made (Lafferty-Whyte et al., 2009). However, those signatures are applicable to specific data (Lafferty-Whyte et al., 2009) and they do not provide mechanistic details about TMM pathway activation.

Here we were set to develop a complementary approach to TMM detection that utilizes widely available gene expression data in combination with molecular interaction topologies in the TMM pathways. Establishment and analysis of TMM pathway topologies is not a trivial issue, because there is no holistic understanding of the functional context of the molecular factors, of their interactions and of the mechanisms triggering TMM activation. Previous research has identified transcriptional regulators of telomerase complex assembly (Yuan et al., 2019), however, how the enzyme components are processed and brought together (Schmidt and Cech, 2015), how the enzyme is recruited to the telomeres and what promotes final synthesis (Chen et al., 2012), is largely not clear and scattered throughout the literature in the best case. Even less is known about regulation of ALT on a gene level. Although it is considered as a break-induced repair (BIR)-like process, some of the usually accepted BIR factors are not always involved (Jia-Min Zhang et al., 2019). We previously developed a TMM-pathway approach under consideration of TEL and ALT and applied it to colon cancer (Nersisyan et al., 2019). However, overall there is no pathway representation, neither of TEL, nor of ALT TMM, which has been proven in a wider context of cells and/or tissues.

In the first part of the manuscript, we show how gene expression data can be used for TMM phenotyping making use of the TEL and ALT TMM pathways. We have constructed these pathways based on available knowledge about molecular factors and interactions involved in TMMs by further developing our previous work (Nersisyan et al., 2019). For demonstration, we have analyzed available gene expression data on different cancers with independent experimental TMM annotations (Lafferty-Whyte et al., 2009). In the second part, we apply comprehensive bioinformatics analyses to discover details of TEL and ALT activation in healthy human tissues.

RESULTS

Literature Based Reconstruction of Telomere Maintenance Pathways

In order to address the current lack of signaling pathway representations of telomere maintenance mechanisms (TMMs), we have performed a literature search to identify genes involved in TEL and/or ALT TMMs and to define their interaction partners and functional role (**Figure 1**). Overall, we identified 38 (ALT) and 27 (TEL) genes derived from 19 and 13 references, respectively (**Tables 1**, **2**). We have considered interactions among these genes in terms of pathway topologies and took into account complex formation and other molecular events with possible impact for TEL or the ALT TMM in order to describe pathway activation in time and space (**Figure 1**).

The ALT pathway is represented through a series of branches describing DNA damage and assembly of APB bodies, which then promote separation of one of the telomere strands and invasion to another telomeric template from sister chromatids, other chromosomes or extra-chromosomal telomeric sequences, followed by DNA polymerase delta assisted telomere synthesis and ultimate processing of Holliday junctions, which are formed during strand invasion (**Figure 1A**; Henson et al., 2002; Pickett and Reddel, 2015; Jia-Min Zhang et al., 2019; Sobinoff and Pickett, 2020). The ALT pathway thus takes into account not only APB formation, but also many other events leading to telomere synthesis. We have included the molecular factors involved in those events based on studies where both presence of APBs and C-circle assays were used for ALT-detection (**Supplementary Table S4**). The TEL pathway describes expression, post-transcriptional modifications, recruitment and assembly of different components of the telomerase complex, namely hTERT, hTR, and dyskerin, followed by the formation of a catalytically active telomerase complex, its recruitment to telomeres and telomere synthesis by telomerase and DNA polymerase alpha (**Figure 1B**; Hug and Lingner, 2006; Tseng et al., 2015; Rice and Skordalakes, 2016).

Our pathway model is based on the following general assumptions: (i) TMM pathways are subnets of a global network of cellular processes. We only include genes that have unambiguous effect on either TEL or ALT pathway. (ii) Alternative splicing and/or post-transcriptional regulatory effects mediated, e.g., via transcription factors, non-coding RNA (miRNA or lncRNA) or interactions on protein level, are not explicitly considered in our TMM pathways. We assume that these factors to a certain degree are implicitly taken into account by the measured expression levels of the genes in the TMM pathways. (iii) The TMM pathways are treated as directed graphs to mimic signal transduction from source to sink as described below.

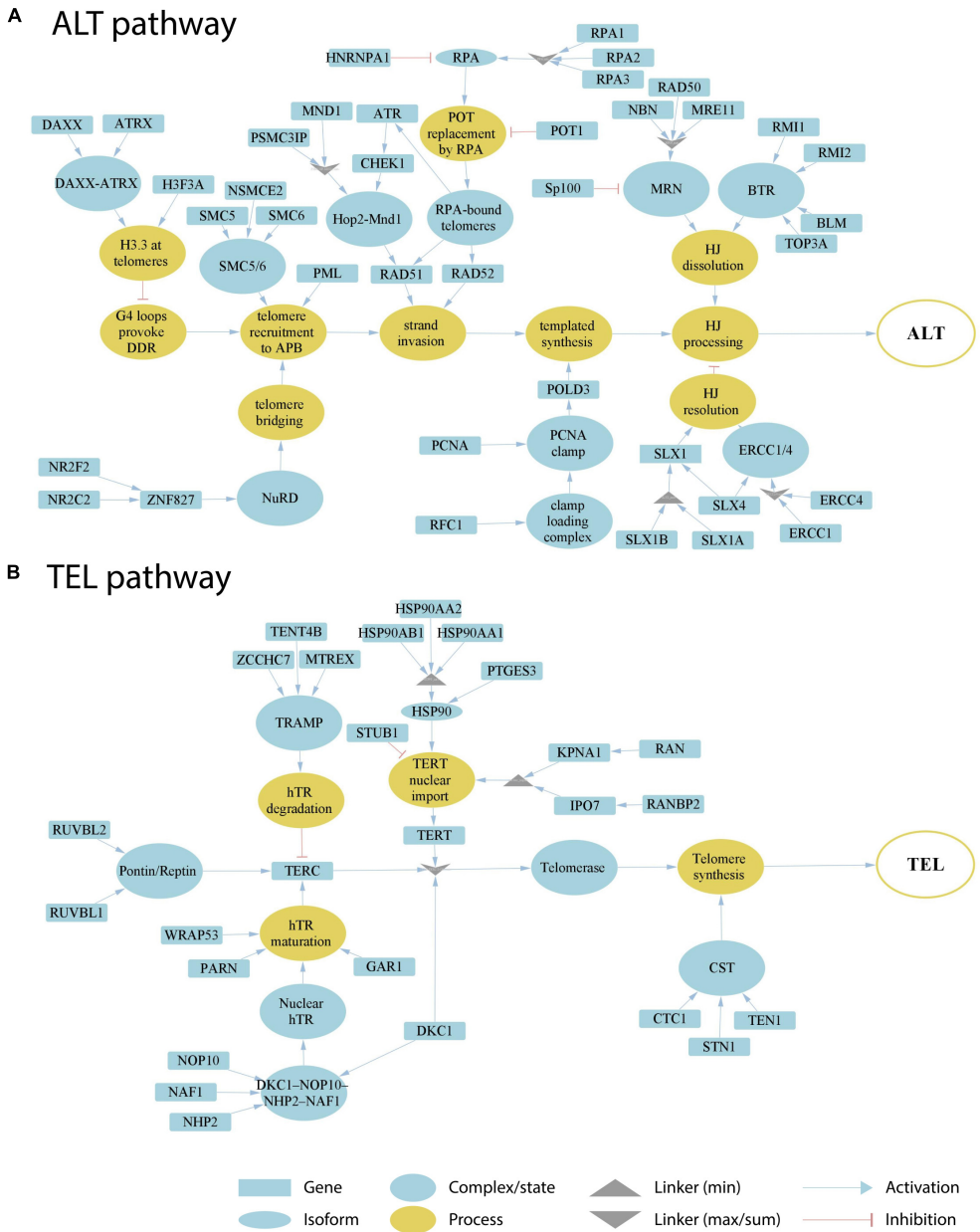


FIGURE 1 | Literature based reconstruction of the ALT **(A)** and the TEL **(B)** pathways of telomere maintenance. The ALT and TEL pathways include 37 and 26 genes based on 19 and 13 unique citations, respectively. The events leading to telomere maintenance in each pathway converge at respective final sink nodes (open circles at the right). Types of nodes and edges are defined in the figure. Linker nodes apply different operators for describing signal transduction by taking minimum or maximum values of all the input signals or their sum (see legend in the figure).

TMM Pathway Activities Are Supported by Experimental Assays

For assessment of TEL and ALT TMM pathway activities in a given sample we use gene expression data and the pathway signal flow (PSF) algorithm as implemented in Cytoscape (apps *PSFC* and *TMM*; see "Materials and Methods" section for details). The algorithm computes the PSF score in each of the pathway nodes by considering signal propagation through all upstream activating and inhibiting interactions and complex and linker node types making use of the fold change (FC) expression values of the involved pathway genes with respect to their mean expression in the respective data set. The PSF score thus reflects the activity of all upstream events and of their topology in contrast to gene set overexpression measures often used alone for functional assessment (Nersisyan et al., 2014, 2017). The PSF scores of the final sink nodes then estimate the overall activity of the TEL and the ALT pathway, respectively.

TABLE 1 | The ALT pathway nodes with activating (+) or inhibiting (–) effects.

Name (alias)	Effect	References	Name (alias)	Effect	References
G4 formation; DDR provocation			**Template directed synthesis**		
H3F3A	–	Lovejoy et al., 2012; Clynes et al., 2015; Dyer et al., 2017	POLD3	+	Dilley et al., 2016
ATRX/DAXX	–		PCNA	+	
DAXX	–		RFC1	+	
ATRX	–		**Holiday Junction (HJ) processing (HJ dissolution)**		
Telomere bridge formation (NTB)			MRN complex	+	Dimitrova and de Lange, 2009; Clynes et al., 2015
NR2C2 (TRF4)	+	Conomos et al., 2014	MRE11	+	Lafrance-Vanasse et al., 2015
NR2F2 (COUP-TF2)	+		RAD50	+	
ZNF827	+		NBS1	+	
NuRD complex	+		SP100	–	Jiang et al., 2005
Recruitment of telomeres to APBs (APB)			BTR complex	+	Sobinoff et al., 2017; Min et al., 2019
PML	+	Chung et al., 2012	BLM	+	Sobinoff et al., 2017
SMC5/6 complex	+	Potts and Yu, 2007; Aragón, 2018	TOP3A	+	
SMC5 (RAD18)	+		RMI1	+	
SMC6 (Spr18)	+		RMI2	+	
NSMCE2 (NSE2)	+		**Holiday Junction (HJ) processing (HJ resolution**		
Strand invasion (SI)			SLX4-SLX1-ERRC4 complex	–	Sobinoff et al., 2017
POT1	–	Flynn et al., 2012	SLX1	–	
RPA	+		SLX4	–	
RPA1	+		SLX1A	–	
RPA2	+		SLX1B	–	
RPA3	+		ERCC1	–	Zhu et al., 2003
HNRNPA1	–		ERCC4	–	
ATR	+	Flynn et al., 2012, 2015; Deeg et al., 2016; Dilley et al., 2016			
CHEK1	+	Dilley et al., 2016			
RAD51	+	Cho et al., 2014; Jia-Min Zhang et al., 2019			
HOP2	+				
MND1	+				
RAD52	+	Jia-Min Zhang et al., 2019; Min et al., 2019			

Application to two publicly available microarray gene expression datasets from cell lines and liposarcoma tissues delivers an ALT/TEL-PSF data couple for each sample, which is then plotted into an ALT-*versus*-TEL PSF coordinate systems (**Figures 2A,B**). Each sample is color-coded according to its assignment to TEL^+/ALT^- or TEL^+/ALT^- or double negative TEL^-/ALT^- phenotypes which were determined by independent experimental assays (APB assay for ALT and TRAP assay for TEL) alongside with the gene expression measurements. Using support vector machine (SVM) learning, TEL positive and negative samples were separated by a vertical line, and ALT positive and negative samples by a horizontal line in both data sets (**Figures 2A,B**).

Detailed inspection revealed that all double negative healthy mesenchymal stem cells and mortal cell lines (green circles) indeed locate in the ALT^-/TEL^- quadrant formed by the perpendicular lines of our classification scheme thus indicating agreement with experimental assays. For single positive ALT^+/TEL^- (red) and ALT^-/TEL^+ (blue) samples one finds their accumulation in the respective top-left and down-right quadrants, respectively, as expected. A certain fraction of theses phenotypes is found "displaced" in the left-down and top-right quadrants assigning them to double negative and double positive TMM cases, respectively. For some samples, we observed discordance in TMM PSF values between technical replicates, which was also noticeable on the level of gene expression (**Supplementary Data 2**), suggesting possible technical issues during microarray processing. Overall, we obtained 80–85% agreement between our computational assessment of TMM activity as ALT or TEL single positive,

TABLE 2 | The TEL pathway nodes with activating (+) or inhibiting (–) effects.

Name (alias)	Effect	References	Name (alias)	Effect	References
Nuclear TERT (TERT activation and recruitment)			**Telomerase assembly**		
TERT	+	Cohen et al., 2007	TERT	+	Cohen et al., 2007
KPNA1	+	Jeong et al., 2015	TERC	+	Cayuela et al., 2005; Cohen et al., 2007
RAN	+		DKC1	+	Schmidt and Cech, 2015
IPO7	+	Frohnert et al., 2014	Pontin/Reptin	+	Venteicher et al., 2008
RANBP2	+		RUVBL1	+	
HSP90	+	Jeong et al., 2015	RUVBL2	+	
HSP90AA1	+		**Recruitment to telomeres and synthesis**		
HSP90AA1	+		CST complex	+	Chen et al., 2012
HSP90AB1	+		CTC1	+	
PTGES3	+		STN1	+	
STUB1	–	Lee et al., 2010	TEN1	+	
hTR maturation					
TERC	+	Cayuela et al., 2005; Cohen et al., 2007			
PARN	+	Moon et al., 2015; Boyraz et al., 2016			
TRAMP	–	Boyraz et al., 2016			
TENT4B (PAPD5)	–	Moon et al., 2015; Tseng et al., 2015			
MTREX (MTR4)	–	Tseng et al., 2015			
AIR2 (ZCCHC7)	–				
DKC1-NOP10-NHP2-NAF1	+	Schmidt and Cech, 2015			
DKC1	+				
NOP10	+				
NHP2	+				
NAF1	+				
GAR1	+				
WRAP53	+	Chen et al., 2018			

and full agreement between double negative samples and the experimental annotations taken from the original publication.

Analysis of Pathway Activation Patterns at Single Gene and Single-Sample Resolution

For a better understanding of particular reasons of the diversity observed in the ALT/TEL plots we visualized pathway activity patterns for selected samples in **Figures 2C,D** by coloring the major event nodes in light-to-dark blue (TEL PSF) or red (ALT PSF) (the nodes were annotated in the part **Figures 2C,D**; the full gallery of pathway activation patterns for all samples is provided in **Supplementary Figures S2,S3**).

SUSM1 (A4), a transformed ALT-activated (according to experimental assignment) cell line derived from fetal liver, shows activation of all the nodes involved in TEL and ALT (**Figure 2C**), which pushes one replicate to the upper right quadrant corresponding to the double positive TEL^+/ALT^+ pathway phenotype. The C33 cell line (T2), derived from cervical squamous cell carcinoma, shows clear activation of the major TEL pathway nodes, while the main ALT events, such as telomere recruitment to APB, strand invasion and telomere synthesis, are suppressed, pushing it to the TEL^+/ALT^- quadrant in agreement with experimental assignment (**Figure 2C**). The bladder carcinoma cell line 5,637 (T4), experimentally assessed to have high telomerase activity, showed relatively low TEL PSF values. As seen in **Figure 2C**, the expression and processing of the two telomerase complex subunits hTERT (TEL pathway branch TERT) and dyskerin (branch DKC1) were highly activated in this cell line, however, the expression of the factors processing the RNA template hTR was low, thus being a bottleneck for the telomerase complex formation, explaining the observed discrepancy. We could also identify the genes responsible for this, as described in detail below.

Looking at the patterns of the A7 sample in the dataset of liposarcoma tissues (**Figure 2D**), we observed high activity of the APB branch of the ALT pathway, indicating that both the computational annotation and the experimental assay point on accumulation of APB bodies in this tissue. However, as the expression of downstream factors in the ALT pathway was low (**Figure 2D**), it compromised the ultimate activation of this pathway. This result supports recent studies showing that the mere presence of APB bodies doesn't necessarily lead to the activation of ALT (Jia-Min Zhang et al., 2019). Similar branch-activation patterns were observed for the A6 and A9 tissue samples (**Supplementary Figure S3**). It is important to note that these samples had high TEL pathway activity, which is in line with previous observations of high false-positive rate of the APB assay associated with overexpression of *TERC* or *TERT* (Henson and Reddel, 2010). In summary, inspection of the PSF patterns along the pathways identifies genes and branches contributing to activation or deactivation of ALT and TEL with single sample

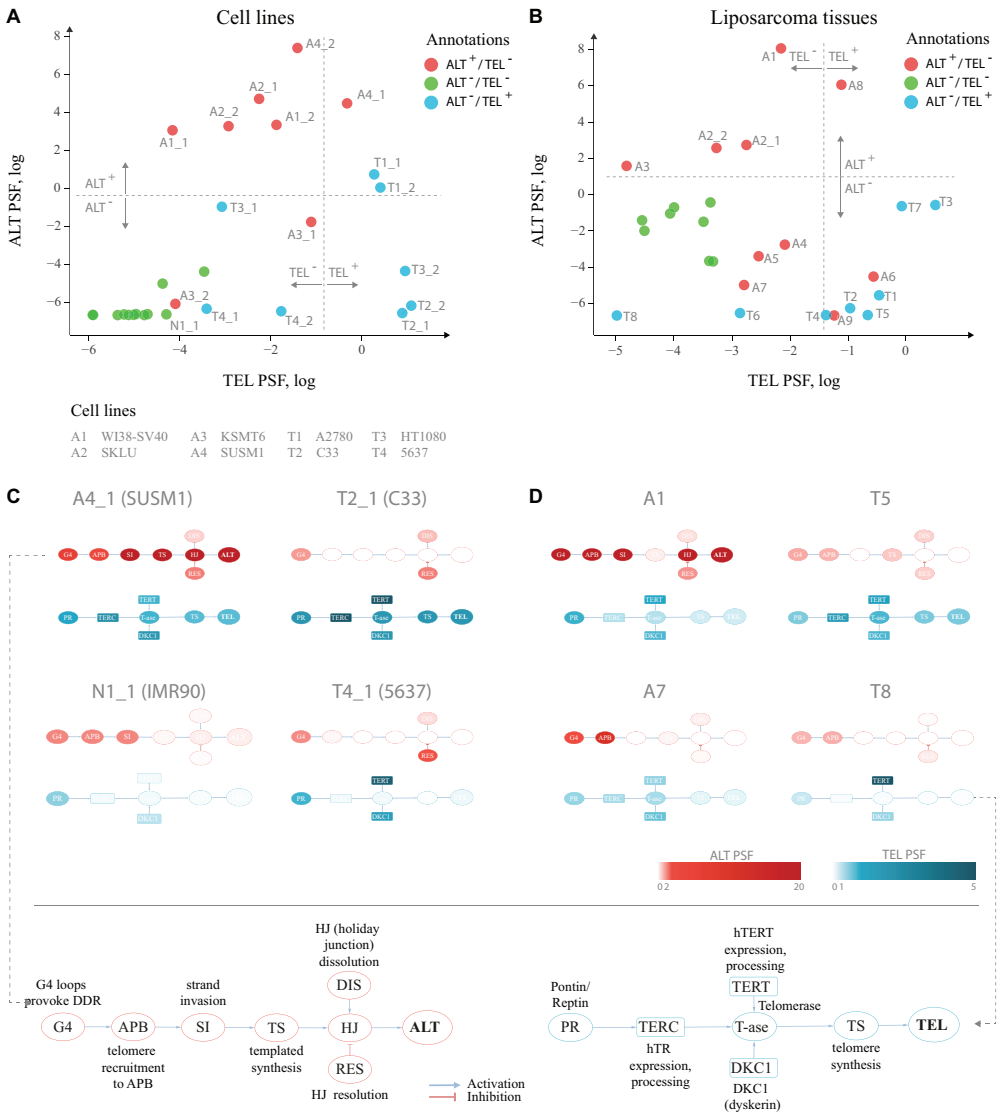


FIGURE 2 | TEL and ALT pathway activity plots for cell lines and liposarcoma tissues. **(A)** Cell lines and hMSCs plotted according to the TEL PSF (*x*-axis) and ALT PSF (*y*-axis) values. Color coding corresponds to experimental annotation of TMM states. Horizontal and vertical dash lines separate ALT$^+$ from ALT$^-$ and TEL$^+$ from TEL$^-$ experimentally annotated samples based on support vector machine classification on the ALT and the TEL PSF values. Technical replicates are distinguished with _1 and _2 suffices. **(B)** Similar representation for liposarcoma tissues and hMSCs. **(C,D)** Examples of pathway activation patterns of some cell lines **(C)** and liposarcoma tissues **(D)**. PSF values of process nodes are indicated on a light-to-dark color scale. Node abbreviations are explained in the bottom schemes. Pathway activity patterns for all the samples are shown in **Supplementary Figures S2,S3**.

resolution. More information regarding gene-level activation patterns can be explored in the full pathway PSF activation patterns (**Supplementary Data 3**).

Partial Influence (PI) Analysis Identifies Gene-Specific Triggers of Pathway Activities

As a second additional option of extracting gene level information responsible for pathway activation changes, we analyzed the partial influence (PI) of each gene on TMM pathway activities ("Materials and Methods" section). PI enables understanding of which genes act as triggers to activate or to deactivate selected nodes in the pathways. Genes, increasing or decreasing the PSF of the target node have positive or negative PI's, respectively. Activating nodes with log fold change (FC) expression above or below zero in a given sample will thus have a positive or negative PI, respectively. Nodes with inhibitory effect, on the other hand, will have the reverse association with PI (**Figure 3A**).

We have examined the PIs in the TEL pathway of the 5,637 (T4) bladder carcinoma cell line (**Figure 2A**), where branch-level analyses showed suppression of the hTR maturation branch (**Figure 2C**). The results of the PI analysis show that the genes

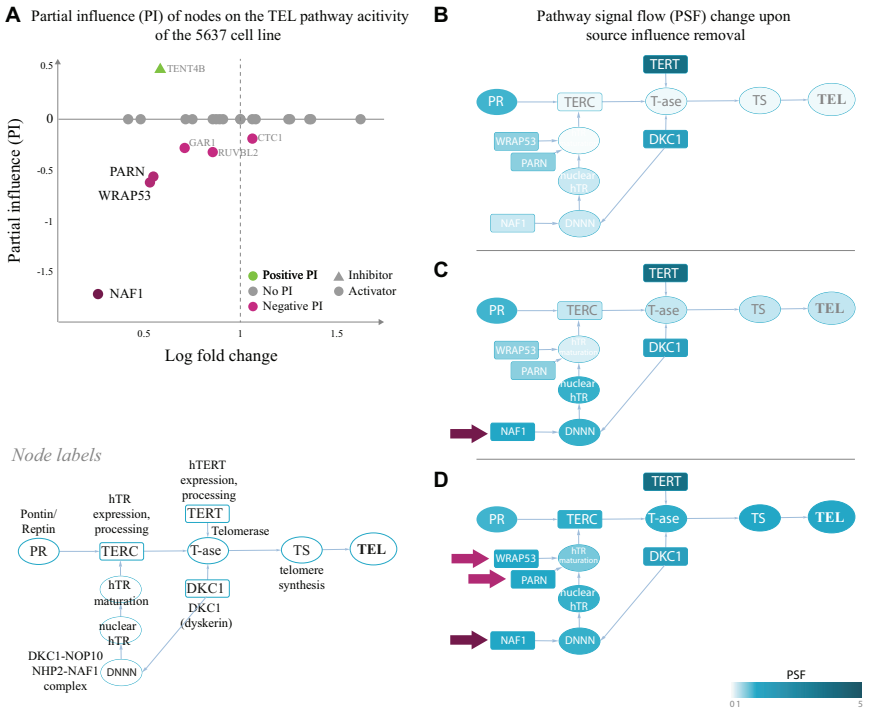


FIGURE 3 | Partial influence of nodes on the TEL pathway activity of the 5,637 cell line. **(A)** Partial influence (PI) of each node is computed as the difference of the TEL node PSF value when the node is set to a fold change (FC) value of 1. In the first replicate of the 5,637 cell line, the genes involved in maturation of the telomerase RNA component hTR (*NAF1*, *WRAP53,* and *PARN*), have the largest negative influence on the TEL pathway activity. Low expression of these genes leads to inactivation of the TERC branch, which is the bottleneck in telomerase complex formation **(B)**. The pathway becomes activated when the influence of *NAF1* **(C)** or *NAF1, WRAP53, and PARN* **(D)** is removed by setting their FC to 1 (see section "Materials and Methods").

encoding the hTR maturation factors *NAF1, WRAP53* and *PARN* (**Figure 3**) were responsible for low activity of the hTR branch in this sample. Indeed, bringing the relative expression of *NAF1* to a fold change value of 1 (**Figure 3B**) was sufficient to push the activity of the TEL pathway over the threshold for classifying it as TEL^{+} (log PSF of 0). It is important to mention that the used microarray gene expression datasets did not contain expression values for the *TERC* gene itself, and only the expression of hTR processing factors contributed to the hTR branch in this case.

A similar branch activation and PI pattern was observed for the liposarcoma tissue T8, where the TMM-PSF diverged from the experimental assay results (**Figure 2B** and **Supplementary Figure S6**). In the other misclassified sample (T6), we observed high PSF activity of the telomerase complex, however, the low TEL pathway activity was driven by low expression of STN1, which stimulates polymerase alpha in synthesis of the complementary telomeric strand after the action of the telomerase complex (**Supplementary Figure S6**).

PI analysis also identified that various genes were responsible for activation of the APB branch of the ALT pathway in the A6 and A7 liposarcoma tissues, while down-regulation of RAD51 and CHEK1 led to suppression of the downstream strand invasion events leading to low ALT PSF activity in both samples (**Supplementary Figure S5**). Hence, PI analysis extracts genes which act as triggers for switching TMM on or off with possible impact for altering between ALT and TEL and *vice versa*.

Comparison With TelNet Genes

Our curated TMM-pathway approach considered in total 63 genes extracted from recent publications (see above and "Materials and Methods" section). As an alternative option we made use of TelNet database (Braun et al., 2018) which collected 2,094 genes with impact for telomere biology and extracted 336 genes annotated as ALT- or TEL-associated (**Supplementary Table S3**). Separate hierarchical clustering of the expression values of the TEL- and ALT-genes in the cell line and liposarcoma samples analyzed above, well separates double negative ALT^{-}/TEL^{-} from the single positive ALT and TEL samples on one hand, and also the latters each from another. Between 82% and 85% of the samples were properly annotated compared with the experimental annotations. Agreement with PSF-based annotations is high (96 and 100%, respectively). Hence, gene clustering and PSF based classifications and experimental annotations are well-aligned (**Figure 4**), and those samples that were misclassified by the PSF algorithm were also misclassified with the TelNet gene set clustering. This simple comparison served as an independent validation for the selection of genes in our TMM-PSF approach using TelNet. Note that the overlap between both collections is 42 genes, meaning that 67% of the TMM-pathway genes are considered in TelNet what we attribute to our more recent curation. Moreover, the TMM pathway approach clearly makes use of a markedly reduced number of genes after strict curation and,

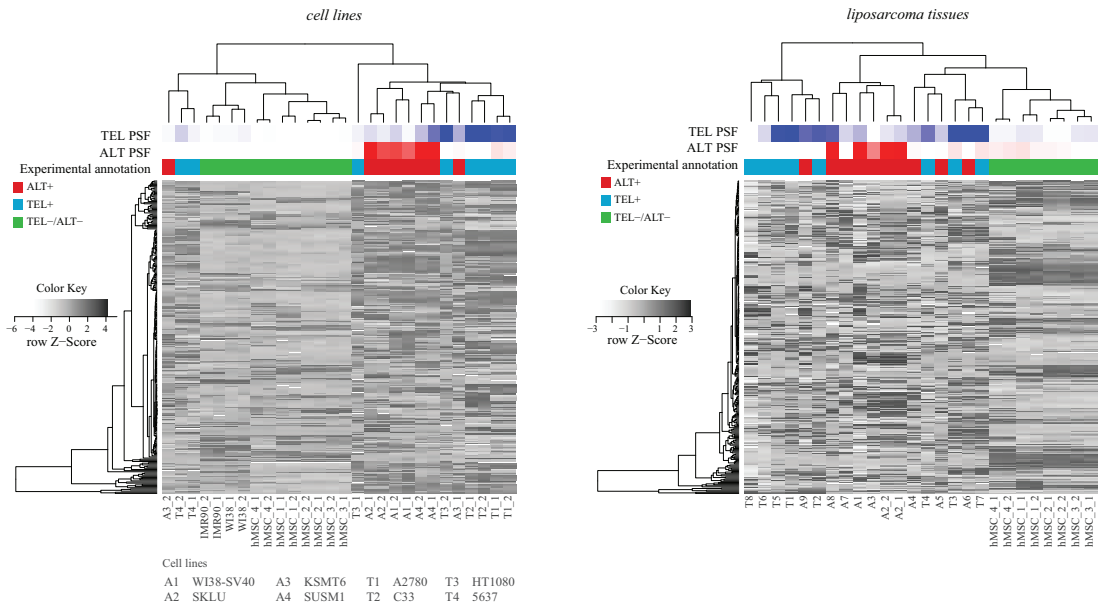


FIGURE 4 | Hierarchical clustering of cell lines and tissues based on expression of genes retrieved from the TelNet database. TEL and ALT PSF values are depicted with blue and red color gradients, and experimental annotations are drawn below. PSF and TelNet assignments closely agree each with another.

moreover, enables topology-based analysis not only to extract genes, but also relevant branches and triggers for switching between TMM pathways.

We have also explored the TMM related 12 gene sets from the C5 collection of the Molecular Signatures Database[1]. Together they contained 138 unique genes, either involved in the pathways curated by us or in the TelNet database with ambiguous role related to TMM and not included in the pathways by design (see section "Materials and Methods").

It is worth mentioning that gene sets are typically applied using enrichment or overexpression signature metrics, which are conceptually different but complementary to the PSF-approach that explicitly considers interaction topologies (see Nersisyan et al., 2014 for details). We have previously applied the TMM PSF method together with several telomere maintenance gene signatures such as the Reactome "extension of telomeres" or "telomere maintenance" and found that the latter were able to estimate the overall activity of the TEL TMM, however without providing details of the activating mechanisms as opposed to our approach (Nersisyan et al., 2019). Pathway analysis thus provides an alternative option of estimating TMM activity with an overall detailed resolution of involved sub-processes affecting telomere maintenance.

Telomere Maintenance in Healthy Human Tissues

We have previously applied our PSF approach to study telomere maintenance states in Lynch syndrome and sporadic colorectal cancer subtypes (Nersisyan et al., 2019). Here we expand this method to evaluate the state of telomere maintenance mechanisms in healthy human tissues.

We estimated TEL and ALT activities in units of PSF across a series of fifteen tissues (overall N = 17–80 samples per tissue taken from donors evenly distributed across age and sex groups) making use of RNA-seq data taken from the GTEx portal (**Figure 5**). The TEL pathway activity was generally low across the tissues (**Figure 5A**), while ALT shows slightly enhanced PSF-values (**Figure 5B**). Interestingly, testis showed considerably increased activity of ALT consistently in all of the samples, and also activation of TEL, however, only in half of them, while the other half showed virtually deactivated TEL (**Figures 5A,B**).

For a more detailed view on the TMM activation patterns in testis, we stratified the PSF-values of the full sample set (N = 129) according to the age of the donors (**Figures 5C,D**). No clear aging-trend of TEL and ALT PSF was observed except for a shift of the TEL-PSF distribution toward larger values for younger men of age 20–29 years (Mann-Whitney U test p = 0.01). On the other hand, we observed a clear bi-modal distribution of samples with high or low TEL pathway activities regardless of age (**Figure 5E**). The detailed analysis of TEL pathway activation patterns in terms of PI showed that the observed variability was mainly driven by the expression of *TERC* (**Figure 5F**). This gene was not expressed in half of the testis samples leading to very low TEL activities. Importantly, *TERC* was generally found as the main limiting factor for TEL pathway activity across all the tissues studied, meaning that low *TERC* expression strongly downscales TEL-PSF.

PI analysis of the ALT pathway in testis mainly implicated the importance of RAD51, suggesting RAD51-dependent ALT activation. In the RAD51-dependent pathway, telomere-bound POT1 is replaced by RPA, followed by RAD51 recruitment

[1]https://www.gsea-msigdb.org/gsea/msigdb

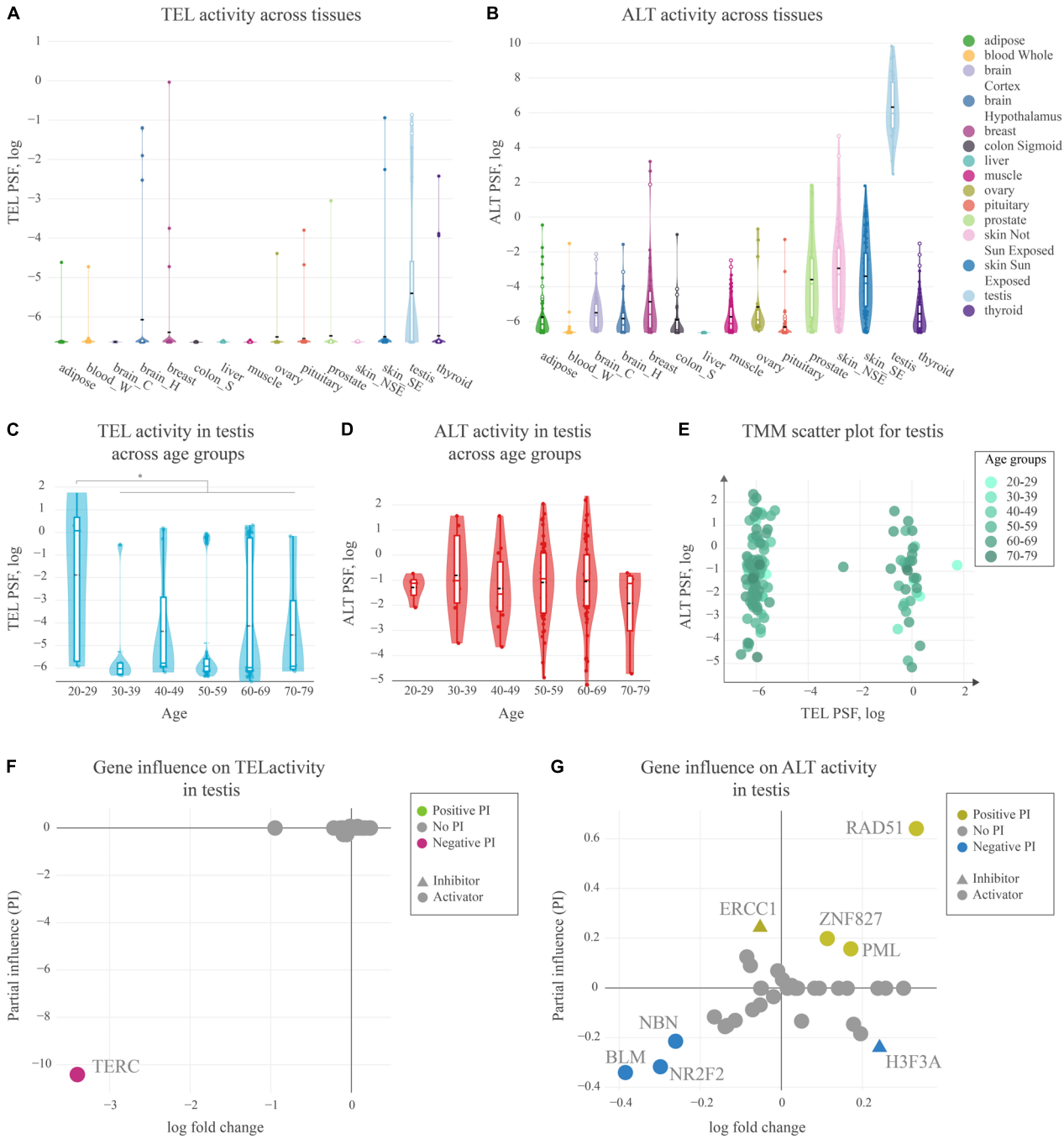


FIGURE 5 | Activity of telomere maintenance mechanisms in healthy human tissues. **(A)** TEL pathway activity in healthy human tissues. The black dashed lines represent means, the colored lines represent medians. **(B)** ALT pathway activity in human tissues. **(C)** TEL pathway activity in testis across age groups. **(D)** ALT pathway activity in testis across age groups. **(E)** ALT and TEL pathway activity scatter-plot in testis colored by age-groups. **(F)** Partial influence (PI) of TEL pathway genes on the pathway activity. **(G)** PI of ALT pathway genes on the pathway activity.

and strand invasion. Low expression of BLM helicase and of the nuclear receptor NR2F2 potentially limits hyperactivity of ALT in this tissue (Conomos et al., 2014; Sobinoff et al., 2017; Zhang et al., 2021), although it has recently be shown that in some cases NR2F2 may not be directly involved in ALT (Alhendi and Royle, 2020). Interestingly, a recent study of Episkopou et al. (2019) have identified that the expression of the testis-specific Y-encoded-like protein 5 (*TSPYL5*), a previously unrecognized APB body component, is crucial for survival of ALT^{+} cells, as it protects replaced POT1 from proteasomal

degradation. We also found high expression of *TSPYL5* in our testis samples showing slight correlation with ALT activity (Pearson $R = 0.3$, **Supplementary Figure S7**). This supports possible involvement of *TSPYL5* in maintaining viability of ALT^+ cells in healthy human testis.

In summary, our analyses show that both TEL and ALT pathways of telomere maintenance may be activated in human testis. In this tissue, the main driver for ALT pathway activation is *RAD51*, while the limiting factor for TEL pathway activity is the expression of *TERC,* which was observed only in a subset of samples, notably more pronounced in younger subjects.

DISCUSSION

The activation of telomere maintenance mechanisms can serve as a phenotypic biomarker for prognostic purposes and for choosing chemotherapies, e.g., by direct targeting of the TMM pathways (Villa et al., 2008; Sugarman et al., 2019; Chen et al., 2020). However, little is known about molecular triggers leading to activation of telomere maintenance mechanisms via the TEL or the ALT pathways. Some cancer tissues are prone to activation of the ALT pathway, such as the mesenchyme-originating liposarcomas, osteosarcomas and glioblastomas, while others are more permissive of TEL activation (Henson et al., 2002). At the same time, some tumors do not activate neither TEL nor ALT TMM pathways, while others may activate both or switch activation from TEL to ALT or vice versa (Costa et al., 2006; Gocha et al., 2013). The molecular mechanisms behind such cellular decisions are mostly unknown. Owing to the role of TMM in cancer prognosis and the promise of telomere-targeting therapies (Villa et al., 2008; Sugarman et al., 2019; Chen et al., 2020), it is of paramount importance to investigate these activation patterns, as well as to come up with better approaches to assess or predict TMM states of tissues and individual cells.

Our study aimed at combining information about molecular factors involved in the TMMs to study the mechanisms of activation of either the TEL or the ALT pathways in cancers and healthy human tissues, and to provide a complementary bioinformatics method for identification of TMM phenotypes from gene expression data. To reach this goal, we have reconstructed TEL and ALT pathways of telomere maintenance taking into account, first of all, comprehensive review and recent original articles of the last 3 years (**Supplementary Tables S1,S2**). To the best of our knowledge this is the first attempt of providing a holistic view and quantitative analysis of signaling events involved in TMM.

The TMM pathway topologies proved by quantitative assessment of the TEL and ALT TMMs activity in cancer cell lines and tissues based on two gene expression data taken from a previous study (Lafferty-Whyte et al., 2009). Our pathway approach not only considered the expression values of the genes in each pathway, but also their mutual (activating or inhibiting) interactions, complex formation, as well as linking operator nodes, alltogether potentially influencing the final pathway activity states. According to the combinations of the activity of TEL and ALT, we have classified the samples into four TMM phenotypes (TEL^-/ALT^-, TEL^+/ALT^-, TEL^-/ALT^+, TEL^+/ALT^+) in 80–85% agreement with independent experimental annotations of TMM in the samples. The absence of *TERC* expression data due to the lack of microarray probes for this gene may have limited the accuracy of the TEL pathway activity estimation in part of the samples.

Our TMM-PSF method stands out with a couple of advantages: (a) it helps to easily annotate samples based on ALT/TEL pathway activity values, and (b) it provides molecular details for dissecting the role of genes and sub-events in the overall activation of the pathway. For example, we could show that for some samples despite their low ALT activity, the APB-pathway branch was highly active, which may explain why these samples were detected as ALT positive in the independent APB assay. This is in agreement with recent studies showing that the existence of APBs does not ensure telomere synthesis and many APBs in the cell may lack ALT activity (Jia-Min Zhang et al., 2019). In consequence APB-based ALT tests may lead to false positives in samples with telomerase overexpression (Henson and Reddel, 2010).

It is important to note that there is no gold-standard method for TMM phenotyping of cells/tissues. All currently available experimental assays have their drawbacks (Pickett and Reddel, 2015; Jia-Min Zhang et al., 2019). The activity of the telomerase enzyme is usually assessed by the TRAP assay that measures the amount of DNA synthesized from a telomere-like template *in vitro*. However, it has low sensitivity, is not well suited for single-cell analysis and does not account for downstream processes, such as recruitment of the enzyme to the telomeres and synthesis of the complementary strand (Fajkus, 2006). ALT activity is assessed by assays that measure the abundance of extra-telomeric C-circles or of APB's, or heterogeneity of telomere length (Sobinoff and Pickett, 2017). However, while these biomarkers may be common in many ALT positive cells/tissues, it has been shown that they alone are not sufficient for promoting ALT (Jia-Min Zhang et al., 2019). New methods for direct monitoring of telomere elongation in ALT cells are still being adapted (Verma et al., 2018). There is an urging need for novel assays to detect TMM states that could help in understanding cellular response to chemotherapies and for development of TMM targeted therapies (Villa et al., 2008; Sugarman et al., 2019; Chen et al., 2020; Recagni et al., 2020). Particularly useful will be assays that allow for TMM assessment in single cells. As it has been shown previously, and confirmed in this study, both TEL and ALT pathway may be co-activated in the same tumor. It will be extremely important to assess this issue by single cell transcriptomics and our pathway approach whether those pathways co-exist in the same cell or show mosaic activation in the tissue (Costa et al., 2006; Gocha et al., 2013). Additionally, switching from TEL to ALT as a cellular response to a therapy is also possible (Gan et al., 2002; Bechter et al., 2004; Shay et al., 2012; Recagni et al., 2020). It is possible that the poorer agreement with experimental assays in the liposarcoma tissue samples, compared with the cell lines, was caused by the presence of different cell types in those tissues. In this sense, using RNA-seq gene expression

data to assess TMM activity from single cells is a promising future direction.

We have previously applied our approach to dissect molecular factors involved in TMM activation in Lynch syndrome and sporadic colorectal cancer subtypes in order to study association of ALT with microsatellite and chromosomal instability in those cancers (Nersisyan et al., 2019). Here we expanded beyond these studies to investigate TMM activation patterns in healthy human tissues. The experimental assays have so far been used to detect ALT activity in cancers only. Previous studies using telomeric DNA tagging have found evidence for telomere elongation via the ALT pathway in the absence of high telomerase activity in mammalian somatic tissues during early development (Liu et al., 2007; Neumann et al., 2013). A recent study by Novakovic et al. (2016) has identified elevated TERRA levels, elongated telomeres and some evidence for ALT-specific C-circles in human placenta. The authors note, however, that the amount of C-cicles was low compared to ALT positive cancers, suggesting that a mild ALT phenotype may exist in specific placental cells. Overall, these studies show that more sensitive methods to detect low ALT levels are needed to further investigate TMM states in healthy human tissues with higher resolution. In this study, gene expression datasets from healthy human tissues from the GTEx portal were used to assess TMM activity states. We found high ALT and TEL pathway activities, first of all, in testis. Interestingly, TEL was activated only in a subset of testicular tissues paralleled with marked *TERC* expression, while depleted *TERC* levels lead to low TEL phenotypes. Indications for analogous binary mosaicism effects were reported in previous studies. In humans, *TERC* is mainly expressed in the primary spermatocytes, however at a lower level than in other spermatogenic cells (Paradis et al., 1999). In addition, *TERT* expression also shows mosaic-like regulation in testis, depending on the cells of the tissue or on the stage of spermatogenesis (Ozturk, 2015). Another study has shown that telomere length increases during the development of male germ cells from spermatogonia to spermatozoa, inversely correlated with enzymatic activity of telomerase (Achi et al., 2000). This could provide the link between the observed TEL mosaicism, and activation of the ALT pathway in testis.

The high ALT activity observed in testis was largely conditioned by upregulation of *RAD51*, suggesting that the ALT pathway is activated in a RAD51-dependent, rather than independent manner in testis (Jia-Min Zhang et al., 2019). Interestingly, the expression of the testis-specific Y-encoded-like protein 5 (*TSPYL5*), a recently identified APB body component that is crucial for survival of ALT$^+$ cells (Episkopou et al., 2019) was associated with ALT activity in healthy human testis in our study. Finally, our results indicate elevated TMM activity in testis in agreement with the recent study, also performed on GTEx datasets. Accordingly, telomeres are the longest in healthy human testis compared with other tissues and are in weak negative correlation with age (Demanelis et al., 2020).

There are several limitations to this study that we would like to mention. The accuracy of pathway activity measurements depends on the accuracy of expression values of TMM genes. In the absence of protein abundance measures, one can also include gene-specific translation regulators, such as miRNAs (Santambrogio et al., 2014; Salamati et al., 2020), alternative splicing or functional variants (Ludlow et al., 2019), and the lack of measurements for TERRA abundance (Novakovic et al., 2016).

Another limitation of our study is the lack of other assays as alternative measures for TMM phenotype. In particular a dataset coupled with C-circle assay based measurement of ALT could serve as an additional validation and to estimate generalizability of the curated ALT pathway. However, as the presence of APB was not always indicative of ALT pathway activity in our study, and as we had curated the ALT pathway genes from studies that were originally based on both APB and C-circle based assays (see **Supplementary Table S4**), we believe that our approach should not have systematic biases toward only capturing the presence of APBs. The latter is evident, since we do have samples in datasets showing high scores at the node describing for telomere recruitment to APBs, for which we have computed low ALT pathway activities (A7 in **Figure 2D** and A6,A9 in **Supplementary Figure S3**) and *vice versa* (A2 and A3 in **Supplementary Figure S3**).

In summary, we have reconstructed the TEL and ALT TMM pathways from previous literature. It has been carefully curated relying on reported molecular ingredients contributing to TMM and their interactions. Pathway signal flow activity estimates obtained from gene expression data have been shown to reliably estimate the TMM phenotype as a novel complementary approach to experimental assays. The main advantage of our approach is its "white box" (in contrast to "black box") nature, meaning that the resulting TMM phenotype can be "dissected" at gene level. In other words, we can explore gene-specific effects on sub-processes or events triggering activation of TMM. The method estimates TEL and ALT pathway activity in the same sample, thus enabling to establish its state in a TEL/ALT continuum with impact for investigating co-activation and switching events between the two pathways, e.g., upon cancer treatment and development. Of importance, owing to its sensitivity, it detects subtle activation of TEL and ALT in healthy human tissues. Despite the actuality of our pathway topologies it is important to note that new studies about additional factors affecting TMMs are permanently appearing with possible consequences for the pathways reconstructed here. Notably, the topologies were reconstructed in a generic manner: as the pathways can be activated via different mechanisms, not all the genes are required for the pathway activation in different situation. Future applications will show what branches and components are important in different situations. Single cell transcriptomics is one important field to essentially improve resolution of our TMM pathway method. Expression based methods thus potentially provide an independent and complementary approach to assess the TMM state. Because of the complex nature of TEL and ALT TMM, single gene expression markers or gene set signatures are potentially insufficient in most cases. Our TMM pathway method paves a new way for omics-based evaluation of TMM

with a series of applications ranging from cell experiments to cancer diagnostics.

MATERIALS AND METHODS

Literature Search and Pathway Construction

TEL and ALT pathways presented in this publication were carefully constructed based on comprehensive literature search resulting in 31 publications which provided relevant knowledge about TMM. To find and select publications describing factors involved in the telomere maintenance mechanisms, we searched the PubMed database with terms "telomerase" or "alternative lengthening of telomeres". The first phase was based on "review" articles published before 2020. These reviews provided the order of molecular processes involved in the pathways and thus their basic topology. The TMM genes mentioned in the review articles were chosen and then used in an extended search of the form (gene or protein) AND ("telomerase" OR "alternative lengthening of telomeres"). We also included genes not yet mentioned in review articles, by repeating the initial search with "telomerase" or "alternative lengthening of telomeres", concentrating on the original research articles of the last 3 years (2017–2020).

The articles were read in chronological order and in case of current consensus about the functional role of the mentioned genes they were included in the respective pathway, otherwise they were ignored. To ensure quality, we confirmed each interaction by inter-researcher agreement between the two authors of this manuscript that have curated the pathways. The present version of ALT and TEL pathways (TMMv2.0) is based on a previous version (TMMv1.0) (Nersisyan, 2017; Nersisyan et al., 2019), and makes use of an improved topology, branch and interaction structure according to updated literature knowledge. The network in.xgmml format may be accessed from **Supplementary Data 4**.

Data Sources and Preprocessing

For approval of ALT and TEL TMM pathways we made use of gene expression data on cell lines and liposarcoma tissues taken from the study by Lafferty-Whyte et al. (2009) which were annotated as double negative ALT^-/TEL^- (ALT and TEL inactive) or single positive ALT^+/TEL^-(ALT active) or ALT^-/TEL^+ (TEL active) using independent experiments. The gene expression matrix files were downloaded from the Gene Expression Omnibus (GEO) repository (accession GSE14533). This dataset contains microarray gene expression profiling data for ten cell lines cultured from different tissues, and seventeen liposarcoma tumor samples along with four human Mesenchymal Stem Cells (hMSC) samples isolated from the bone marrow of healthy individuals. ALT activation was assessed by the presence of ALT-associated promyelocytic leukemia bodies (APBs) (Costa et al., 2006; Cairney et al., 2008b), while TEL activation was measured using the telomeric-repeat amplification protocol for telomerase activity detection (TRAP-assay) (Kim et al., 1994; Cairney et al., 2008a). Among the cell lines, four were ALT positive (ALT^+/TEL^-), four were telomerase positive (ALT^-/TEL^+), and two were ALT/TEL double inactive (ALT^-/TEL^-). Among the liposarcoma samples, nine were ALT positive and eight were telomerase positive. Most of the samples had two technical replicates. In case of multiple microarray probes mapping to the same gene, the probe with highest standard deviation of values was considered. Cell line and tissue data were processed separately. Gene expression values higher than the 0.9 percentile in each of these sets were limited to that percentile value.

Healthy human tissue RNA-seq gene expression data was obtained from the GTEx portal (release V8) in units of transcripts per million (TPM). Fold changes were computed in comparison to the average of non-zero TPM values per gene. Data of the top 15 most common tissues extracted from subjects suffered violent or fast death from natural causes were selected. They were grouped by age and sex. For cross-tissue analysis we selected a maximum number of 10 subjects per age and sex group). For age-dependent analysis of the testis transcriptome all 129 available testis samples were chosen.

Pathway Signal Flow (PSF) Activity and Partial Influence (PI)

The pathway signal flow (PSF) algorithm (Arakelyan and Nersisyan, 2013; Nersisyan et al., 2014, 2017) was used to asses TMM pathway activity. It computes the activation along the whole pathway based on relative expression values of its member genes and of their interactions. Details are described in the supplement (section Pathway Signal Flow algorithm and **Supplementary Figure S1**). The PSF algorithm is implemented in the Cytoscape app PSFC (v1.1.8) (Nersisyan et al., 2017). For the specific tasks applied in this work, we have implemented a higher-level app, TMM (v0.8)[2]. It compares TMM pathway activation patterns with experimental annotations, uses PSFC for pathway activity computation and it also produces reports for TMM phenotype comparison across samples. The app is written in Java (major version 8), the source code is available at https://github.com/lilit-nersisyan/tmm. The app user guide, along with the example datasets and network files can be accessed at the project homepage http://big.sci.am/software/tmm/.

The partial influence (PI) of a source gene estimates the extent to which its expression affects the activity of a downstream target node in the pathway. The PI-value depends on the expression value of the source gene, pathway topology and the expression of other pathway members. It is computed by neutralizing the fold change of the gene to $FC = 1$, and calculating the log ratio of PSF at the target node before and after neutralizing its expression (**Supplementary Figure S4**). To compute the mean PI across all the samples in the testis, we have generated a mean sample by averaging fold change values for each gene, and performed PI analyses on it.

[2]http://apps.cytoscape.org/apps/tmm

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE14533 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14533); GTeX portal RNA-seq data (https://storage.googleapis.com/gtex_analysis_v8/rna_seq_data/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz).

AUTHOR CONTRIBUTIONS

LN, AA, and HB conceived the study. LN and AS performed pathway curation and data analysis. LN designed the software packages and prepared the figures. All authors contributed to data interpretation and manuscript preparation.

ACKNOWLEDGMENTS

We acknowledge support from the German Research Foundation (DFG) and Universität Leipzig within the program of Open Access Publishing. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 09/07/2020.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.662464/full#supplementary-material

REFERENCES

Achi, M. V., Ravindranath, N., and Dym, M. (2000). Telomere Length in Male Germ Cells Is Inversely Correlated with Telomerase Activity1. *Biol. Reprod.* 63, 591–598. doi: 10.1095/biolreprod63.2.591

Alhendi, A. S. N., and Royle, N. J. (2020). The absence of (TCAGGG)n repeats in some telomeres, combined with variable responses to NR2F2 depletion, suggest that this nuclear receptor plays an indirect role in the alternative lengthening of telomeres. *Sci. Rep.* 10:20597. doi: 10.1038/s41598-020-77606-w

Aragón, L. (2018). The Smc5/6 Complex: New and Old Functions of the Enigmatic Long-Distance Relative. *Annu. Rev. Genet.* 52, 89–107. doi: 10.1146/annurev-genet-120417-031353

Arakelyan, A., and Nersisyan, L. (2013). KEGGParser: Parsing and editing KEGG pathway maps in Matlab. *Bioinformatics* 29, 518–519.

Bechter, O. E., Zou, Y., Walker, W., Wright, W. E., and Shay, J. W. (2004). Telomeric recombination in mismatch repair deficient human colon cancer cells after telomerase inhibition. *Cancer Res.* 64, 3444–3451. doi: 10.1158/0008-5472.CAN-04-0323

Blackburn, E. H. (1991). Structure and function of telomeres. *Nature* 350, 569–573. doi: 10.1038/350569a0

Boyraz, B., Moon, D. H., Segal, M., Muosieyiri, M. Z., Aykanat, A., Tai, A. K., et al. (2016). Posttranscriptional manipulation of TERC reverses molecular hallmarks of telomere disease. *J. Clin. Invest.* 126, 3377–3382. doi: 10.1172/JCI87547

Braun, D. M., Chung, I., Kepper, N., Deeg, K. I., and Rippe, K. (2018). TelNet - a database for human and yeast genes involved in telomere maintenance. *BMC Genet.* 19:32. doi: 10.1186/s12863-018-0617-8

Cairney, C. J., Hoare, S. F., Daidone, M.-G., Zaffaroni, N., and Keith, W. N. (2008a). High level of telomerase RNA gene expression is associated with chromatin modification, the ALT phenotype and poor prognosis in liposarcoma. *Br. J. Cancer* 98, 1467–1474. doi: 10.1038/sj.bjc.6604328

Cairney, C. J., Hoare, S. F., Daidone, M. G., Zaffaroni, N., and Keith, W. N. (2008b). High level of telomerase RNA gene expression is associated with chromatin modification, the ALT phenotype and poor prognosis in liposarcoma. *Br. J. Cancer* 98, 1467–1474. doi: 10.1038/sj.bjc.6604328

Cayuela, M. L., Flores, J. M., and Blasco, M. A. (2005). The telomerase RNA component Terc is required for the tumour-promoting effects of Tert overexpression. *EMBO Rep.* 6, 268–274. doi: 10.1038/sj.embor.7400359

Cesare, A. J., and Reddel, R. R. (2008). Telomere uncapping and alternative lengthening of telomeres. *Mech. Ageing Dev.* 129, 99–108. doi: 10.1016/j.mad.2007.11.006

Chen, L.-Y., Redon, S., and Lingner, J. (2012). The human CST complex is a terminator of telomerase activity. *Nature* 488, 540–544. doi: 10.1038/nature11269

Chen, L., Roake, C. M., Freund, A., Batista, P. J., Tian, S., Yin, Y. A., et al. (2018). An Activity Switch in Human Telomerase Based on RNA Conformation and Shaped by TCAB1. *Cell* 174, 218.e–230.e. doi: 10.1016/j.cell.2018.04.039

Chen, X., Tang, W., Shi, J. B., Liu, M. M., and Liu, X. (2020). Therapeutic strategies for targeting telomerase in cancer. *Med. Res. Rev.* 40, 532–585. doi: 10.1002/med.21626

Cho, N. W., Dilley, R. L., Lampson, M. A., and Greenberg, R. A. (2014). Interchromosomal homology searches drive directional ALT telomere movement and synapsis. *Cell* 159, 108–121. doi: 10.1016/j.cell.2014.08.030

Chow, T. T., Zhao, Y., Mak, S. S., Shay, J. W., and Wright, W. E. (2012). Early and late steps in telomere overhang processing in normal human cells: The position of the final RNA primer drives telomere shortening. *Genes Dev.* 26, 1167–1178. doi: 10.1101/gad.187211.112

Chung, I., Osterwald, S., Deeg, K. I., and Rippe, K. (2012). PML body meets telomere: the beginning of an ALTernate ending? *Nucleus* 3, 263–275. doi: 10.4161/nucl.20326

Clynes, D., Jelinska, C., Xella, B., Ayyub, H., Scott, C., Mitson, M., et al. (2015). Suppression of the alternative lengthening of telomere pathway by the chromatin remodelling factor ATRX. *Nat. Commun.* 6:7538. doi: 10.1038/ncomms8538

Cohen, S. B., Graham, M. E., Lovrecz, G. O., Bache, N., Robinson, P. J., and Reddel, R. R. (2007). Protein composition of catalytically active human telomerase from immortal cells. *Science* 315, 1850–1853. doi: 10.1126/science.1138596

Conomos, D., Reddel, R. R., and Pickett, H. A. (2014). NuRD-ZNF827 recruitment to telomeres creates a molecular scaffold for homologous recombination. *Nat. Struct. Mol. Biol.* 21, 760–770. doi: 10.1038/nsmb.2877

Costa, A., Daidone, M. G., Daprai, L., Villa, R., Cantù, S., Pilotti, S., et al. (2006). Telomere maintenance mechanisms in liposarcomas: Association with histologic subtypes and disease progression. *Cancer Res.* 66, 8918–8924. doi: 10.1158/0008-5472.CAN-06-0273

Dagg, R. A., Pickett, H. A., Neumann, A. A., Napier, C. E., Henson, J. D., Teber, E. T., et al. (2017). Extensive Proliferation of Human Cancer Cells with Ever-Shorter Telomeres. *Cell Rep.* 19, 2544–2556. doi: 10.1016/j.celrep.2017.05.087

Deeg, K. I., Chung, I., Bauer, C., and Rippe, K. (2016). Cancer Cells with Alternative Lengthening of Telomeres Do Not Display a General Hypersensitivity to ATR Inhibition. *Front. Oncol.* 6:186. doi: 10.3389/fonc.2016.00186

Demanelis, K., Jasmine, F., Chen, L. S., Chernoff, M., Tong, L., Delgado, D., et al. (2020). Determinants of telomere length across human tissues. *Science* 369:AAZ6876. doi: 10.1126/SCIENCE.AAZ6876

Dilley, R. L., Verma, P., Cho, N. W., Winters, H. D., Wondisford, A. R., and Greenberg, R. A. (2016). Break-induced telomere synthesis underlies alternative telomere maintenance. *Nature* 539, 54–58. doi: 10.1038/nature20099

Dimitrova, N., and de Lange, T. (2009). Cell Cycle-Dependent Role of MRN at Dysfunctional Telomeres: ATM Signaling-Dependent

Induction of Nonhomologous End Joining (NHEJ) in G1 and Resection-Mediated Inhibition of NHEJ in G2. *Mol. Cell. Biol.* 29, 5552–5563. doi: 10.1128/mcb.00476-09

Dyer, M. A., Qadeer, Z. A., Valle-Garcia, D., and Bernstein, E. (2017). ATRX and DAXX: Mechanisms and Mutations. *Cold Spring Harb. Perspect. Med.* 7:a026567. doi: 10.1101/cshperspect.a026567

Episkopou, H., Diman, A., Claude, E., Viceconte, N., and Decottignies, A. (2019). TSPYL5 Depletion Induces Specific Death of ALT Cells through USP7-Dependent Proteasomal Degradation of POT1. *Mol. Cell* 75, 469.e–482.e. doi: 10.1016/j.molcel.2019.05.027

Fajkus, J. (2006). Detection of telomerase activity by the TRAP assay and its variants and alternatives. *Clin. Chim. Acta* 371, 25–31. doi: 10.1016/j.cca.2006.02.039

Flynn, R. L., Chang, S., and Zou, L. (2012). RPA and POT1: friends or foes at telomeres? *Cell Cycle* 11, 652–657. doi: 10.4161/cc.11.4.19061

Flynn, R. L., Cox, K. E., Jeitany, M., Wakimoto, H., Bryll, A. R., Ganem, N. J., et al. (2015). Alternative lengthening of telomeres renders cancer cells hypersensitive to ATR inhibitors. *Science* 347, 273–277. doi: 10.1126/science.1257216

Frohnert, C., Hutten, S., Wälde, S., Nath, A., and Kehlenbach, R. H. (2014). Importin 7 and Nup358 promote nuclear import of the protein component of human telomerase. *PLoS One* 9:0088887. doi: 10.1371/journal.pone.0088887

Gan, Y., Mo, Y., Johnston, J., Lu, J., Wientjes, M. G., and Au, J. L.-S. (2002). Telomere maintenance in telomerase-positive human ovarian SKOV-3 cells cannot be retarded by complete inhibition of telomerase. *FEBS Lett.* 527, 10–14. doi: 10.1016/S0014-5793(02)03141-1

Gocha, A. R. S., Nuovo, G., Iwenofu, O. H., and Groden, J. (2013). Human sarcomas are mosaic for telomerase-dependent and telomerase- independent telomere maintenance mechanisms: Implications for telomere-based therapies. *Am. J. Pathol.* 182, 41–48. doi: 10.1016/j.ajpath.2012.10.001

Greenberg, R. A. (2005). Telomeres, crisis and cancer. *Curr. Mol. Med.* 5, 213–218. doi: 10.2174/1566524053586590

Henson, J. D., Neumann, A. A., Yeager, T. R., and Reddel, R. R. (2002). Alternative lengthening of telomeres in mammalian cells. *Oncogene* 21, 598–610. doi: 10.1038/sj.onc.1205058

Henson, J. D., and Reddel, R. R. (2010). Assaying and investigating Alternative Lengthening of Telomeres activity in human cells and cancers. *FEBS Lett.* 584, 3800–3811. doi: 10.1016/j.febslet.2010.06.009

Hug, N., and Lingner, J. (2006). Telomere length homeostasis. *Chromosoma* 115, 413–425. doi: 10.1007/s00412-006-0067-3

Jeong, S. A., Kim, K., Lee, J. H., Cha, J. S., Khadka, P., Cho, H. S., et al. (2015). Akt-mediated phosphorylation increases the binding affinity of hTERT for importin α to promote nuclear translocation. *J. Cell Sci.* 2015:jcs.166132. doi: 10.1242/jcs.166132

Jia-Min Zhang, A., Yadav, T., Ouyang, J., Lan, L., and Zou Correspondence, L. (2019). Alternative Lengthening of Telomeres through Two Distinct Break-Induced Replication Pathways. *CellReports* 26, 955.e–968.e. doi: 10.1016/j.celrep.2018.12.102

Jiang, W.-Q., Zhong, Z.-H., Henson, J. D., Neumann, A. A., Chang, A. C.-M., and Reddel, R. R. (2005). Suppression of Alternative Lengthening of Telomeres by Sp100-Mediated Sequestration of the MRE11/RAD50/NBS1 Complex. *Mol. Cell. Biol.* 25, 2708–2721. doi: 10.1128/MCB.25.7.2708-2721.2005

Kim, N. W., Piatyszek, M. A., Prowse, K. R., Harley, C. B., West, M. D., Ho, P. L. C., et al. (1994). Specific association of human telomerase activity with immortal cells and cancer. *Science* 266, 2011–2015. doi: 10.1126/science.7605428

Lafferty-Whyte, K., Cairney, C. J., Will, M. B., Serakinci, N., Daidone, M.-G., Zaffaroni, N., et al. (2009). A gene expression signature classifying telomerase and ALT immortalization reveals an hTERT regulatory network and suggests a mesenchymal stem cell origin for ALT. *Oncogene* 28, 3765–3774. doi: 10.1038/onc.2009.238

Lafrance-Vanasse, J., Williams, G. J., and Tainer, J. A. (2015). Envisioning the dynamics and flexibility of Mre11-Rad50-Nbs1 complex to decipher its roles in DNA replication and repair. *Prog. Biophys. Mol. Biol.* 117, 182–193. doi: 10.1016/j.pbiomolbio.2014.12.004

Lee, J. H., Khadka, P., Baek, S. H., and Chung, I. K. (2010). CHIP promotes hTERT degradation and negatively regulates telomerase activity. *J. Biol. Chem.* 285, 42033–42045. doi: 10.1074/jbc.M110.149831

Liu, L., Bailey, S. M., Okuka, M., Muñoz, P., Li, C., Zhou, L., et al. (2007). Telomere lengthening early in development. *Nat. Cell Biol.* 9, 1436–1441. doi: 10.1038/ncb1664

Lovejoy, C. A., Li, W., Reisenweber, S., Thongthip, S., Bruno, J., de Lange, T., et al. (2012). Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of telomeres pathway. *PLoS Genet.* 8:e1002772. doi: 10.1371/journal.pgen.1002772

Ludlow, A. T., Slusher, A. L., and Sayed, M. E. (2019). Insights into telomerase/hTERT alternative splicing regulation using bioinformatics and network analysis in cancer. *Cancers* 11, 1–15. doi: 10.3390/cancers11050666

Martínez, P., and Blasco, M. A. (2015). Replicating through telomeres: a means to an end. *Trends Biochem. Sci.* 40, 504–515. doi: 10.1016/j.tibs.2015.06.003

Min, J., Wright, W. E., and Shay, J. W. (2019). Clustered telomeres in phase-separated nuclear condensates engage mitotic DNA synthesis through BLM and RAD52. *Genes Dev.* 33, 814–827. doi: 10.1101/gad.324905.119

Moon, D. H., Segal, M., Boyraz, B., Guinan, E., Hofmann, I., Cahan, P., et al. (2015). Poly(A)-specific ribonuclease (PARN) mediates 3′-end maturation of the telomerase RNA component. *Nat. Genet.* 47, 1482–1488. doi: 10.1038/ng.3423

Nersisyan, L. (2017). *Telomere Analysis Based on High-Throughput Multi -Omics Data*. Available online at: https://nbn-resolving.org/urn:nbn:de:bsz:15-qucosa2-162974 (accessed March 24, 2021).

Nersisyan, L., Hopp, L., Loeffler-Wirth, H., Galle, J., Loeffler, M., Arakelyan, A., et al. (2019). Telomere Length Maintenance and Its Transcriptional Regulation in Lynch Syndrome and Sporadic Colorectal Carcinoma. *Front. Oncol.* 9:1172. doi: 10.3389/fonc.2019.01172

Nersisyan, L., Johnson, G., Riel-Mehan, M., Pico, A. R., and Arakelyan, A. (2017). PSFC: a Pathway Signal Flow Calculator App for Cytoscape [version 2; peer review: 2 approved]. *F1000Research* 4:6706.2. doi: 10.12688/f1000research.6706.2

Nersisyan, L., Löffler-Wirth, H., Arakelyan, A., and Binder, H. (2014). Gene Set- and Pathway- Centered Knowledge Discovery Assigns Transcriptional Activation Patterns in Brain, Blood, and Colon Cancer: A Bioinformatics Perspective. *Int. J. Knowl. Discov. Bioinforma.* 4:24. doi: 10.4018/IJKDB.2014070104

Neumann, A. A., Watson, C. M., Noble, J. R., Pickett, H. A., Tam, P. P. L., and Reddel, R. R. (2013). Alternative lengthening of telomeres in normal mammalian somatic cells. *Genes Dev.* 27, 18–23. doi: 10.1101/gad.205062.112

Novakovic, B., Napier, C. E., Vryer, R., Dimitriadis, E., Manuelpillai, U., Sharkey, A., et al. (2016). DNA methylation mediated up-regulation of *TERRA* non-coding RNA is coincident with elongated telomeres in the human placenta. *Mol. Hum. Reprod.* 22, 791–799. doi: 10.1093/molehr/gaw053

Ozturk, S. (2015). Telomerase activity and telomere length in male germ cells. *Biol. Reprod.* 92, 53–54. doi: 10.1095/biolreprod.114.124008

Paradis, V., Dargère, D., Laurendeau, I., Benoît, G., Vidaud, M., Jardin, A., et al. (1999). Expression of the RNA component of human telomerase (hTR) in prostate cancer, prostatic intraepithelial neoplasia, and normal prostate tissue. *J. Pathol.* 189, 213–218. doi: 10.1002/(SICI)1096-9896(199910)189:2<213::AID-PATH417<3.0.CO;2-A

Pickett, H. A., and Reddel, R. R. (2015). Molecular mechanisms of activity and derepression of alternative lengthening of telomeres. *Nat. Struct. Mol. Biol.* 22, 875–880. doi: 10.1038/nsmb.3106

Potts, P. R., and Yu, H. (2007). The SMC5/6 complex maintains telomere length in ALT cancer cells through SUMOylation of telomere-binding proteins. *Nat. Struct. Mol. Biol.* 14, 581–590. doi: 10.1038/nsmb1259

Recagni, M., Bidzinska, J., Zaffaroni, N., and Folini, M. (2020). The Role of Alternative Lengthening of Telomeres Mechanism in Cancer: Translational and Therapeutic Implications. *Cancers* 12:949. doi: 10.3390/cancers12040949

Rice, C., and Skordalakes, E. (2016). Structure and function of the telomeric CST complex. *Comput. Struct. Biotechnol. J.* 14, 161–167. doi: 10.1016/j.csbj.2016.04.002

Salamati, A., Majidinia, M., Asemi, Z., Sadeghpour, A., Oskoii, M. A., Shanebandi, D., et al. (2020). Modulation of telomerase expression and function by miRNAs: Anti-cancer potential. *Life Sci.* 259:118387. doi: 10.1016/j.lfs.2020.118387

Santambrogio, F., Gandellini, P., Cimino-Reale, G., Zaffaroni, N., and Folini, M. (2014). MicroRNA-dependent Regulation of Telomere Maintenance Mechanisms: A Field as Much Unexplored as Potentially Promising. *Curr Pharm Des.* 20, 6404–6421.

Schmidt, J. C., and Cech, T. R. (2015). Human telomerase: Biogenesis, trafficking, recruitment, and activation. *Genes Dev.* 29, 1095–1105. doi: 10.1101/gad.263863.115

Shay, J. W. (2016). Role of Telomeres and Telomerase in Aging and Cancer. *Cancer Discov.* 6, 584–593. doi: 10.1158/2159-8290.CD-16-0062

Shay, J. W., Reddel, R. R., and Wright, W. E. (2012). Cancer: Cancer and telomeres - An alternative to telomerase. *Science* 2012:1222394. doi: 10.1126/science.1222394

Sobinoff, A. P., Allen, J. A., Neumann, A. A., Yang, S. F., Walsh, M. E., Henson, J. D., et al. (2017). BLM and SLX4 play opposing roles in recombination−dependent replication at human telomeres. *EMBO J.* 36, 2907–2919. doi: 10.15252/embj.201796889

Sobinoff, A. P., and Pickett, H. A. (2017). Alternative Lengthening of Telomeres: DNA Repair Pathways Converge. *Trends Genet.* 33, 921–932. doi: 10.1016/J.TIG.2017.09.003

Sobinoff, A. P., and Pickett, H. A. (2020). Mechanisms that drive telomere maintenance and recombination in human cancers. *Curr. Opin. Genet. Dev.* 60, 25–30. doi: 10.1016/j.gde.2020.02.006

Sugarman, E. T., Zhang, G., and Shay, J. W. (2019). In perspective: An update on telomere targeting in cancer. *Mol. Carcinog.* 58, 1581–1588. doi: 10.1002/mc.23035

Tseng, C.-K., Wang, H.-F., Burns, A. M., Schroeder, M. R., Gaspari, M., and Baumann, P. (2015). Human Telomerase RNA Processing and Quality Control. *Cell Rep.* 13, 2232–2243. doi: 10.1016/j.celrep.2015.10.075

Venteicher, A. S., Meng, Z., Mason, P. J., Veenstra, T. D., and Artandi, S. E. (2008). Identification of ATPases Pontin and Reptin as Telomerase Components Essential for Holoenzyme Assembly. *Cell* 132, 945–957. doi: 10.1016/j.cell.2008.01.019

Verma, P., Dilley, R. L., Gyparaki, M. T., and Greenberg, R. A. (2018). "Direct Quantitative Monitoring of Homology-Directed DNA Repair of Damaged Telomeres," in *Methods in Enzymology*,eds J. Abelson, M. Simon, G. Verdine, A. Pyle (Amsterdam: Elsevier)doi: 10.1016/bs.mie.2017.11.010

Villa, R., Daidone, M. G., Motta, R., Venturini, L., De Marco, C., Vannelli, A., et al. (2008). Multiple mechanisms of telomere maintenance exist and differentially affect clinical outcome in diffuse malignant peritoneal mesothelioma. *Clin. Cancer Res.* 14, 4134–4140. doi: 10.1158/1078-0432.CCR-08-0099

Yuan, X., Larsson, C., and Xu, D. (2019). Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: old actors and new players. *Oncogene* 38, 6172–6183. doi: 10.1038/s41388-019-0872-9

Zhang, J.-M., Genois, M.-M., Ouyang, J., Lan, L., and Zou, L. (2021). Alternative lengthening of telomeres is a self-perpetuating process in ALT-associated PML bodies. *Mol. Cell* 2021:030. doi: 10.1016/j.molcel.2020.12.030

Zhu, X. D., Niedernhofer, L., Kuster, B., Mann, M., Hoeijmakers, J. H. J., and De Lange, T. (2003). ERCC1/XPF Removes the 3' Overhang from Uncapped Telomeres and Represses Formation of Telomeric DNA-Containing Double Minute Chromosomes. *Mol. Cell* 2003, 478–477. doi: 10.1016/S1097-2765(03)00478-7

Conflict of Interest: LN and AA were co-founders of a start-up team Pathverse, which uses the algorithms described herein for further developments.

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.

frontiers in Genetics

TECHNOLOGY AND CODE
published: 11 May 2021
doi: 10.3389/fgene.2021.618189

MPRAdecoder: Processing of the Raw MPRA Data With *a priori* Unknown Sequences of the Region of Interest and Associated Barcodes

Anna E. Letiagina[1,2†], Evgeniya S. Omelina[1†], Anton V. Ivankin[1] and Alexey V. Pindyurin[1]*

[1] *Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia,*
[2] *Faculty of Natural Sciences, Novosibirsk State University, Novosibirsk, Russia*

Massively parallel reporter assays (MPRAs) enable high-throughput functional evaluation of numerous DNA regulatory elements and/or their mutant variants. The assays are based on the construction of reporter plasmid libraries containing two variable parts, a region of interest (ROI) and a barcode (BC), located outside and within the transcription unit, respectively. Importantly, each plasmid molecule in a such a highly diverse library is characterized by a unique BC–ROI association. The reporter constructs are delivered to target cells and expression of BCs at the transcript level is assayed by RT-PCR followed by next-generation sequencing (NGS). The obtained values are normalized to the abundance of BCs in the plasmid DNA sample. Altogether, this allows evaluating the regulatory potential of the associated ROI sequences. However, depending on the MPRA library construction design, the BC and ROI sequences as well as their associations can be *a priori* unknown. In such a case, the BC and ROI sequences, their possible mutant variants, and unambiguous BC–ROI associations have to be identified, whereas all uncertain cases have to be excluded from the analysis. Besides the preparation of additional "mapping" samples for NGS, this also requires specific bioinformatics tools. Here, we present a pipeline for processing raw MPRA data obtained by NGS for reporter construct libraries with *a priori* unknown sequences of BCs and ROIs. The pipeline robustly identifies unambiguous (so-called genuine) BCs and ROIs associated with them, calculates the normalized expression level for each BC and the averaged values for each ROI, and provides a graphical visualization of the processed data.

Keywords: massively parallel reporter assay, MPRA, reporter constructs, region of interest, barcodes, next-generation sequencing, NGS data processing, pipeline

OPEN ACCESS

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

Reviewed by:
Nariman Battulin,
The Siberian Branch of the Russian Academy of Sciences, Russia
Ilias Georgakopoulos-Soares,
University of California, San Francisco, United States

****Correspondence:***
Alexey V. Pindyurin
a.pindyurin@mcb.nsc.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: *16 October 2020*
Accepted: *25 March 2021*
Published: *11 May 2021*

Citation:
Letiagina AE, Omelina ES, Ivankin AV and Pindyurin AV (2021) MPRAdecoder: Processing of the Raw MPRA Data With a priori Unknown Sequences of the Region of Interest and Associated Barcodes. Front. Genet. 12:618189. doi: 10.3389/fgene.2021.618189

INTRODUCTION

Although numerous regulatory elements have been identified in eukaryotic genomes (Narlikar and Ovcharenko, 2009; Taher et al., 2011; Kellis et al., 2014), so far there is no complete understanding of why these elements are active in specific cell types and at specific levels. Accordingly, the effect of a particular mutation within a regulatory element can be hardly predicted, especially for a particular cell type (1000 Genomes Project Consortium et al., 2015; Albert and Kruglyak, 2015;

Rojano et al., 2019). The recent development of massively parallel reporter assays (MPRAs) allows high-throughput functional characterization of native transcriptional regulatory elements (first of all, enhancers and promoters) as well as their mutant variants (reviewed in Haberle and Lenhard, 2012; Inoue and Ahituv, 2015; Trauernicht et al., 2020; Mulvey et al., 2021). In an MPRA, regions of interests (ROIs), e.g., putative enhancers or promoters, together with unique barcodes (BCs) are assembled into reporter constructs to obtain MPRA plasmid libraries that consist of thousands or even millions of individual molecules (Kheradpour et al., 2013; Kwasnieski et al., 2014; van Arensbergen et al., 2019). Specific MPRA libraries can also be packaged in lentiviruses to deliver reporter constructs into the target genome (O'Connell et al., 2016; Inoue et al., 2017; Maricque et al., 2017; Gordon et al., 2020).

From the structural point of view, BCs are always placed within the transcription unit [usually in the 5′ or 3′ untranslated region (UTR)], whereas ROIs are typically outside this unit (**Figure 1A**). As a result, the BC sequences are present in the reporter mRNA molecules and, thus, allow quantitative evaluation of the regulatory effects caused by their *cis*-paired ROI variants using next-generation sequencing (NGS) (**Figure 1B** and **Supplementary Figure 1**). For that, cells of interest are transfected by an MPRA plasmid library or transduced by a lentiviral MPRA library, and subsequently, transcriptional activity levels of barcoded reporters are assessed on episomal plasmids and/or after stable integration of the constructs at random or specific genomic loci (Melnikov et al., 2012; Sharon et al., 2012; Kheradpour et al., 2013; White et al., 2013; O'Connell et al., 2016; Tewhey et al., 2016; Ulirsch et al., 2016; Maricque et al., 2017; Inoue et al., 2019). More specifically, the "expression" and "normalization" samples are prepared by PCR amplification of the BC sequences from cDNA synthesized on total RNA isolated from the transfected/transduced cells and the plasmid DNA used to transfect cells or total DNA isolated from the transduced cells, respectively. These samples are subjected to NGS to determine the normalized expression level of each BC, which is calculated as the ratio between the BC abundance in the expression and normalization samples.

It should be noted that ROIs can be either (i) preselected native, mutant, and/or synthetic sequences (e.g., minimal core elements of enhancers and promoters) usually of the same length (Melnikov et al., 2012; Sharon et al., 2012; Kheradpour et al., 2013; Smith et al., 2013) or (ii) somehow experimentally enriched genomic fragments, random genomic fragments, or synthetic sequences of varying length (Mogno et al., 2013; Verfaillie et al., 2016; van Arensbergen et al., 2017). In particular cases, the ROI can be just a fixed segment within the cloned regulatory element (Patwardhan et al., 2009; Vvedenskaya et al., 2015; Omelina et al., 2019). On the other hand, BCs are most frequently sequences of fixed length between 9 and 20 nucleotides (nts) (Kwasnieski et al., 2012; Melnikov et al., 2012; Patwardhan et al., 2012; Mogno et al., 2013; Verfaillie et al., 2016).

Depending on the MPRA library design, the ROI and BC sequences as well as their associations can be either *a priori* known or not. Completely predetermined MPRA libraries are generated by using sequences synthesized on custom high-density DNA microarrays (Patwardhan et al., 2009; Melnikov et al., 2012; Sharon et al., 2012; Kwasnieski et al., 2014). MPRA libraries with unknown sequences of ROIs and BCs are made by cloning randomly sheared genomic fragments or pooled synthetic DNA fragments or by PCR-mediated mutagenesis and/or by cloning oligonucleotides containing randomized stretches of nucleotides (Patwardhan et al., 2012; Mogno et al., 2013; Vvedenskaya et al., 2015; Verfaillie et al., 2016; van Arensbergen et al., 2017; Kircher et al., 2019; Omelina et al., 2019). In some cases, the ROI sequences are predetermined although associated BCs are not known in advance (Smith et al., 2013; O'Connell et al., 2016; Tewhey et al., 2016; Grossman et al., 2017; Gordon et al., 2020). For the libraries that are not completely predetermined, there is a need to identify cloned ROI and/or BC sequences as well as their associations. Hereafter, the procedure of finding unique BC–ROI associations is referred to as "mapping" by analogy with the thousands of reporters integrated in parallel (TRIP) experiments (Akhtar et al., 2013, 2014). The mapping is typically done by PCR amplification of BC–ROI regions of MPRA constructs followed by Illumina NGS (Patwardhan et al., 2012; Mogno et al., 2013; Tewhey et al., 2016; Omelina et al., 2019). Importantly, associations of the same BC with different ROI sequences are excluded from the further analysis although the association of the same ROI with different BCs allows revealing and excluding the possible influence of particular BC sequences on the measurements.

The MPRAdecoder pipeline described in this study was developed for the processing of NGS data generated for MPRA libraries with *a priori* unknown sequences of ROIs and BCs, for example, those cloned by the usage of oligonucleotides with randomized stretches of nucleotides. The pipeline (i) robustly identifies unambiguous (hereafter genuine) BCs and their mutant variants as well as associated ROIs, (ii) calculates the normalized expression level for each genuine BC and the averaged values for each ROI, and (iii) provides a graphical visualization of the processed data. The functionality of the pipeline was demonstrated using a data set obtained for an MPRA library designed to study the effects of sequence variations located at a certain distance downstream of the transcription termination site (TTS) of the *eGFP* reporter on its expression at the transcription level.

MATERIALS AND METHODS

Preparation of the MPRA Mapping, Expression, and Normalization Samples and Illumina NGS

The MPRA plasmid library, in which random-sequence BC and ROI are separated by an 83-nt fixed-sequence region and located, respectively, in 3′ UTR and downstream of the TTS of the *eGFP* reporter, was generated earlier (Omelina et al., 2019). The wild-type and mutant deltaC (Boldyreva et al., 2021) reporter plasmids carrying specific 20-nt BCs were constructed by standard molecular cloning procedures and verified by

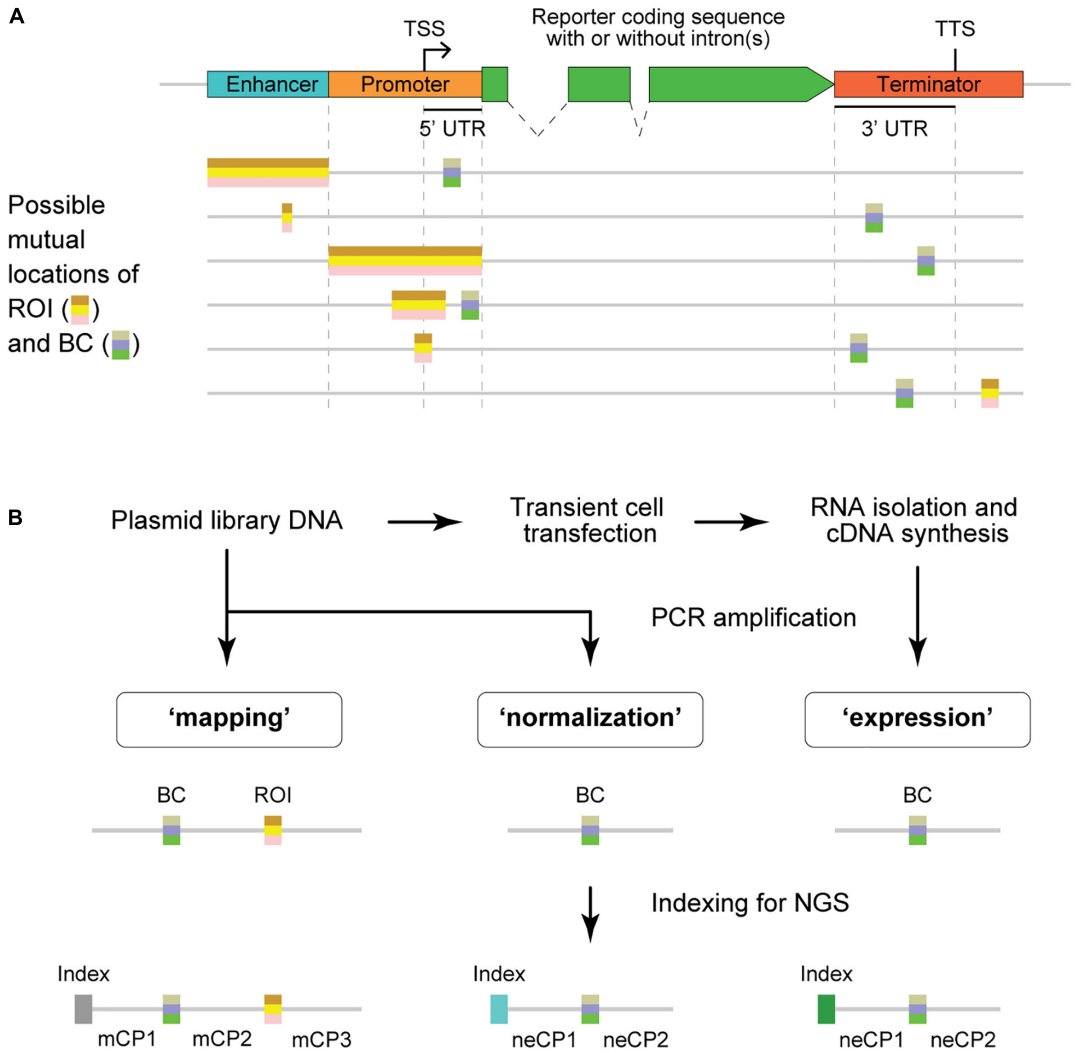


FIGURE 1 | Structure of MPRA constructs and samples. **(A)** Schematic representation of typical MPRA reporter constructs with two variable regions, the region of interest (ROI) and barcode (BC). An enhancer element may be missing in particular MPRA constructs. The ROI is usually located outside the coding sequence, whereas the BC must be within the transcribed unit to ensure measurements of the ROI influence on the reporter expression level. TSS, transcription start site; TTS, transcription termination site; UTR, untranslated region. **(B)** Schematic showing the experimental steps involved in the preparation of MPRA samples and their structures. Note that the normalization and expression samples have identical structures; the mapping sample is required only for MPRA libraries in which BC–ROI associations are not predetermined. mCP, constant part present in the sequence of the mapping sample; neCP, constant part present in the sequence of the normalization/expression sample.

sequencing. An equimolar pool of two such wild-type and two deltaC mutant plasmids was mixed in a 1:99 molar ratio with the MPRA plasmid library. Immortalized human embryonic kidney (HEK293T) cells were obtained from ATCC (United States) and were maintained and transfected as described previously (Boldyreva et al., 2021).

The mapping samples were prepared according to a previously reported two-round conventional PCR procedure that prevents the formation of chimeric products (Omelina et al., 2019). Briefly, primers specific to the ends of fixed sequences mCP1 and mCP3 (**Figure 1B** and **Table 1**) were used, and a specific, custom-designed 8-nt index along with other sequences necessary for Illumina NGS was introduced in the PCR products of each sample replicate. The normalization samples were obtained in the same way, using primers specific to the ends of fixed sequences neCP1 and neCP2 (**Figure 1B** and **Table 1**) and 2.5 ng of the plasmid library as a template. To prepare expression samples, BCs were amplified as specified above but using 1/20 of cDNA prepared from the transfected cells as described earlier (Boldyreva et al., 2021) as a template. Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific) was used for all amplification reactions. All obtained PCR products were purified on spin columns, mixed together, and sequenced on an Illumina MiSeq instrument as 151-nt single-end reads. Notice that the read length was shorter than the amplified plasmid fragments for all samples. Therefore, there was no need to remove Illumina

adapter sequences from the reads. Finally, to prepare an example data set, a representative subset of the reads was randomly selected from the obtained fastq file. A copy of this subset was demultiplexed using Cutadapt (Martin, 2011).

Pipeline Code and Documentation Availability

The MPRAdecoder pipeline source code written in Python, the example data set, and the corresponding expected outputs as well as detailed documentation are publicly available on GitHub repository[1].

Hardware and Software Requirements

The MPRAdecoder installation and analyses were performed on a computer with an Intel® Core™ i7-3770 processor, 31.4 Gb RAM, Linux Ubuntu 14.04 64-bit system, and Python version 3.8.6.

RESULTS

Overview of the MPRAdecoder Pipeline

A workflow of the MPRAdecoder pipeline is shown in **Figure 2**. Briefly, after providing details of a particular MPRA data set to be analyzed, the pipeline parses the input fastq file(s) and demultiplexes them if required. Next, all expected parts of the mapping, normalization, and expression reads are detected, particularly the sequences of BCs and ROIs. Then, a list of BCs common for all samples is generated with the assumption that some BCs have zero counts in the expression data. After that, genuine BCs and their mutant variants as well as associated ROIs are identified. Finally, the data are averaged over expression and normalization replicates, normalized, and averaged over ROIs, and the results are visualized in different plots. Below, these steps are described in more detail with the help of the example MPRA data set.

[1] https://github.com/Code-master2020/MPRAdecoder

Characteristics of the Example Data Set

To demonstrate the capabilities of the MPRAdecoder pipeline, we used a data set consisting of two biological replicates of mapping, normalization, and expression samples obtained using an MPRA library, in which the BC and ROI (both cloned by using oligonucleotides containing fully randomized sequences) are located in 3′ UTR and downstream of TTS, respectively (the option is shown at the bottom of **Figure 1A**), being separated by 83 nts of fixed sequence (Omelina et al., 2019). The samples were sequenced as 151-nt single-end reads on the Illumina MiSeq platform and were indexed with custom-designed 8-nt sequences located at the beginning of the reads (**Figure 1B**). Important features of the data set are listed in **Table 1**. Note that the BC sequences were in forward and reverse-complement orientations in the mapping and normalization/expression samples, respectively. In addition, about 1% of the reads in each sample contained four unique 20-nt BCs associated with spiked-in reference constructs; the TTCCAAGTGCAGGTTAGGCG and TGTGTACGGCTTGCTCTCAA sequences tagged the wild-type construct, whereas GAGCCCGGATCCACTCCAAG and TGTCACGTCAGCTAACCCAC sequences marked the deltaC mutant construct that is characterized by a higher expression level than the wild-type one (Boldyreva et al., 2021). The substantially longer length of the BC (18 nts) compared to the ROI (8 nts) ensures that each ROI is associated with multiple different BCs in a representative large plasmid library. This allows

TABLE 1 | Specific features of the example MPRA data set.

Part[a]	Length, nts	Strand[b]	Sequence	Note
"Mapping" sample				
index	8	Plus	AGCGAGCT, CTGCACGT	Fixed
mCP1	17	Plus	GACACTCGAGGATCGAG	Fixed
BC	18[c]	Plus	$(N)_{18}$	Random
mCP2	83	Plus	GAGTTGTGGCCGGCCCTTGTGACTGGGAAAACCCTGGCGTAAAT AAAATACGAAATGACTAGTCATGCGTCAATTTTACGCAT	Fixed
ROI	8	Plus	$(N)_8$	Random
mCP3	17[d]	Plus	TTAACGTACGTCACAATATGATTATCTTTCTAGGG[e]	Fixed
"Normalization" and "Expression" samples				
index	8	Plus	CCTATGGT, AACGTCGT, ACAATTCG, TACTTGTC	Fixed
neCP1	39	Minus	CGCCAGGGTTTTCCCAGTCACAAGGGCCGGCCACAACTC	Fixed
BC	18[c]	Minus	$(N)_{18}$	Random
neCP2	86[d]	Minus	CTCGATCCTCGAGTGTCACCTAAATCGTATGCGGCCG CGAATTCTTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCGGCGGC GGTCACGAACTCCAGCAGGAC[e]	Fixed

[a] *In the order of presence in the sample (starting from immediately after the Illumina forward sequencing primer).*
[b] *Orientation relative to the plasmid reporter construct, except the index introduced during PCR amplification.*
[c] *The length of the BCs present in the spiked-in reference constructs was 20 nts (see the text for details).*
[d] *The expected length of the fragment of the component in 151-nt single-end reads; the length is shorter by 2 nts for the reads containing reference 20-nt BCs.*
[e] *The complete sequence of the component in the PCR amplified sample is shown; the fragment expected in 151-nt single-end reads is underlined.*

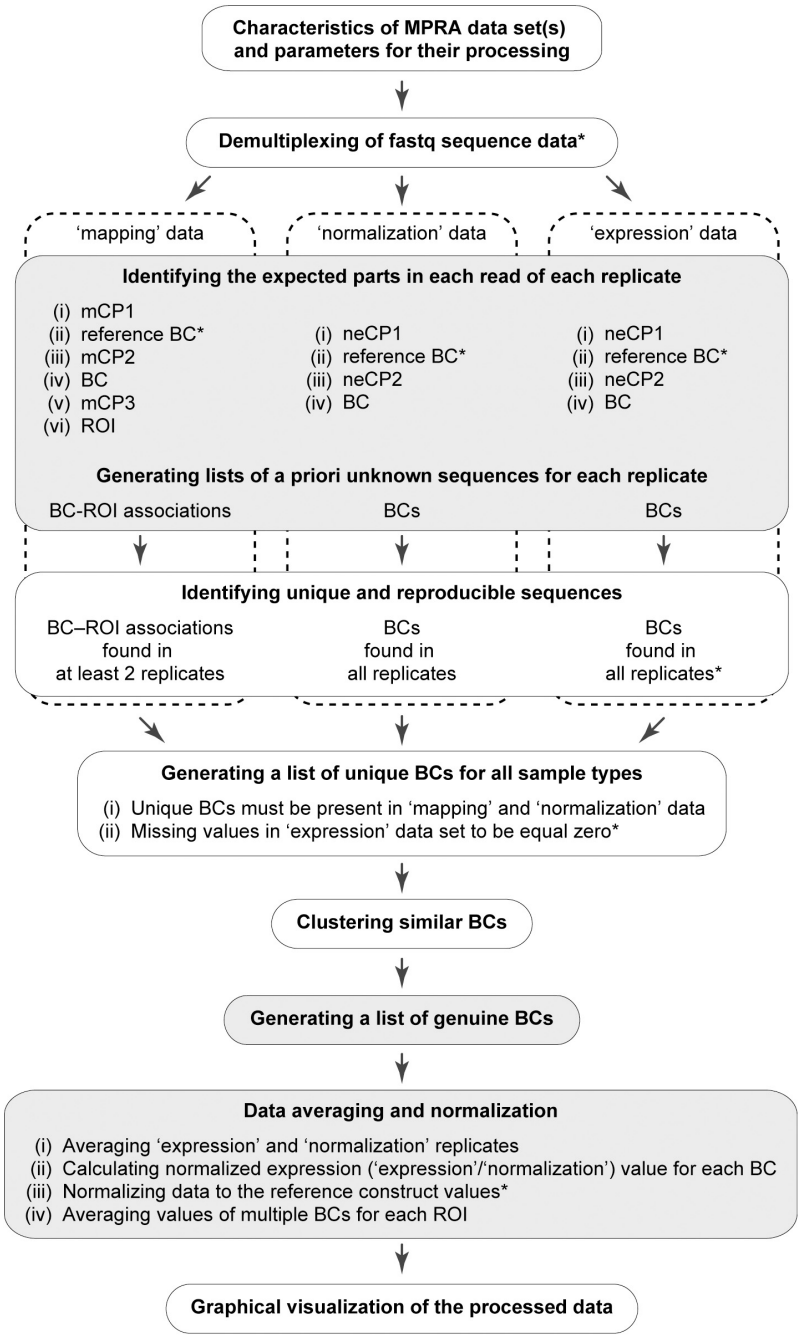


FIGURE 2 | Schematic overview of the MPRAdecoder pipeline. The optional steps are marked by an asterisk. The steps at which the pipeline saves intermediate or final results are marked by gray.

controlling the potential influence of individual BC sequences on the studied phenomenon.

Specifying Characteristics of an MPRA Data Set to Be Analyzed

The information on the input MPRA data set is provided in the two complementary forms. First, most details, such as (i) names and lengths of all expected parts in the mapping and normalization/expression reads for each MPRA library (including indexes), (ii) sequences of the predetermined parts (including indexes and optional reference BCs), (iii) relative orientation of BC sequences in mapping and normalization/expression reads, (iv) a maximum allowed error rate and the Phred quality score threshold for different parts, (v) a minimum number of read counts required for a BC

and a BC–ROI association, and (vi) settings for identification of genuine BCs and associated ROIs, are specified in the configuration file. A detailed description of this file is available on the GitHub page of this project. Second, a user has to manually input the following details in the command prompt: (vii) names of the appropriate fastq file(s) and their locations as well as a location for output files, (viii) a number of replicates of each sample for each MPRA library, (ix) names of indexes used for sample multiplexing and (x) information on whether the fastq file(s) should be demultiplexed by the pipeline.

MPRA Data Demultiplexing by Pairwise Sequence Alignment

The pipeline is able to process either fastq files that are already demultiplexed, for example, by the Illumina software, or fastq files containing custom-designed index sequences at the beginning of the reads. In the latter case, detection of a predetermined index sequence in each read is performed using a pairwise sequence alignment tool from Biopython (Cock et al., 2009). For that, all index sequences specified in the configuration file are aligned, one by one, against the beginning of a read. The following alignment scoring system is used: +1 for a match, 0 for a mismatch, and −1 for an indel. If the maximum alignment score is higher than or equal to the threshold value (calculated as the index length - the maximum allowed error rate + 1 for each insertion) and the Phred quality score for each base (Cock et al., 2010) is higher than a threshold (equal to 10 for the example data set), the corresponding index sequence is considered to be identified; otherwise, the read is discarded. To generate the example data set, 8-nt index sequences differing from each other by at least 2 nts were used as suggested for the short (5–10 nts) predefined BCs (Patwardhan et al., 2009; Sharon et al., 2012). At the same time, the maximum allowed error rate was set to ∼10% based on our experience with PCR-amplification and subsequent NGS of predetermined sequences under experimental conditions identical to those used in this study (including the quality of oligonucleotide primers). Together, these factors ensure that one allowed single-base mutation (substitution, deletion, or insertion) in the index sequence cannot lead to an error in its identification. At the end, the reads are divided into an appropriate number of groups based on the detected indexes.

Identification of the BC and ROI Sequences in the Reads

Detection of the mCP1, mCP2, mCP3, neCP1, neCP2 (**Figure 1B** and **Table 1**), and reference BC sequences in the reads is performed for each replicate of each sample by using the pairwise sequence alignment approach described above for the index, taking into account location(s) of the preceding part(s), which can be already identified (e.g., the mCP1/neCP1) or just estimated (e.g., the BC). Sequences of BCs and ROIs are defined as spacers between the appropriate constant parts. By default, the Phred quality scores are ignored for the mCP1, mCP2, mCP3, neCP1, and neCP2 sequences. For the BCs (including the reference ones) and ROIs, the quality score for each base should be higher than a threshold (e.g., set to 10 for the example data set); otherwise, reads are discarded from the downstream analysis. More specifically, in the case of the mapping reads, the process includes the following sequential steps. First, the mCP1 sequence is detected. Second, if sequences of the reference BCs are specified in the configuration file, the reads with such BCs are identified and excluded from the subsequent structural analysis. This is done because the functional sequences (e.g., wild-type or deltaC in the example data set) associated with the reference BCs might be located outside the ROI (e.g., within the mCP2 sequence as in the example data set). Third, the mCP2 sequence is detected, and the sequence between mCP1 and mCP2 is recognized as the BC if its length is within the range set in the configuration file (e.g., $\geq$16 and $\leq$20 nts for the example data set). Fourth, the mCP3 sequence is identified, and the sequence between mCP2 and mCP3 is recognized as the ROI if its length is within the range defined in the configuration file (e.g., $\geq$7 and $\leq$9 nts for the example data set). In the case of the normalization and expression reads, the last step is omitted. Lastly, if the ROI and/or BC sequences are in reverse-complement orientations in the mapping or normalization/expression samples (this is specified in the configuration file), they are converted to their forward counterparts.

Data Filtering and Generation of a List of Unique BCs

At the next step, the number of supporting reads for each BC (with a random or reference sequence) is counted for each replicate of all samples. Then, these numbers are divided by the total number of effective reads (i.e., those that passed all filters described above) in a replicate and multiplied by 1×10^6 to calculate the reads per million (RPM) values. After that, unique BC–ROI associations and BCs are assessed for reproducibility and robustness. Although preliminary results can be obtained using single replicates of the mapping, normalization, and expression samples, at least two replicates of each sample are strongly recommended. Under such conditions, only the BC–ROI associations that are revealed with at least m raw read counts (e.g., one for the example data set) in at least two out of any available number of replicates of the mapping data are retained for further analysis. Also, only the BCs with n raw read counts (e.g., three for the example data set) in each replicate of the normalization data are kept. For the expression data, the threshold read count e is set by default to zero, as some BCs might be present with very low frequency or even completely absent in the reporter transcripts due to the properties of particular ROI sequences. The threshold values m, n, and e are arbitrarily set in the configuration file. Finally, a list of BCs that are common for all samples is generated considering that some BCs might have zero counts in some or all replicates of the expression data.

Identification of Genuine BCs

Oligonucleotides with a totally randomized part (characterized by an equal representation of all four nucleotides at each position) of 15–20 nts in length can ensure cloning of $\sim 1 \times 10^9$ to 1×10^{12} unique BCs, some of which might be different from each other just at one position. However, in practice, the size of

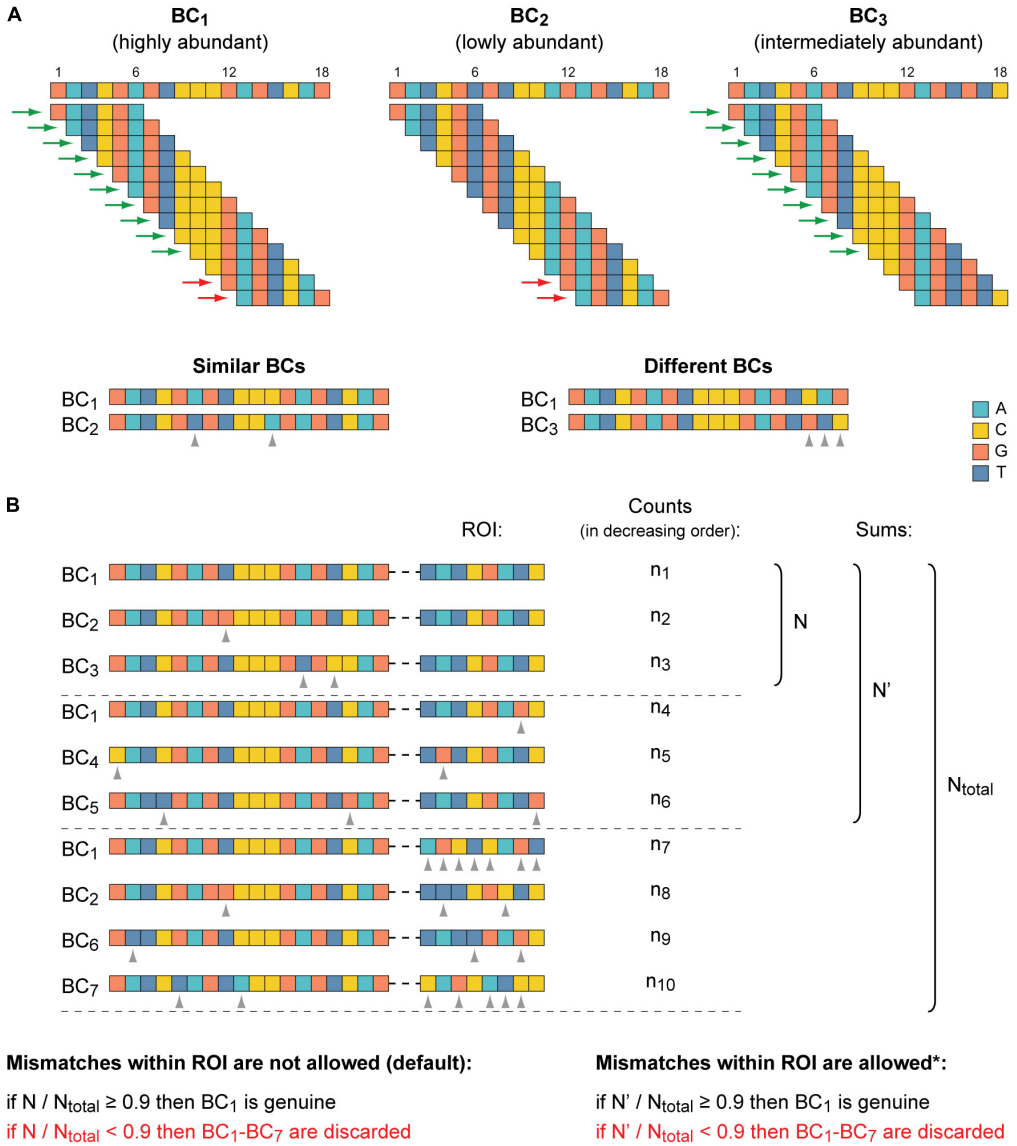


FIGURE 3 | Identification of genuine BCs, their mutant variants, and associated ROIs. **(A)** The clustering of similar BC sequences is achieved by their decomposition into overlapping k-mers and by the subsequent pairwise alignment of BCs that share identical k-mers. At the top, three BCs are shown as an example. K-mers (6-mers) shared by BC_1 and BC_2 and by BC_1 and BC_3 are indicated by red and green arrows, respectively. At the bottom, the pairwise sequence alignment of the candidate similar BC sequences is depicted. The BC_1 and BC_2 are recognized to be similar because their sequences differ from each other only at two positions ($\leq$the maximum allowed error rate). BC_1 and BC_3 are considered to be different because their sequences differ from each other at three positions ($>$the maximum allowed error rate) even though these BCs share more common k-mers than the BC_1 and BC_2. **(B)** Identification of genuine BCs. One cluster of seven similar BCs along with the associated ROI sequences is shown as an example. BC_1 is the most abundant BC (as in **A**) and the ROI sequence, which is associated with it most frequently ($n_1 > n_4$ and $n_1 > n_7$), is considered as the putative ROI for the cluster. By default, if the putative ROI is supported by at least 90% of normalized read counts calculated for all ROI sequences found in the cluster, the BC_1 becomes genuine. Otherwise, the entire cluster is excluded from the subsequent analysis. Optionally (indicated by an asterisk), if mismatches within the ROI are permitted (e.g., a difference at one position could be allowed for the example data set), then normalized read counts for the putative ROI and its allowed mutants are summed. Notice that differences between the ROI sequences associated with similar BCs should be allowed with caution, especially for very short ROIs. Dashed horizontal lines separate different groups of ROIs: the putative sequence, its allowed mutants, and all other sequences. Gray arrowheads denote mismatches in both panels.

a typical MPRA plasmid library is significantly less (by orders of magnitude) than the theoretical values. Nevertheless, in MPRA data sets, BCs with similar sequences do appear, partly due to errors introduced during PCR amplification and NGS steps. Thus, there is a need to find similar BC sequences, group them, and identify the genuine BCs in each such group (referred to below as a cluster). Two BC sequences are considered to be similar if they differ at no more than *s* positions (by substitutions, deletions, and/or insertions), where *s* is equal to the maximum allowed error rate for this part. By default, up to two mismatches

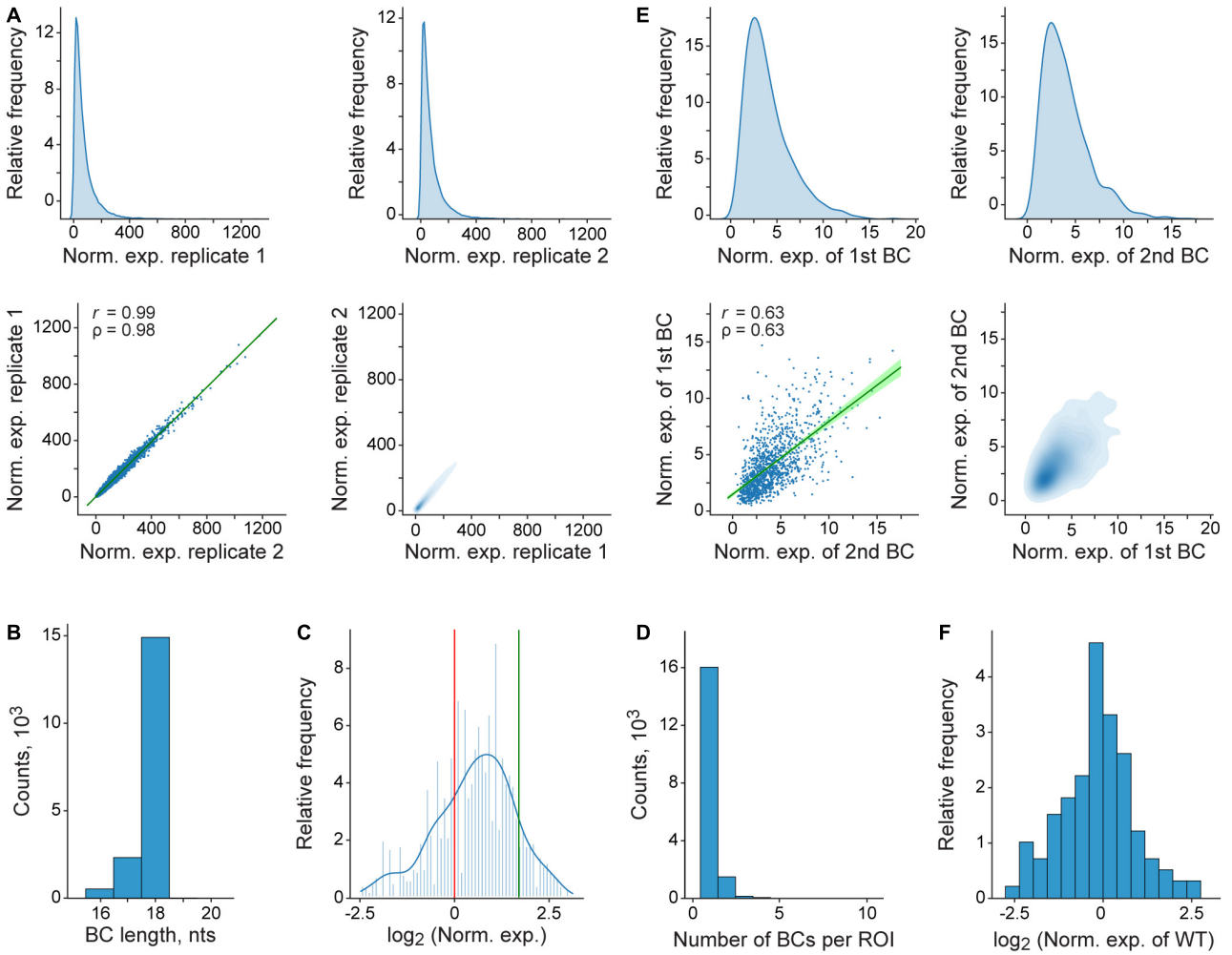


FIGURE 4 | Representative plots generated by the pipeline. **(A)** Distribution and correlation of normalized expression values for the entire MPRA library. At the top, density plots of normalized expression values obtained for genuine BCs in each of the two replicates are shown. At the bottom, the correlation of these values between the replicates is visualized as regular (left) and density (right) scatterplots. In the left scatterplot, the regression line and 99% confidence intervals are shown in green; *r* and ρ denote Pearson's correlation coefficient and Spearman's rank correlation coefficient, respectively. **(B)** Genuine BC length distribution for the entire MPRA library. **(C)** Histogram and a kernel density estimation of normalized expression values for the entire MPRA library that were averaged over replicates and normalized to the wild-type reference construct. The expression levels of the wild-type and the deltaC mutant references are indicated by red and green vertical lines, respectively. **(D)** Distribution of the genuine BC number per ROI for the entire MPRA library. **(E)** Effect of the BC sequences on normalized expression values as estimated by using a subset of the ROIs, each associated with more than one BC. For each such ROI, only two different BCs, which are randomly assigned to groups "1st BC" and "2nd BC" are used for the comparison (for the ROIs associated with three or more BCs, only two of them are randomly sampled). At the top, density plots of normalized expression values obtained for BCs from the groups "1st BC" and "2nd BC" are shown. At the bottom, the correlation of these values between the groups is visualized as regular (left) and density (right) scatterplots. The rest of the description of the left scatterplot is as in **(A)**. **(F)** Distribution of normalized expression values of genuine BCs, each associated with the ROI of the wild-type (WT) sequence. For **(A–F)**, it is worthwhile noting that the plots shown were generated by using the entire fastq file obtained (see section "Materials and Methods"), from which the example data set was randomly sampled.

are allowed for BCs of the example data set, as suggested previously (Akhtar et al., 2013).

Because identification of similar BCs by the means of alignment approaches is rather time-consuming, especially for thousands or even millions of sequences to compare (Song et al., 2014; Zielezinski et al., 2017), the MPRAdecoder pipeline first preselects candidate BCs for their subsequent pairwise sequence alignment (**Figure 3A**). The preselection is achieved by decomposing all unique BC sequences into overlapping k-mers and then revealing BCs that share identical k-mers (Haubold, 2014; Zielezinski et al., 2017). The length of k-mers (e.g., six for the example data set) is calculated as the BC length/(s+ 1) rounded down to the nearest whole number. Next, BCs sharing each particular k-mer are directly compared by using the pairwise sequence alignment (see above), taking into account their normalized read counts (RPM values). Then, similar BCs are grouped into clusters, and a number of quality control steps are applied to ensure the absence of overlap between the clusters (ambiguous cases are removed).

After that, for each cluster, it is verified whether the most abundant ROI associated with the most abundant BC is supported by the majority of normalized read counts obtained for all ROI sequences present in a cluster (**Figure 3B**). As a default setting, an arbitrary cutoff at ≥ 0.9 (specified in the configuration file) is used, similar to earlier studies (Akhtar et al., 2013; Mogno et al., 2013). If the criterion is not satisfied, probably due to associations of the same BC with different ROIs during the cloning by a chance or formation of chimeric molecules during PCR amplification of the mapping samples (Omelina et al., 2019), the entire cluster is excluded from the downstream analysis. If the criterion is satisfied, the most abundant BC and all other BCs are recognized as genuine and its mutant variants, respectively (the appropriate information is saved in a tab-delimited text file), and the RPM values of all BCs in such cluster are summed for each replicate of each sample. Eventually, all genuine BC sequences are different from each other by at least $s + 1$ position(s) (e.g., three for the example data set).

Data Normalization and Visualization

Once genuine BCs are identified, their RPM values in the normalization and expression replicates are averaged. Next, for each genuine BC, the normalized expression value is calculated as a ratio between its expression and normalization RPM values. Then, if reference constructs were spiked in the plasmid library, the pipeline can further normalize data by dividing them by the value obtained for one of these references (specified in the configuration file; e.g., for the wild-type construct in the case of the example data set). After that, values obtained with different genuine BCs but for the same ROI sequence are averaged. The raw and normalized read counts per unique BC–ROI association for each replicate of the mapping samples and per unique BC for each replicate of the expression and normalization samples, the RPM values averaged over these replicates as well as the ultimate expression values obtained for genuine BCs after each step of the data normalization and averaging are saved as tab-delimited text files. Also, the important details of data processing are reported in additional files. Among them are the numbers of allowed mismatches in the expected parts of the reads; the list of input fastq files used for a run; and statistics on (i) total and effective read counts per fastq file, (ii) numbers of unique and genuine BCs, and (iii) numbers of genuine BCs per ROI.

Finally, the pipeline generates a number of plots to help evaluate data quality and interpret the results (**Figure 4**). In particular, the reproducibility of the measurements between the replicates of the expression and normalization samples, the potential influence of the BC sequences on the measurements, and the sequence peculiarities of the ROIs with different properties are visualized.

Performance of the Pipeline

The pipeline can process 1 million reads of a non-demultiplexed fastq file in ∼20 min using the hardware and software specified in Materials and Methods. For larger data sets, the processing time can be estimated by assuming a linear dependence on the read number.

DISCUSSION

MPRAs are becoming widely used as an effective tool to assess functionality of *cis*-regulatory DNA elements in a high-throughput manner (Ernst et al., 2016; Rabani et al., 2017; Mattioli et al., 2019; Shigaki et al., 2019; Choi et al., 2020; Davis et al., 2020; Ireland et al., 2020; King et al., 2020; Klein et al., 2020; Morgan et al., 2020; Renganaath et al., 2020). In addition, several modifications to the approach have been described that broaden its applicability (Rosenberg et al., 2015; Shen et al., 2016; Safra et al., 2017). Accordingly, to simplify the design of the MPRA experiments as well as to analyze their results, a number of bioinformatics pipelines have been developed, the majority of which were, however, so far validated primarily for studies with predetermined sequences of both ROIs and BCs or, at least, ROIs (Georgakopoulos-Soares et al., 2017; Ghazi et al., 2018; Kalita et al., 2018; Ashuach et al., 2019; Myint et al., 2019; Niroula et al., 2019; Gordon et al., 2020; Qiao et al., 2020; Yang et al., 2021).

The MPRAdecoder pipeline is primarily intended for the processing of data obtained for MPRA libraries generated using oligonucleotides with randomized stretches of nucleotides for cloning the ROI and BC sequences. Such libraries are most suitable for the investigation of the properties of all possible sequence variants within a certain small region of a regulatory element. Considering the current capabilities of NGS as well as the necessity for several different BCs per ROI, the length of the region that can be subjected to saturation mutagenesis is in the range of 8–10 nts. The need for multiple BCs per ROI is dictated by the following two main factors. First, the BC sequences themselves might influence the measurements performed (Ernst et al., 2016; Ulirsch et al., 2016; **Figure 4F**), most probably due to occasional occurrence of binding sites for specific DNA- or RNA-binding proteins or microRNA in them. Therefore, in order to identify and exclude such cases, it is necessary to analyze each ROI sequence in combination with different BCs. Second, mutations may appear in both the ROI and BC sequences due to errors in PCR amplification and NGS although the frequency of such events was previously estimated to be relatively low (the error rate per nt $\leq 0.3\%$) (Pfeiffer et al., 2018; Ma et al., 2019). At the same time, all possible variants of the short ROI sequence are expected to be present in a high-quality MPRA library, making identification of mutant ROI variants in the reads practically impossible. However, the use of multiple BCs for each ROI allows detecting outliers, which can be, in particular, caused by mutated ROI sequences, and excluding them from the analysis.

Multiple BCs per ROI can be simply ensured by a longer sequence of the BCs compared to the ROIs (e.g., 18 and 8 nts, respectively, in the example MPRA library). In addition, such design allows excluding as much as possible mutant or just very similar BC sequences from the analysis. Namely, only such BCs (referred to as genuine) (Akhtar et al., 2013; Omelina et al., 2019) are used, which sequences differ from each other by at least a certain number of nts. For example, when predefined BCs up to 20 nts in length are used, the difference between each pair of them of at least at two to three positions is typically set (Patwardhan et al., 2009; Sharon et al., 2012). For BCs with random sequences of 16 nts in length, the minimum difference at three positions

also provides reliable measurements (Akhtar et al., 2013, 2014). In our case, we linked the allowed error rate in the BC sequences (as well as in all other parts of the reads, except for the ROI, in which we do not allow errors by default) with the experimentally determined error rate detected for fixed sequences amplified and sequenced in same conditions. Note that with the ROI length of 8 nts, a total of 4^8 = 65,536 sequence variants are possible, whereas the BC length of 18 nts provides 4^{18} = 68,719,476,736 variants. Of the latter, obviously, not all can be genuine BCs (satisfy the Levenshtein distance $\geq$ 3) (Faircloth and Glenn, 2012; Hawkins et al., 2018), but nevertheless, each ROI can be associated with more than enough number of different BCs.

The use of oligonucleotides with randomized stretches of nucleotides to clone the ROIs and BCs as well as regular primers to amplify the mapping, normalization, and expression samples means that the following considerations should be taken into account during the processing of raw MPRA data. First, although synthetic oligonucleotides are purified by polyacrylamide gel electrophoresis (PAGE) or high-performance liquid chromatography (HPLC), their actual length in the preparation may vary due to the presence of deletions (more often) and insertions (less often) (**Figure 4B**). Second, our experience shows that most errors found in the reads come from imperfection in oligonucleotide primer synthesis and purification (however, this could strongly depend on a supplier). Therefore, substitutions, deletions, and insertions are quite possible in the sequences of the ROIs and BCs as well as in the regions of the constant parts flanking them (that were generated by oligonucleotides used at the plasmid library cloning step). The same is true for the edges of PCR-amplified products, which are introduced by appropriate primer pairs. Along with the general drop in the quality of sequencing toward the end of the reads, this is the main reason why we allow a fairly high percentage of errors ($\sim$10%) in all expected parts of the reads. The described issues with the use of synthesized oligonucleotides are generally consistent with previous studies (Faircloth and Glenn, 2012; Hawkins et al., 2018). In addition, considering the possible variation in the BC length, especially its shortening (**Figure 4B**), it seems reasonable to equip the reference constructs that can be spiked into an MPRA library with slightly longer BC sequences (e.g., 20 nts in the example MPRA library). This could minimize the chance of accidental coincidence of sequences of the reference BC and a random BC.

Because many of the pipeline settings are arbitrary (set in the configuration file), it is important to note the following. First, of course, it is possible to set the allowed error level for all expected parts of reads to 0%; however, in the case of the example data set, this leads to a decrease in the number of genuine BCs by more than two times compared with the default settings described above. Second, because it is well known that the quality of sequencing gradually decreases toward the end of the reads, it seems appropriate to map the mCP3 and neCP2 regions in the reads not completely, but only by their beginnings. In particular, the use of only 10 instead of 17 nts for mCP3 and 20 instead of 86 nts for neCP2 for the example data set ultimately makes it possible to detect more than $\sim$1.5 times more genuine BCs with the error level in all parts of the reads set to 0%, but this gives only negligible gain (<0.1%) with the default settings described above. Third, the difference in the number of minimum reads, in which unique BCs should be detected in replicates of the mapping and normalization samples (parameters *m* and *n*), is associated with the fact that, when performing the mapping procedure, it is more important to identify the fact of different BC–ROI association(s) although data from the normalization samples are eventually quantified. Moreover, both of these parameters, as well as the parameter *e*, which determines the minimum number of reads for each unique BC in replicates of the expression samples, largely depend on both the complexity of a particular MPRA library (the number of unique clones in it) and the sequencing depth of the samples. Fourth, the threshold level of 0.9 controlling the identification of genuine BCs can be increased if necessary. This parameter is also highly dependent on the expected number of unique BC–ROI associations in the samples and their sequencing depth.

Although it is strongly recommended to obtain at least two biological replicates of the mapping, normalization, and expression samples, we notice that the pipeline nevertheless can process single replicates of these samples as well. This option can be useful when performing pilot experiments for a quick and preliminary evaluation of the results. Also, it is possible to load raw data obtained for different MPRA libraries into the pipeline simultaneously.

Finally, the results obtained for the example data set (**Figure 4C**) indicate that sequence variations in the region located after the TTS (which is not present in mature mRNA molecules) are able to substantially influence the reporter transcript level. This suggests a potentially high regulatory potential of the sequences located at the 3′-ends of genes, which has not yet been systematically studied.

DATA AVAILABILITY STATEMENT

The MPRAcode code written in Python is publicly available at https://github.com/Code-master2020/MPRAdecoder. The example input data as well as expected outputs are included in the GitHub repository. Detailed information on program can be found in the GitHub repository.

AUTHOR CONTRIBUTIONS

AL and AP conceived the study. AL, EO, and AI developed the pipeline. EO and AL performed experiments and applied the pipeline to the obtained data sets. AP supervised the project. AP, EO, and AL wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was mainly supported by the Russian Science Foundation Grant 16-14-10288 and in part of the preparation and deposition of the materials to the GitHub repository by the Russian Science Foundation Grant 20-74-00137.

ACKNOWLEDGMENTS

We thank Lyubov A. Yarinich and Mikhail O. Lebedev for the generation of the MPRA plasmid library, Lyubov A. Yarinich for critical reading of the manuscript, and Petr P. Laktionov and Daniil A. Maksimov for the assistance with the Illumina DNA sequencing that was performed at the Molecular and Cellular Biology core facility of the Institute of Molecular and Cellular Biology of the Siberian Branch of the Russian Academy of Sciences.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.618189/full#supplementary-material

REFERENCES

1000 Genomes Project Consortium; Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Akhtar, W., de Jong, J., Pindyurin, A. V., Pagie, L., Meuleman, W., de Ridder, J., et al. (2013). Chromatin position effects assayed by thousands of reporters integrated in parallel. *Cell* 154, 914–927. doi: 10.1016/j.cell.2013.07.018

Akhtar, W., Pindyurin, A. V., de Jong, J., Pagie, L., ten Hoeve, J., Berns, A., et al. (2014). Using TRIP for genome-wide position effect analysis in cultured cells. *Nat. Protoc.* 9, 1255–1281. doi: 10.1038/nprot.2014.072

Albert, F. W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16, 197–212. doi: 10.1038/nrg3891

Ashuach, T., Fischer, D. S., Kreimer, A., Ahituv, N., Theis, F. J., and Yosef, N. (2019). MPRAnalyze: statistical framework for massively parallel reporter assays. *Genome Biol.* 20:183. doi: 10.1186/s13059-019-1787-z

Boldyreva, L. V., Yarinich, L. A., Kozhevnikova, E. N., Ivankin, A. V., Lebedev, M. O., and Pindyurin, A. V. (2021). Fine gene expression regulation by minor sequence variations downstream of the polyadenylation signal. *Mol. Biol. Rep.* 48, 1539–1547. doi: 10.1007/s11033-021-06160-z

Choi, J., Zhang, T., Vu, A., Ablain, J., Makowski, M. M., Colli, L. M., et al. (2020). Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat. Commun.* 11:2718. doi: 10.1038/s41467-020-16590-1

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771. doi: 10.1093/nar/gkp1137

Davis, J. E., Insigne, K. D., Jones, E. M., Hastings, Q. A., Boldridge, W. C., and Kosuri, S. (2020). Dissection of c-AMP response element architecture by using genomic and episomal massively parallel reporter assays. *Cell Syst.* 11, 75–85. doi: 10.1016/j.cels.2020.05.011

Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., et al. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* 34, 1180–1190. doi: 10.1038/nbt.3678

Faircloth, B. C., and Glenn, T. C. (2012). Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One* 7:e42543. doi: 10.1371/journal.pone.0042543

Georgakopoulos-Soares, I., Jain, N., Gray, J. M., and Hemberg, M. (2017). MPRAnator: a web-based tool for the design of massively parallel reporter assay experiments. *Bioinformatics* 33, 137–138. doi: 10.1093/bioinformatics/btw584

Ghazi, A. R., Chen, E. S., Henke, D. M., Madan, N., Edelstein, L. C., and Shaw, C. A. (2018). Design tools for MPRA experiments. *Bioinformatics* 34, 2682–2683. doi: 10.1093/bioinformatics/bty150

Gordon, M. G., Inoue, F., Martin, B., Schubach, M., Agarwal, V., Whalen, S., et al. (2020). lentiMPRA and MPRAflow for high-throughput functional characterization of gene regulatory elements. *Nat. Protoc.* 15, 2387–2412. doi: 10.1038/s41596-020-0333-5

Grossman, S. R., Zhang, X., Wang, L., Engreitz, J., Melnikov, A., Rogov, P., et al. (2017). Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U.S.A.* 114, E1291–E1300. doi: 10.1073/pnas.1621150114

Haberle, V., and Lenhard, B. (2012). Dissecting genomic regulatory elements in vivo. *Nat. Biotechnol.* 30, 504–506. doi: 10.1038/nbt.2266

Haubold, B. (2014). Alignment-free phylogenetics and population genetics. *Brief. Bioinform.* 15, 407–418. doi: 10.1093/bib/bbt083

Hawkins, J. A., Jones, S. K. Jr., Finkelstein, I. J., and Press, W. H. (2018). Indel-correcting DNA barcodes for high-throughput sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 115, E6217–E6226. doi: 10.1073/pnas.1802640115

Inoue, F., and Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics* 106, 159–164. doi: 10.1016/j.ygeno.2015.06.005

Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., et al. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* 27, 38–52. doi: 10.1101/gr.212092.116

Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N., and Yosef, N. (2019). Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* 25, 713–727. doi: 10.1016/j.stem.2019.09.010

Ireland, W. T., Beeler, S. M., Flores-Bautista, E., McCarty, N. S., Röschinger, T., Belliveau, N. M., et al. (2020). Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. *eLife* 9:e55308. doi: 10.7554/eLife.55308

Kalita, C. A., Moyerbrailean, G. A., Brown, C., Wen, X., Luca, F., and Pique-Regi, R. (2018). QuASAR-MPRA: accurate allele-specific analysis for massively parallel reporter assays. *Bioinformatics* 34, 787–794. doi: 10.1093/bioinformatics/btx598

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., et al. (2014). Defining functional DNA elements in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 111, 6131–6138. doi: 10.1073/pnas.1318948111

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., et al. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.* 23, 800–811. doi: 10.1101/gr.144899.112

King, D. M., Hong, C. K. Y., Shepherdson, J. L., Granas, D. M., Maricque, B. B., and Cohen, B. A. (2020). Synthetic and genomic regulatory elements reveal aspects of cis-regulatory grammar in mouse embryonic stem cells. *eLife* 9:e41279. doi: 10.7554/eLife.41279

Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R. J. A., et al. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* 10:3583. doi: 10.1038/s41467-019-11526-w

Klein, J. C., Agarwal, V., Inoue, F., Keith, A., Martin, B., Kircher, M., et al. (2020). A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat. Methods* 17, 1083–1091. doi: 10.1038/s41592-020-0965-y

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., and Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* 24, 1595–1602. doi: 10.1101/gr.173518.114

Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc. Natl. Acad. Sci. U.S.A.* 109, 19498–19503. doi: 10.1073/pnas.1210678109

Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., et al. (2019). Analysis of error profiles in deep next-generation sequencing data. *Genome Biol.* 20:50. doi: 10.1186/s13059-019-1659-6

Maricque, B. B., Dougherty, J. D., and Cohen, B. A. (2017). A genome-integrated massively parallel reporter assay reveals DNA sequence determinants of cis-regulatory activity in neural cells. *Nucleic Acids Res.* 45:e16. doi: 10.1093/nar/gkw942

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Mattioli, K., Volders, P.-J., Gerhardinger, C., Lee, J. C., Maass, P. G., Melé, M., et al. (2019). High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res.* 29, 344–355. doi: 10.1101/gr.242222.118

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277. doi: 10.1038/nbt.2137

Mogno, I., Kwasnieski, J. C., and Cohen, B. A. (2013). Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res.* 23, 1908–1915. doi: 10.1101/gr.157891.113

Morgan, R. A., Ma, F., Unti, M. J., Brown, D., Ayoub, P. G., Tam, C., et al. (2020). Creating new β-globin-expressing lentiviral vectors by high-resolution mapping of locus control region enhancer sequences. *Mol. Ther. Methods Clin. Dev.* 17, 999–1013. doi: 10.1016/j.omtm.2020.04.006

Mulvey, B., Lagunas, T. Jr., and Dougherty, J. D. (2021). Massively parallel reporter assays: defining functional psychiatric genetic variants across biological contexts. *Biol. Psychiatry* 89, 76–89. doi: 10.1016/j.biopsych.2020.06.011

Myint, L., Avramopoulos, D. G., Goff, L. A., and Hansen, K. D. (2019). Linear models enable powerful differential activity analysis in massively parallel reporter assays. *BMC Genomics* 20:209. doi: 10.1186/s12864-019-5556-x

Narlikar, L., and Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomic. Proteomic.* 8, 215–230. doi: 10.1093/bfgp/elp014

Niroula, A., Ajore, R., and Nilsson, B. (2019). MPRAscore: robust and non-parametric analysis of massively parallel reporter assays. *Bioinformatics* 35, 5351–5353. doi: 10.1093/bioinformatics/btz591

O'Connell, D. J., Kolde, R., Sooknah, M., Graham, D. B., Sundberg, T. B., Latorre, I., et al. (2016). Simultaneous pathway activity inference and gene expression analysis using RNA sequencing. *Cell Syst* 2, 323–334. doi: 10.1016/j.cels.2016.04.011

Omelina, E. S., Ivankin, A. V., Letiagina, A. E., and Pindyurin, A. V. (2019). Optimized PCR conditions minimizing the formation of chimeric DNA molecules from MPRA plasmid libraries. *BMC Genomics* 20:536. doi: 10.1186/s12864-019-5847-2

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30, 265–270. doi: 10.1038/nbt.2136

Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* 27, 1173–1175. doi: 10.1038/nbt.1589

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J. L., et al. (2018). Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.* 8:10950. doi: 10.1038/s41598-018-29325-6

Qiao, D., Zigler, C. M., Cho, M. H., Silverman, E. K., Zhou, X., Castaldi, P. J., et al. (2020). Statistical considerations for the analysis of massively parallel reporter assays data. *Genet. Epidemiol.* 44, 785–794. doi: 10.1002/gepi.22337

Rabani, M., Pieper, L., Chew, G.-L., and Schier, A. F. (2017). A massively parallel reporter assay of 3′ UTR sequences identifies in vivo rules for mRNA degradation. *Mol. Cell* 68, 1083–1094. doi: 10.1016/j.molcel.2017.11.014

Renganaath, K., Cheung, R., Day, L., Kosuri, S., Kruglyak, L., and Albert, F. W. (2020). Systematic identification of cis-regulatory variants that cause gene expression differences in a yeast cross. *eLife* 9:e62669. doi: 10.7554/eLife.62669

Rojano, E., Seoane, P., Ranea, J. A. G., and Perkins, J. R. (2019). Regulatory variants: from detection to predicting impact. *Brief. Bioinform.* 20, 1639–1654. doi: 10.1093/bib/bby039

Rosenberg, A. B., Patwardhan, R. P., Shendure, J., and Seelig, G. (2015). Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell* 163, 698–711. doi: 10.1016/j.cell.2015.09.054

Safra, M., Nir, R., Farouq, D., Vainberg Slutskin, I., and Schwartz, S. (2017). TRUB1 is the predominant pseudouridine synthase acting on mammalian mRNA via a predictable and conserved code. *Genome Res.* 27, 393–406. doi: 10.1101/gr.207613.116

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., et al. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.* 30, 521–530. doi: 10.1038/nbt.2205

Shen, S. Q., Myers, C. A., Hughes, A. E. O., Byrne, L. C., Flannery, J. G., and Corbo, J. C. (2016). Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* 26, 238–255. doi: 10.1101/gr.193789.115

Shigaki, D., Adato, O., Adhikari, A. N., Dong, S., Hawkins-Hooker, A., Inoue, F., et al. (2019). Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.* 40, 1280–1291. doi: 10.1002/humu.23797

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., et al. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.* 45, 1021–1028. doi: 10.1038/ng.2713

Song, K., Ren, J., Reinert, G., Deng, M., Waterman, M. S., and Sun, F. (2014). New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief. Bioinform.* 15, 343–353. doi: 10.1093/bib/bbt067

Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., et al. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res.* 21, 1139–1149. doi: 10.1101/gr.119016.110

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., et al. (2016). Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* 165, 1519–1529. doi: 10.1016/j.cell.2016.04.027

Trauernicht, M., Martinez-Ara, M., and van Steensel, B. (2020). Deciphering gene regulation using massively parallel reporter assays. *Trends Biochem. Sci.* 45, 90–91. doi: 10.1016/j.tibs.2019.10.006

Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., et al. (2016). Systematic functional dissection of common genetic variation affecting red blood cell traits. *Cell* 165, 1530–1545. doi: 10.1016/j.cell.2016.04.048

van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., et al. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nat. Biotechnol.* 35, 145–153. doi: 10.1038/nbt.3754

van Arensbergen, J., Pagie, L., FitzPatrick, V. D., de Haas, M., Baltissen, M. P., Comoglio, F., et al. (2019). High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* 51, 1160–1169. doi: 10.1038/s41588-019-0455-2

Verfaillie, A., Svetlichnyy, D., Imrichova, H., Davie, K., Fiers, M., Kalender Atak, Z., et al. (2016). Multiplex enhancer-reporter assays uncover unsophisticated TP53 enhancer logic. *Genome Res.* 26, 882–895. doi: 10.1101/gr.204149.116

Vvedenskaya, I. O., Zhang, Y., Goldman, S. R., Valenti, A., Visone, V., Taylor, D. M., et al. (2015). Massively systematic transcript end readout, "MASTER": transcription start site selection, transcriptional slippage, and transcript yields. *Mol. Cell* 60, 953–965. doi: 10.1016/j.molcel.2015.10.029

White, M. A., Myers, C. A., Corbo, J. C., and Cohen, B. A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U.S.A.* 110, 11952–11957. doi: 10.1073/pnas.1307449110

Yang, Z., Wang, C., Erjavec, S., Petukhova, L., Christiano, A., and Ionita-Laza, I. (2021). A semisupervised model to predict regulatory effects of genetic variants at single nucleotide resolution using massively parallel reporter assays. *Bioinformatics* doi: 10.1093/bioinformatics/btab040 [Epub ahead of print].

Zielezinski, A., Vinga, S., Almeida, J., and Karlowski, W. M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol.* 18:186. doi: 10.1186/s13059-017-1319-7

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 May 2021
doi: 10.3389/fgene.2021.621049

Association of *CASR, CALCR,* and *ORAI1* Genes Polymorphisms With the Calcium Urolithiasis Development in Russian Population

Maria M. Litvinova[1,2†], Kamil Khafizov[3†], Vitaly I. Korchagin[4†], Anna S. Speranskaya[4†], Aliy Yu. Asanov[1†], Alina D. Matsvay[3,5†], Daniil A. Kiselev[1,5†], Diana V. Svetlichnaya[1,6†], Sevda Z. Nuralieva[1†], Alexey A. Moskalev[7†] and Tamara V. Filippova[1†]*

[1] Department of Medical Genetics, Ministry of Public Health of the Russian Federation, I. M. Sechenov First Moscow State Medical University, Sechenov University, Moscow, Russia, [2] Moscow Health Department, The Loginov Moscow Clinical Scientific Center, Moscow, Russia, [3] Moscow Institute of Physics and Technology, National Research University, Dolgoprudny, Russia, [4] Federal Service on Consumers' Rights Protection and Human Well-Being Surveillance, Central Research Institute for Epidemiology, Moscow, Russia, [5] Center of Strategic Planning of FMBA of Russia, Moscow, Russia, [6] Moscow Regional Research and Clinical Institute (MONIKI), Moscow, Russia, [7] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

OPEN ACCESS

Edited by:
Angel Carlos Roman, University of Extremadura, Spain

Reviewed by:
Dusanka Savic Pavicevic, University of Belgrade, Serbia
Yuriy L. Orlov, I. M. Sechenov First Moscow State Medical University, Russia

***Correspondence:**
Maria M. Litvinova
mariya.litvinova@gmail.com

†ORCID:
Maria M. Litvinova
orcid.org/0000-0002-1863-3768
Kamil Khafizov
orcid.org/0000-0001-5524-0296
Vitaly I. Korchagin
orcid.org/0000-0003-2264-6294
Anna S. Speranskaya
orcid.org/0000-0001-6326-1249
Aliy Yu. Asanov
orcid.org/0000-0002-5388-8133
Alina D. Matsvay
orcid.org/0000-0002-6301-9169
Daniil A. Kiselev
orcid.org/0000-0001-8074-8411
Diana V. Svetlichnaya
orcid.org/0000-0001-6497-8487
Sevda Z. Nuralieva
orcid.org/0000-0003-0030-5157
Alexey A. Moskalev
orcid.org/0000-0002-3248-1633
Tamara V. Filippova
orcid.org/0000-0002-9916-8617

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

Received: *05 November 2020*
Accepted: *16 April 2021*
Published: *12 May 2021*

Kidney stone disease is an urgent medical and social problem. Genetic factors play an important role in the disease development. This study aims to establish an association between polymorphisms in genes coding for proteins involved in calcium metabolism and the development of calcium urolithiasis in Russian population. In this case-control study, we investigated 50 patients with calcium urolithiasis (experimental group) and 50 persons lacking signs of kidney stone disease (control group). For molecular genetic analysis we used a previously developed gene panel consisting of 33 polymorphisms in 15 genes involved in calcium metabolism: *VDR*, *CASR*, *CALCR*, *OPN*, *MGP*, *PLAU*, *AQP1*, *DGKH*, *SLC34A1*, *CLDN14*, *TRPV6*, *KLOTHO*, *ORAI1*, *ALPL*, and *RGS14.* High-throughput target sequencing was utilized to study the loci of interest. Odds ratios and 95% confidence intervals were used to estimate the association between each SNP and risk of urolithiasis development. Multifactor dimensionality reduction analysis was also carried out to analyze the gene-gene interaction. We found statistically significant (unadjusted p-value < 0.05) associations between calcium urolithiasis and the polymorphisms in the following genes: *CASR* rs1042636 (OR = 3.18 for allele A), *CALCR* rs1801197 (OR = 6.84 for allele A), and *ORAI1* rs6486795 (OR = 2.25 for allele C). The maximum OR was shown for AA genotypes in loci rs1042636 (*CASR*) and rs1801197 (*CALCR*) (OR = 4.71, OR = 11.8, respectively). After adjustment by Benjamini-Hochberg FDR we found only *CALCR* (rs1801197) was significantly associated with the risk of calcium urolithiasis development. There was no relationship between recurrent course of the disease and family history of urolithiasis in investigated patients. Thus we found a statistically significant association of polymorphism rs1801197 (gene *CALCR*) with calcium urolithiasis in Russian population.

Keywords: kidney stone disease, urolithiasis, calcium stones, calcium urolithiasis, *CALCR, CASR*, *ORAI1, CLDN14* gene

INTRODUCTION

Kidney stone disease (KSD) has been known to be one of the most excruciating chronic diseases. It is estimated to affect nearly 5% of women and 12% of men during their lifetime, and is considered to be the third most frequent urological disorder (Worcester and Coe, 2010). Multiple studies have revealed that genetics alter the risk of KSD development alongside the environmental factors. It is believed that the vast majority of cases are, in fact, multifactorial.

Deciphering the molecular substrate for the etiopathogenesis of urolithiasis is of outmost importance for developing diagnostic tools and therapy strategies. In most cases, KSD is caused by the formation of calcium concrements, supplying grounds for research into the calcium metabolism impairments in those affected by the disease. Numerous works have been published that elucidate hidden associations between polymorphisms in genes of calcium metabolism and the development of KSD (Filippova et al., 2020). To our knowledge, very few investigations were performed to look into these associations in Russian population (Apolihin et al., 2015; Apolikhin et al., 2016, 2017). According to various reports, the development of calcium urolithiasis has been attributed to polymorphisms in several genes: *VDR, CASR, CALCR, OPN, MGP, PLAU, AQP1, SLC34A1, CLDN14, KLOTHO,* and *ORAI1* (Filippova et al., 2020).

In this article, we examine possible connections between polymorphisms in genes of calcium metabolism and the risk of KSD development in Russian population.

MATERIALS AND METHODS

Patients Characteristics

In this case-control study, the experimental group featured 50 patients with KSD, and the control group consisted of 50 healthy individuals aged 1 to 70. All patients suffered from calcium oxalate urolithiasis (as verified with spectral assay of the concrements). The distribution of patients by gender in both groups is following: 13 male patients (26%), 37 female patients (74%). The mean age at onset of the disease in the group of patients with urolithiasis was 29.6 years (median 24 years). The average age of the subjects at the time of participation in the study in both groups was 38.5 years (median – 34 years). Family history on KSD was collected from all patients. No family history of KSD was found for any people in the control group.

The study was confirmed by the ethics committee of Sechenov University. All participants signed informed consent prior to entering the research program.

Genetic Analysis

A previously developed gene panel was used to evaluate possible correlations between the development of KSD and polymorphisms in the genes of calcium metabolism: *VDR* (rs1544410, rs731236), *CASR* (rs6776158, rs7652589, rs1501899, rs1801725, rs1042636, rs1801726), *CALCR* (rs1042138, rs1801197), *OPN* (rs2853749, rs2853750, rs1126616, rs4754), *MGP* (rs4236), *PLAU* (rs4065), *AQP1* (rs12669187, rs1000597), *DGKH* (rs4142110), *SLC34A1* (rs12654812), *CLDN14* (rs219781, rs219780, rs219779, rs219778, rs219777), *TRPV6* (rs4987667, rs4987682), *KLOTHO* (rs3752472), *ORAI1* (rs12313273, rs6486795, rs7135617), *ALPL* (rs1256328), and *RGS14* (rs11746443) (Filippova et al., 2020).

Peripheral blood was used as a source of genomic DNA. The DNA was extracted with DNeasy Blood & Tissue Kit (Qiagen) on QIAcube automated extraction platform (Qiagen) according to the manufacturer's instructions.

Obtained DNA was PCR-amplified with a primer panel specifically developed for this study. The panel featured 68 primers divided into 2 pools to optimize amplification and minimize possible artifacts. Target PCR was conducted with AmpliSens reagents on QuantStudio 5 real-time PCR system (Thermo Fisher Scientific).

NGS libraries were prepared according to an in-house protocol. T4 Polynucleotide Kinase and T4 DNA Ligase (both New England Biolabs) were utilized in compliance with the manufacturer's directions with slight modifications to maximize the output.

PCR products and NGS libraries at any stage of library preparation were purified with Sera-Mag SpeedBeads (General Electric) according to the manufacturer's protocol to ensure recovery of the fragments of an optimal length. DNA concentrations were measured with Qubit 2.0 Fluorometer (Thermo Fisher Scientific) using Qubit^{TM} dsDNA High Sensitivity Assay Kit (Thermo Fisher Scientific). The quality of the final libraries was assessed with on-chip capillary electrophoresis on Agilent 2100 Bioanalyzer (Agilent Technologies) with Agilent High Sensitivity DNA Kit (Agilent Technologies).

The libraries were sequenced on Ion S5 (Thermo Fisher Scientific) high-throughput sequencing platform with Ion 520 & Ion 530 Kit-Chef (Thermo Fisher Scientific) reagents on Ion 530 Chip (Thermo Fisher Scientific).

Bioinformatic Analysis of Sequencing Data

The analysis of primary sequencing data included several stages: (1) read quality filtering with PRINSEQ-lite (Schmieder and Edwards, 2011); (2) mapping to the reference human genome (GRCh38.p7, PRJNA31257) with Burrows-Wheeler Aligner (BWA-mem, v_0.7.13) (Schmieder and Edwards, 2011); (3) searching for single-nucleotide variants (SNVs) with Genome Analysis Toolkit (GATK version: 4.0.11.0) (McKenna et al., 2010). SAMtools v_1.3.1 (Li et al., 2009) and Picard toolkit v_2.18.17 were used for working with sam/bam files. VEP tool (McLaren et al., 2016) combined with 94_GRCh38 cash was used for primary variant annotation. Identified SNVs were validated manually in Tablet (Milne et al., 2013).

Statistical Analysis

The final matrix of 28 SNPs obtained after removing SNPs with high linkage disequilibrium, was using for association studies by PLINK v1.90b6.9 (Steiß et al., 2012; Chang et al., 2015). Statistical analyses were conducted with the standard functions of the R

environment and packages (R Core Team, 2019). Differences in allelic and genotypic distributions were estimated by Fisher's exact test with Lancaster's mid-p adjustment (Lancaster, 1961). Hardy-Weinberg equilibrium (HWE) in controls was calculated using the chi-squared test with continuity correction (Graffelman, 2015). To estimate the association between each SNP and KSD risk the odds ratios (OR) and 95% confidence intervals (CIs) were calculated using exact methods (median-unbiased estimation (mid-p), maximum likelihood estimation (Fisher) and small sample adjustment) by Epitools package (Tomas, 2017). Differences with the *p*-value less than 0.05 were considered statistically significant.

For detecting multilocus genotype combinations which may predict disease risk, multifactor dimensionality reduction (MDR) approach was used by MDR 3.0.2 (build 2) software package (Hahn et al., 2003). MDR is a non-parametric data mining method that assumes no genetic model and has been supported by numerous studies of gene-gene and gene-environment interactions (Ritchie et al., 2001; Cho et al., 2004; Andrew et al., 2006; Brassat et al., 2006). Cross-validation and 1000-fold permutation testing were used to find optimal models for defining disease risk.

RESULTS

Data analysis revealed a statistically significant association between the development of calcium KSD and polymorphisms in the following genes: *CASR* (rs1042636, OR = 3.18), *CALCR* (rs1801197, OR = 6.84), *ORAI1* (rs6486795, OR = 2.25). Association between SNP rs219780 of the *CLDN14* gene and urolithiasis was characterized by borderline *p*-value (OR = 2.03; $p = 0.05$). After the adjustment by Benjamini-Hochberg procedure we found only *CALCR* (rs1801197) significantly associate with the risk of calcium urolithiasis development (**Table 1**). For other studied loci of the gene panel, no statistically significant differences in allele frequencies were found between the experimental and control groups.

More than a half of the patients from the experimental group (26 patients, 52%) had a family history of KSD. In the group of patients with KSD 26 persons (52%) suffered from recurring urolithiasis. Among patients with recurring urolithiasis, 14 people (53.9%) had a family history of KSD. Among those with non-recurring urolithiasis 12 patients (50%) had a family history of the disease. Thus, no relationship was found between the recurrent course of the disease and the family history of the patients (Pearson's Chi-squared test with Yates' continuity correction, $\chi^2 = 0$, *p*-value = 1).

Comparative characteristics of genotype frequencies of genes loci *CASR* (rs1042636), *CALCR* (rs1801197), *ORAI1* (rs6486795), and *CLDN14* (rs219780), affecting the risk of KSD in our study, is shown in **Table 2**. Genotype distributions for all loci were compatible with HWE in controls. For the loci of the *CASR* and *CALCR* genes, a statistically significant difference was shown between the experimental and control groups, both in the frequency of the alleles and in the frequency of genotypes ($p < 0.05$).

Table 3 shows the significance of the dominant and recessive models for the studied polymorphisms of the *CASR* (rs1042636), *CALCR* (rs1801197), *ORAI1* (rs6486795), and *CLDN14* (rs219780) genes regarding the development of calcium urolithiasis in the Russian population.

Under the recessive model of inheritance, carriers with the AA genotypes of *CASR* (rs1042636) and AA genotypes of *CALCR* (rs1801197) had a 4.71-fold and 11.8-fold increased risk of KSD respectively. Moreover the carriers of CC/CT genotype of rs6486795 *(ORAI1)* have 2.54-fold increased risk of KSD comparing to carriers of TT genotype. After the adjustment by Benjamini-Hochberg procedure no statistically significant differences between KSD patients and controls were found for the rs219780 (*CLDN14*) and rs1042636 *(CASR)*.

Gene-gene interaction analysis using MDR approach showed that a two-locus model consisting of rs1042636 *(CASR)* and rs1801197 *(CALCR)* might have a non-linear association with the susceptibility to the KSD development. This model had an overall accuracy test of 78%, a consistency of cross-validation of 9/10, and a 1000-fold permutation *p*-value = 0.003 (**Table 4**). **Figure 1** summarizes the two-way gene interaction showing the high-risk genotype [AA + AA] of the rs1042636 (*CASR*) and rs1801197 (*CALCR*) associated with an increased KSD risk (OR = 2.59, 95% CI = 1.78-3.86).

DISCUSSION

We detected an association between the polymorphisms of *CASR* (rs1042636), *CALCR* (rs1801197), and *ORAI1* (rs6486795) genes and the development of calcium urolithiasis in the Russian population. However after adjustment by Benjamini-Hochberg FDR we found only *CALCR* (rs1801197) was significantly associated with the risk of calcium urolithiasis development.

It is known that these genes products are involved in calcium metabolism. Thus the *CASR* gene encodes a calcium-sensing receptor which senses changes of calcium concentration in an organism and controls a parathyroid hormone secretion. Activation of the parathyroid hormone synthesis stimulates the calcium release from bone tissue into the bloodstream and decreases the phosphates and calcium reabsorption in the proximal renal tubules (Vezzoli et al., 2013). The *CALCR* gene is attributable for a calcitonin receptor synthesis. CALCR interacts with the hormone calcitonin which is a functional antagonist of a parathyroid hormone and inhibits the activity of osteoclasts in the bone tissue. This in turn decreases calcium release into the bloodstream and also regulates the phosphates and calcium reabsorption in the renal tubules (Shakhssalim et al., 2014). The *ORAI1* gene encodes calcium release–activated calcium modulator type 1. This protein is required for transmembrane calcium metabolism. It is usually activated upon the depletion of internal calcium stores (Chou et al., 2011).

The association between the *CASR*, *CALCR*, and *ORAI1* genes polymorphisms and the urolithiasis development has been shown in a number of studies conducted in Italian, Indian,

TABLE 1 | Associations between the risk of calcium urolithiasis development and polymorphisms of *CASR, CALCR, ORAI1*, and *CLDN14* genes.

Gene	SNP,(risk allele)	Allele frequency		OR (95% CI)	*p*-value*	Permutation *p-value***	FDR BH****
		Case, %	Control, %				
ORAI1	rs6486795 (C)	30%	16%	2.25(1.135–4.462)	0.020	0.036	0.33
CALCR	rs1801197 (A)	94%	69%	6.84(2.87–19.26)	0.000004(<0.0001)	0.00002	0.00012
CASR	rs1042636 (A)	96%	88%	3.18(1.05–12.07)	0.041	0.046	0.43
CLDN14	rs219780 (C)	86%	75%	2.03(0.99–4.31)	0.052	0.059	0.43

**Fisher's exact test with Lancaster's mid-p adjustment was used to obtain p-values.*
***Adaptive Monte Carlo permutation test was performed.*
****Benjamini-Hochberg FDR adjustment for 28 SNP.*

TABLE 2 | Genotypes frequencies of loci associated with KSD development in the group of patients with calcium urolithiasis and in the control group.

Locus	Genotype	Genotype frequency		Fisher's exact test,*p-value*/adjusted *p-value**	HWE in controls
		Case, n (%)	Control, n (%)		
rs1042636 *(CASR)*	AA	47 (94)	38 (76)	0.008/0.016	0.70
	AG	2 (4)	12 (24)		
	GG	1 (2)	0		
rs1801197 *(CALCR)*	AA	46 (92)	24 (48)	0.0000011/0.000016	0.88
	AG	2 (4)	21 (42)		
	GG	2 (4)	5 (10)		
rs6486795 *(ORAI1)*	TT	25 (50)	36 (72)	0.074/0.52	0.74
	TC	20 (40)	12 (24)		
	CC	5 (10)	2 (4)		
rs219780(*CLDN14)*	CC	37 (74)	28 (56)	0.212/0.52	0.82
	CT	12 (24)	19 (38)		
	TT	1 (2)	3 (6)		

**Benjamini-Hochberg FDR*

TABLE 3 | Association of the *CASR*, *CALCR*, *ORAI1*, and *CLDN14* genes genotypes with the risk of calcium urolithiasis development under the different inheriting models.

Locus	Model	Group		OR (95% CI)	*p*-value*
		Case	Control		
rs1042636 *(CASR)*	Dominant(AA + AG vs. GG)	49/1	50/0	0 (0.01–8.2)***	0.5
	Recessive(AA vs. AG + GG)	47/3	38/12	4.71 (1.36–23.0)	0.013/0.052**
rs1801197 *(CALCR)*	Dominant(AA + AG vs. GG)	48/2	45/5	2.54 (0.49–20.5)	0.27
	Recessive(AA vs AG + GG)	46/4	24/26	11.8 (4.0–44.9)	0.000001/0.000008**
rs6486795*(ORAI1)*	Dominant(CC + CT vs. TT)	25/25	14/36	2.54 (1.11–5.97)	0.027/0.072**
	Recessive(CC vs. CT + TT)	5/45	2/48	2.53 (0.49–20.5)	0.274
rs219780(*CLDN14)*	Dominant(CC + CT vs. TT)	49/1	47/3	2.85 (0.32–83.9)	0.367
	Recessive(CC vs. CT + TT)	37/13	28/22	2.21 (0.96–5.28)	0.064

**Fisher's exact test with Lancaster's mid-p adjustment was used to obtain p-values.*
***Benjamini-Hochberg FDR*
****OR was calculated by small sample adjustment.*

TABLE 4 | Interaction analysis of SNPs in *CASR* and *CALCR* and risk of KSD development.

Best candidate models	Training Bal. Acc. (%)	Testing Bal. Acc. (%)	Overall Bal. Acc. (%)	CV consistency	*p*-value
rs1801197	72	72	72	10/10	0.007
rs1042636/rs1801197	78	74	78	9/10	0.003

CV, cross-validation; Bal, balanced; Acc, accuracy.

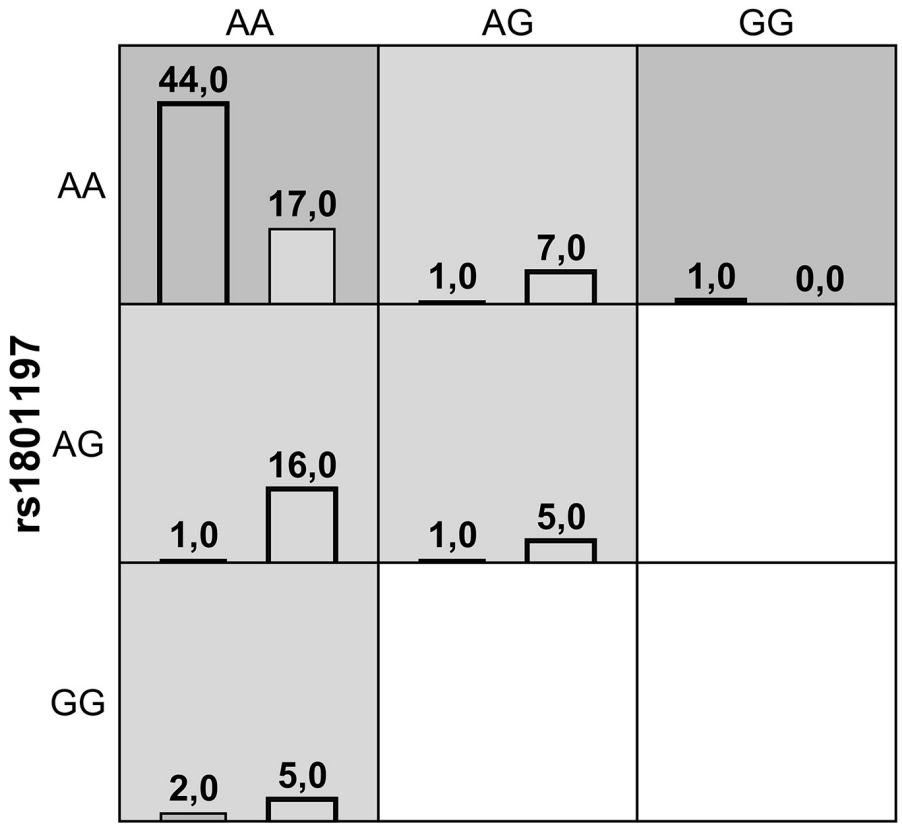


FIGURE 1 | A summary of the best two-way gene-gene interaction analysis by multifactor dimensionality reduction for 9 genotypes [rs1042636 (*CASR*) and rs1801197 (*CALCR*)] associated with increased risk of KSD. The dark shading box represents high-risk combinations and the light shading box shows low-risk combinations. The left and right columns represent the absolute number of the cases and controls, respectively.

and Iranian, Chinece populations by different researchers. The data obtained in this study are generally consistent with the data of the world literature (Corbetta et al., 2006; Scillitani et al., 2007; Shakhssalim et al., 2010; Chou et al., 2011; Vezzoli et al., 2011, 2015; Guha et al., 2015; Apolikhin et al., 2016; Qin et al., 2019). Some differences in the results of the investigations most likely can be explained by the specificity of the genetic characteristics of the Russian population, as well as by the peculiarities of the experimental group formation by different researchers.

An association between the rs1801197 polymorphism of the *CALCR* gene and urolithiasis was shown in the study of Qin et al. (2019). As a result of the meta-analysis (494 patients and 536 healthy individuals) performed by mentioned authors allele A of the locus rs1801197 was significantly associated with the risk of calcium urolithiasis development (OR for allele A was 1.987). According to our data, in the Russian population the OR for the A allele of the rs1801197 locus was 6.84 ($p < 0.0001$).

The relationship between the locus rs1042636 of the *CASR* gene and KSD was studied in populations of Italy, India, and Iran (Corbetta et al., 2006; Scillitani et al., 2007; Vezzoli et al., 2011).

Vezzoli et al. investigated an association between polimorphism rs1042636 (Arg990Gly) of the *CASR* gene and the risk of KSD development in Italian patients with primary hyperparathyroidism (OR for allele G (Gly) was 3.3) (Vezzoli et al., 2015).

Guha et al. showed the influence of the rs1042636 (Arg990Gly) polymorphism of the *CASR* gene at the development of urolithiasis in Indian population (OR for allele G (Gly) 2.21) (Guha et al., 2015).

The data on the role of rs1042636 (Arg990Gly) polymorphism of the *CASR* gene in urolithiasis development obtained by Shakhssalim et al. on the Iranian population are in a good agreement with the results of our study (Shakhssalim et al., 2010). In the mentioned study authors showed that patients with the AA genotype (Arg/Arg) at the rs1042636 locus showed a significantly higher serum ionized calcium compared to the patients with the Arg/Gly or Gly/Gly genotypes (OR for the Arg allele was 8.06).

The frequency of the rs1042636G allele according to dbSNP data[1] in Europe varies from 7 to 10%, which corresponds to the data obtained in this study (the allele rs1042636G frequency

[1]https://www.ncbi.nlm.nih.gov/snp/rs1042636#frequency_tab

in the control group in the current investigation was 12%). According to the results of our study, in Russian population the rare G allele of the locus rs1042636 may have a protective effect in relation to the KSD development. Thus, to date, in different populations different alleles of the rs1042636 locus of the *CASR* gene demonstrate an association with the risk of the urolithiasis development.

A number of studies in different countries were devoted to the investigation of the association between *ORAI1* gene polymorphisms and KSD development (Chou et al., 2011; Apolikhin et al., 2016). Thus, a study conducted in the Russian population by Apolikhin et al. revealed an association between the G allele of the *ORAI1* rs7135617 locus and an increased risk of a recurrence-free urolithiasis development (OR = 1.049). However, in the mentioned study the role of other polymorphisms of the *ORAI1* gene in the KSD was not investigated (Apolikhin et al., 2016).

In a study performed in Thai population Chou et al. studied the effect of 5 polymorphisms of the *ORAI1* gene (rs12313273, rs6486795, rs7135617, rs12320939, and rs712853) on the risk of the calcium urolithiasis development. As a result of their investigation, the higher risk of KSD development was established for the rs12313273 and rs6486795 polymorphisms carriers. For the C allele of the rs12313273 polymorphism, the odds ratio turned out to be the most significant (OR = 2.10). At the same time, the maximum risk of the nephrolithiasis development was demonstrated for the combination of C/T/C alleles at the rs12313273/rs7135617/rs6486795 polymorphic loci (OR = 2.54) (Chou et al., 2011).

In the current study all three mentioned above polymorphisms (rs12313273, rs6486795, and rs7135617) of the *ORAI1* gene were tested. Our results suggest an association of the C allele of the rs6486795 locus (OR = 2.25) with KSD development. The difference in the frequency of the alternative C allele of the rs12313273 locus between the experimental and control groups was pronounced, but did not reach a statistically significant level (25% versus 15%, $\chi 2 = 3.125$, $p = 0.078$). This is possibly due to the size limitation of the studied groups. When applying a comprehensive assessment of the cumulative effect of the rs12313273, rs6486795, and rs7135617 polymorphisms of the *ORAI1* gene on the risk of urolithiasis development, no significant data for their cumulative effect were obtained.

Thus, the data presented in the current study are suggestive for an association between the rs6486795 polymorphism of the *ORAI1* gene and the risk of calcium urolithiasis development in Russia. The results of our investigation do not contradict the data obtained by the above mentioned authors.

The analysis of the dominant and recessive inheritance models of the polymorphisms *CASR* (rs1042636), *CALCR* (rs1801197) and *ORAI1* (rs6486795) genes is important for assessing the risk of calcium urolithiasis development, and therefore it is important for the prevention of KSD. The recessive model for the *CASR* (rs1042636) and *CALCR* (rs1801197) polymorphisms, which was confirmed for this loci in the current study, allows us to predict a higher risk of urolithiasis development in patients homozygous for the risk alleles of these genes (rs1042636A in *CASR* and rs1801197A in *CALCR*).

Studying of the gene-gene interactions and investigating of the complex impact of gene polymorphisms are not less important for determining of the KSD risk development. In our study, the relationship between the loci rs1042636 of the *CASR* gene and rs1801197 of the *CALCR* gene was shown. This phenomenon requires, further, investigation.

Analysis of the association between the rs219780 polymorphism of the *CLDN14* gene and calcium urolithiasis in Russian population showed borderline *p*-value. Further, study of this association is needed to confirm the effect of rs219780 on the risk of KSD development.

CONCLUSION

Thus, we showed the strong association between polymorphism rs1801197 of the *CALCR* gene and the risk of calcium urolithiasis development in Russian population. Further, investigation of the risk loci is necessary in order to assess molecular pathogenesis of calcium urolithiasis and will help to identify additional genetic factors of KSD development for better diagnostics of this complex disease.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Committee of Sechenov University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

ML and TF: development of the concept and design of the study, collection of the samples, counseling of the patients, data analysis, statistics analysis, supervision, writing the text, and approval of the final version of the article. DS: collecting the samples for the study. KK, AS, AMa, and DK: molecular genetic testing and bioinformatic analysis. VK: statistics analysis and visualization. SN: participation in the article text preparation and analysis of the genetic results. AMo: review and editing. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.621049/full#supplementary-material

REFERENCES

Worcester, E. M., and Coe, F. L. (2010). Clinical practice. Calcium kidney stones. *N Engl J Med.* 363, 954–963. doi: 10.1056/NEJMcp1001011

Filippova, T. V., Khafizov, KF., Rudenko, V. I., Rapoport, L. M., Tsarichenko, D. G., Enikeev, D. V., et al. (2020). Genetic factors of polygenic urolithiasis. *Urologia Journal* 2, 1–8. doi: 10.1177/0391560319898375

Apolikhin, O. I., Sivkov, A. V., Konstantinova, O. V., Slominsky, P. A., Tupitsyna, T. V., and Kalinichenko, D. N. (2016). Genetic risk factors for recurrence-free urolithiasis in the russian population. *Urologiia* 4, 20–23. Russian,

Apolikhin, O. I., Sivkov, A. V., Konstantinova, O. V., Slominskii, P. A., Tupitsyna, T. V., and Kalinichenko, D. N. (2017). Early diagnosis of risk for developing calcium oxalate urolithiasis. *Urologiia* 3, 5–8. doi: 10.18565/urol.2017.3.5-8 Russian,

Apolihin, O. I., Sivkov, A. V., Konstantinova, O. V., Slominskij, P. A., Tupicyna, T. V., and Kalinichenko, D. N. (2015). GENETIC RISK FACTORS FOR MULTIPLE KIDNEY STONE FORMATION IN THE RUSSIAN POPULATION. *Urologiia* 4, 4–6. Russian,

Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research.* 20, 1297–1303. doi: 10.1101/gr.107524.110

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

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., et al. (2016). The Ensembl variant effect predictor. *Genome Biol.* 17, 122. doi: 10.1186/s13059-016-0974-4

Milne, I., Stephen, G., Bayer, M., Cock, P. J., Pritchard, L., Cardle, L., et al. (2013). Using tablet for visual exploration of second-generation sequencing data. *Brief Bioinform* 14, 193–202. doi: 10.1093/bib/bbs012

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Steiß, V., Letschert, T., Schäfer, H., and Pahl, R. (2012). PERMORY-MPI: a program for high-speed parallel permutation testing in genome-wide association studies. *Bioinformatics.* 28, 1168–1169. doi: 10.1093/bioinformatics/bts086

R Core Team. (2019). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Lancaster, H. O. (1961). Significance Tests in Discrete Distributions. *Journal of the American Statistical Association* 56, 223–234.

Graffelman, J. (2015). Exploring diallelelic genetic markers: the HardyWeinberg package. *Journal of Statistical Software* 64, 1–23.

Tomas, J. (2017). *Aragon: epitools: Epidemiology Tools. R package version 0.5-9.* https://CRAN.R-project.org/package=epitools***.

Hahn, L. W., Ritchie, M. D., and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics.* 19, 376–382. doi: 10.1093/bioinformatics/btf869

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet.* 69, 138–147. doi: 10.1086/321276

Cho, Y. M., Ritchie, M. D., Moore, J. H., Park, J. Y., Lee, K. U., Shin, H. D., et al. (2004). Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia.* 47, 549–554. doi: 10.1007/s00125-003-1321-3

Brassat, D., Motsinger, A. A., Caillier, S. J., Erlich, H. A., Walker, K., Steiner, L. L., et al. (2006). Multifactor dimensionality reduction reveals gene-gene interactions associated with multiple sclerosis susceptibility in African Americans. *Genes Immun.* 7, 310–315. doi: 10.1038/sj.gene.6364299

Andrew, A. S., Nelson, H. H., Kelsey, K. T., Moore, J. H., Meng, A. C., Casella, D. P., et al. (2006). Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility. *Carcinogenesis.* 27, 1030–1037. doi: 10.1093/carcin/bgi284

Vezzoli, G., Terranegra, A., Aloia, A., Arcidiacono, T., Milanesi, L., Mosca, E., et al. (2013). Decreased transcriptional activity of calcium-sensing receptor gene promoter 1 is associated with calcium nephrolithiasis. *J Clin Endocrinol Metab.* 98, 3839–3847. doi: 10.1210/jc.2013-1834

Shakhssalim, N., Basiri, A., Houshmand, M., Pakmanesh, H., Golestan, B., Azadvari, M., et al. (2014). Genetic polymorphisms in calcitonin receptor gene and risk for recurrent kidney calcium stone disease. *Urol Int.* 92, 356–362. doi: 10.1159/000353348

Chou, Y. H., Juo, S. H., Chiu, Y. C., Liu, M. E., Chen, W. C., Chang, C. C., et al. (2011). A polymorphism of the ORAI1 gene is associated with the risk and recurrence of calcium nephrolithiasis. *J Urol.* 185, 1742–1746. doi: 10.1016/j.juro.2010.12.094

Qin, J., Cai, Z., Xing, J., Duan, B., and Bai, P. (2019). Association between calcitonin receptor gene polymorphisms and calcium stone urolithiasis: A meta-analysis. *Int Braz J Urol.* 45, 901–909. doi: 10.1590/S1677-5538.IBJU.2019.0061

Vezzoli, G., Scillitani, A., Corbetta, S., Terranegra, A., Dogliotti, E., Guarnieri, V., et al. (2011). Polymorphisms at the regulatory regions of the CASR gene influence stone risk in primary hyperparathyroidism. *Eur J Endocrinol.* 164, 421–427. doi: 10.1530/EJE-10-0915

Corbetta, S., Eller-Vainicher, C., Filopanti, M., Saeli, P., Vezzoli, G., Arcidiacono, T., et al. (2006). R990G polymorphism of the calcium-sensing receptor and renal calcium excretion in patients with primary hyperparathyroidism. *Eur J Endocrinol.* 155, 687–692. doi: 10.1530/eje.1.02286

Scillitani, A., Guarnieri, V., Battista, C., De Geronimo, S., Muscarella, L. A., Chiodini, I., et al. (2007). Primary hyperparathyroidism and the presence of kidney stones are associated with different haplotypes of the calcium-sensing receptor. *J Clin Endocrinol Metab.* 92, 277–283. doi: 10.1210/jc.2006-0857

Vezzoli, G., Scillitani, A., Corbetta, S., Terranegra, A., Dogliotti, E., Guarnieri, V., et al. (2015). Risk of nephrolithiasis in primary hyperparathyroidism is associated with two polymorphisms of the calcium-sensing receptor gene. *J Nephrol.* 28, 67–72. doi: 10.1007/s40620-014-0106-8

Guha, M., Bankura, B., Ghosh, S., Pattanayak, A. K., Ghosh, S., Pal, D. K., et al. (2015). Polymorphisms in CaSR and CLDN14 Genes Associated with Increased Risk of Kidney Stone Disease in Patients from the Eastern Part of India. *PLoS One.* 10:e0130790. doi: 10.1371/journal.pone.0130790

Shakhssalim, N., Kazemi, B., Basiri, A., Houshmand, M., Pakmanesh, H., Golestan, B., et al. (2010). Association between calcium-sensing receptor gene polymorphisms and recurrent calcium kidney stone disease: a comprehensive gene analysis. *Scand J Urol Nephrol.* 44, 406–412. doi: 10.3109/00365599.2010.497770

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.

Citation: Litvinova MM, Khafizov K, Korchagin VI, Speranskaya AS, Asanov AY, Matsvay AD, Kiselev DA, Svetlichnaya DV, Nuralieva SZ, Moskalev AA and Filippova TV (2021) Association of CASR, CALCR, and ORAI1 Genes Polymorphisms With the Calcium Urolithiasis Development in Russian Population. Front. Genet. 12:621049. doi: 10.3389/fgene.2021.621049

frontiers in Genetics

ORIGINAL RESEARCH
published: 17 May 2021
doi: 10.3389/fgene.2021.632359

Construction of a circRNA-miRNA-mRNA Regulatory Network Reveals Potential Mechanism and Treatment Options for Osteosarcoma

Yi He[1†], Haiting Zhou[2†], Wei Wang[1], Haoran Xu[1] and Hao Cheng[1]*

[1] Department of Orthopedics, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2] Department of Oncology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

OPEN ACCESS

Edited by:
Anastasia Anashkina, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences (RAS), Russia

Reviewed by:
Fuhai Li, Washington University in St. Louis, United States
Yuriy L. Orlov, I.M. Sechenov First Moscow State Medical University, Russia

****Correspondence:***
Hao Cheng
chenghao@tjh.tjmu.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: *23 November 2020*
Accepted: *20 April 2021*
Published: *17 May 2021*

Citation:
He Y, Zhou H, Wang W, Xu H and Cheng H (2021) Construction of a circRNA-miRNA-mRNA Regulatory Network Reveals Potential Mechanism and Treatment Options for Osteosarcoma. Front. Genet. 12:632359. doi: 10.3389/fgene.2021.632359

Background: Osteosarcoma is a common malignant primary bone tumor in adolescents and children. Numerous studies have shown that circRNAs were involved in the proliferation and invasion of various tumors. However, the role of circRNAs in osteosarcoma remains unclear. Here, we aimed to explore the regulatory network among circRNA-miRNA-mRNA in osteosarcoma.

Methods: The circRNA (GSE140256), microRNA (GSE28423), and mRNA (GSE99671) expression profiles of osteosarcoma were collected from the Gene Expression Omnibus (GEO) database. Differentially expressed circRNAs, miRNAs and mRNAs were identified. CircRNA-miRNA interactions and miRNA-mRNA interactions were determined by Circular RNA Interactome (CircInteractome) database and microRNA Data Integration Portal (mirDIP) database, respectively. Then, we constructed a regulatory network. Function enrichment analysis of miRNA and mRNA was performed by DIANA-miRPath v3.0 and Metascape database, respectively. mRNAs with significant prognostic value were identified based on expression profiles from The Cancer Genome Atlas (TCGA) database, and we constructed a subnetwork for them. To make the most of the network, we used the CLUE database to predict potential drugs for the treatment of osteosarcoma based on mRNA expression in the network. And we used the STITCH database to analyze and validate the interactions among these drugs and mRNAs, and to further screen for potential drugs.

Results: A total of 9 circRNAs, 19 miRNAs, 67 mRNAs, 54 pairs of circRNA-miRNA interactions and 110 pairs of miRNA-mRNA interactions were identified. A circRNA-miRNA-mRNA network was constructed. Function enrichment analysis indicated that these miRNAs and mRNAs in the network were involved in the process of tumorigenesis and immune response. Among these mRNAs, STC2 and RASGRP2 with significantly prognostic value were identified, and we constructed a subnetwork for them. Based on mRNA expression in the network, three potential drugs, quinacridine, thalidomide

and zonisamide, were screened for the treatment of osteosarcoma. Among them, quinacridine and thalidomide have been proved to have anti-tumor effects in previous studies, while zonisamide has not been reported. And a corresponding drug-protein interaction network was constructed.

Conclusion: Overall, we constructed a circRNA-miRNA-mRNA regulatory network to investigate the possible mechanism in osteosarcoma, and predicted that quinacridine, thalidomide and zonisamide could be potential drugs for the treatment of osteosarcoma.

Keywords: circRNA, osteosarcoma, network, GEO, TCGA

INTRODUCTION

Osteosarcoma is a common malignant primary bone tumor in adolescents and children (Luetke et al., 2014). This tumor is most likely to happen in the metaphysis regions of long bones (Cortini et al., 2017). Medical advances have significantly improved the survival rate for osteosarcoma. However, early screening and diagnosis of osteosarcoma are arduous due to the lack of adequate diagnostic markers (Li and Wang, 2020). Metastasis and drug resistance also worsened the prognosis (Mirabello et al., 2009; Harrison et al., 2018). It is urgent to figure out the underlying pathogenesis of osteosarcoma, and to discover potential targets for earlier diagnosis and potential drugs for better treatment.

CircRNA is a new class of endogenous and regulatory non-coding RNA with a covalent closed-loop structure. CircRNAs were discovered as early as the 1990s (Nigro et al., 1991). However, due to limitations in knowledge and technology at that time, they were not be fully investigated and thought to be less abundant *in vivo* due to splicing errors (Jeck and Sharpless, 2014). With the development of high-throughput sequencing technology, thousands of circRNAs have been recognized (Zhang et al., 2014). Numerous studies have shown that circRNAs were involved in tumor proliferation and invasion, and could become molecular markers of various tumors (Cheng et al., 2019; Liu et al., 2019; Li et al., 2020).

In the present study, we collected the expression profiles of circRNAs (GSE140256), miRNAs (GSE28423), and mRNAs (GSE99671) of osteosarcoma from GEO database. We also collected another mRNA expression profiles with survival information from TCGA database. The process flow chart is shown in **Figure 1**. Using expression profiles from GEO to perform differential expression analysis and targets prediction, we determined the circRNA-miRNA interactions and the miRNA-mRNA interactions, and finally constructed a circRNA-miRNA-mRNA regulatory network. Using expression profiles from the TCGA database, we identified mRNAs with significant prognostic value. Furthermore, we performed functional enrichment analysis to reveal the potential mechanism of osteosarcoma. Particularly, we predicted potential drugs for the treatment of osteosarcoma, which may provide new insight into osteosarcoma treatment.

Abbreviations: TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; CircInteractome, Circular RNA Interactome; MirDIP, MicroRNA Data Integration Portal; RBP, RNA-binding protein; GO, Gene Ontology; CeRNA, Competitive endogenous RNA; ROS, Reactive oxygen species; STC2, Stanniocalcin 2; RASGRP2, Ras guanyl nucleotide releasing protein 2.

MATERIALS AND METHODS

Data Sets

Gene Expression Omnibus (GEO)[1] is a public functional genomics data repository that helps researchers query and download experiments and curated gene expression profiles. Three datasets (GSE140256, GSE28423, and GSE99671) from the GEO database were collected in our study. GSE140256 is a microarray chip dataset of circRNAs containing 3 primary osteosarcoma tissues and 3 adjacent tissues, and we used this dataset to screen for differentially expressed circRNAs. GSE28423 is a microarray chip dataset of miRNAs containing samples from 19 human osteosarcoma cell lines and 4 normal bones, which was used to screen for differentially expressed miRNAs. GSE99671 is a dataset of expression profiling by high throughput sequencing, containing 18 osteosarcoma samples and 18 corresponding normal bone samples, which was used to screen for differentially expressed mRNAs.

The Cancer Genome Atlas (TCGA)[2] is a landmark cancer genomics program demonstrating the genomic alterations associated with 33 cancer types. 88 osteosarcoma samples were obtained from the database. After data processing, 3 samples were removed due to incomplete survival information or follow-up time less than 30 days (Subramanian et al., 2005). Finally, 85 samples from TCGA with complete clinical information were included in our study for survival analysis.

The Acquisition of Differential Expression of circRNAs, miRNAs, and mRNAs

The expression profiles of GSE140256 and GSE28423 were normalized by limma package, and that of GSE99671 was normalized with TMM (Trimmed Mean of M values) methods by edgeR package (Robinson et al., 2010). Then differentially expressed circRNAs, miRNAs and mRNAs were obtained from GSE140256, GSE28423, and GSE99671, respectively. Differentially expressed circRNAs, miRNAs, mRNAs were defined as adjusted $P < 0.05$ and $|\log2 \text{ fold-change (FC)}| > 1.5$ (Wang et al., 2019).

[1]https://www.ncbi.nlm.nih.gov/geo/

[2]https://portal.gdc.cancer.gov/

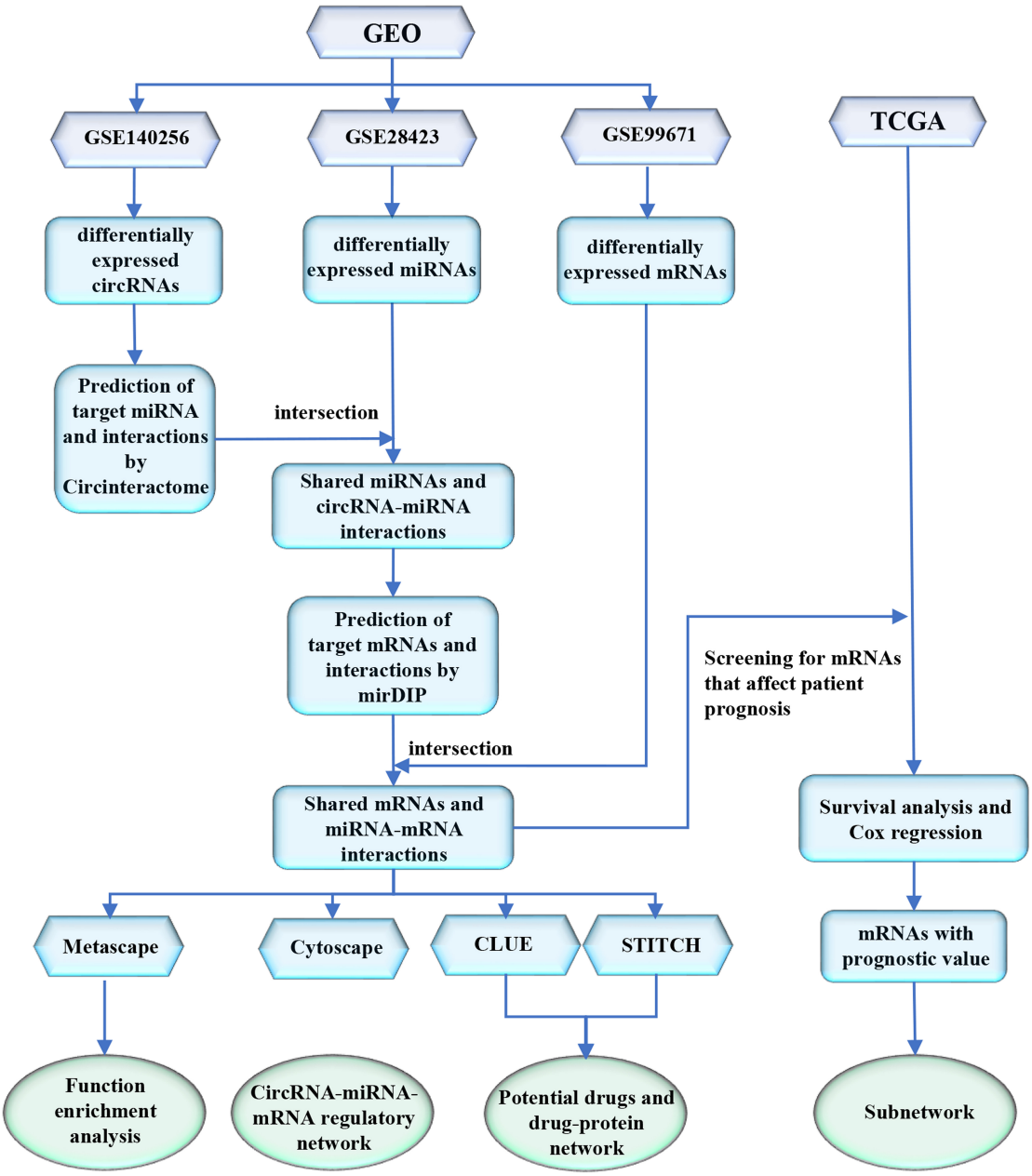


FIGURE 1 | A flow chart of the analytical process.

Prediction of circRNA Targeted miRNAs

Circular RNA Interactome (CircInteractome)[3] is an online tool based on 109 datasets of RNA-binding proteins (RBPs) and queries circRNAs for RNA-binding sites, which enables to predict RNA-binding proteins and miRNAs on circRNAs (Dudekula et al., 2016). CircRNA-targeted miRNAs were predicted by the database, from which we also acquired the interactions among circRNAs and miRNAs. Then these predicted miRNAs were intersected with differentially expressed miRNAs from GSE28423 dataset to obtained shared miRNAs and their interactions with circRNAs.

Prediction of miRNA Targeted mRNAs

MicroRNA Data Integration Portal (mirDIP)[4] is a web tool to predict miRNA-mRNA interactions which integrated information on human miRNA targeted mRNAs from 30 databases (Tokar et al., 2018). We uploaded the shared miRNAs obtained in the previous step to the database to acquire the

[3]https://circinteractome.irp.nia.nih.gov/

[4]http://ophid.utoronto.ca/mirDIP/index.jsp

predicted mRNAs and miRNA-mRNA interactions. In order to make the prediction more accurate, the score class for the confidence was set at the top 1%. After that, these predicted mRNAs were selected to intersect with the differentially expressed mRNAs from GSE99671 dataset to obtained shared mRNAs and their interactions with miRNAs.

Functional Enrichment Analysis

DIANA-miRPath v3.0[5] is an online-server providing evaluation of the regulatory role of miRNAs (Vlachos et al., 2015). We used the database to analyze the function of these miRNAs. The Metascape database[6] is a powerful gene function annotation analysis tool (Zhou et al., 2019). Gene enrichment analysis was performed on Metascape database. The min overlap was set at 3, which required enriched terms to include ≥ 3 candidates. The min enrichment was set at 1.5, which indicated at least 1.5 times more given pathway members are found in uploaded gene list compared to what would have been expected by chance. The *P*-value was set at < 0.01.

Survival Analysis in TCGA Profiles

Since the GEO datasets did not contain survival information, we estimated the prognostic value of shared mRNAs using Kaplan-Meier method and Cox regression analysis based on expression profiles from TCGA dataset. The expression profiles of osteosarcoma from TCGA database were transformed into TPM format for normalization. The threshold of statistical significance of log-rank test was set at *P*-value < 0.05.

Construction of circRNA-miRNA-mRNA Network

In the previous steps, we obtained differentially expressed circRNAs, shared miRNAs, shared mRNAs and their interactions. Then we used their interactions to construct a circRNA-miRNA-mRNA regulatory network and we visualized it by using Cytoscape software (Version 3.8.0).

Prediction of Potential Drugs and the Construction of Drug-Protein Interaction Network

To make the most of the established network, we performed prediction of potential therapeutic drugs. The Connectivity Map Linked User Environment (CLUE)[7] is the world's largest perturbation-driven gene expression dataset (Subramanian et al., 2017). We used it to identify compounds whose administration to cancer cells resulted in an opposite expression profile of these mRNAs in the network. The database would calculate a score (−100∼100) for each compound within database, called connectivity score, based on the expression of queried mRNAs. In particular, a negative score would indicate that the compound was antagonistic to the expression of queried mRNAs, that is, the expression of queried mRNAs was reduced by treatment of the compound. Since the mRNAs we queried were associated with tumor development, reversing their expression by the drugs might inhibit tumor growth. We upload 67 mRNAs from the network to the CLUE database and collated the results to filter out compounds with the connectivity score < -85 (Chen Y.T. et al., 2019).

Search Tool for Interacting Chemicals (STITCH)[8] is a database for predicting interactions between chemical substances and proteins (Szklarczyk et al., 2016). To further screen the compounds, we looked for compounds that interacted with the proteins, which were coded by mRNAs from our network, by using STITCH database. We set the interaction score > 0.7, which indicated a high confidence. Predicted functional partners < 5, which indicated less than 5 other predicted proteins would be involved in the drug and protein interactions. The results were visualized by a drug-protein interaction network by using Cytoscape software.

RESULTS

The Acquisition of Differential Expression of circRNAs, miRNAs, and mRNAs

The basic information of three microarray datasets (GSE140256, GSE28423, and GSE99671) was listed in **Table 1**. In GSE140256 dataset, 9 differentially expressed circRNAs were screened out. The basic information for 9 circRNAs was showed in **Table 2**. Among them, hsa_circ_0000253, hsa_circ_0010220, and hsa_circ_0020378 were up-regulated and might be tumor promotors in osteosarcoma. While hsa_circ_0049271, has_circ_0000006, hsa_circ_0078767, hsa_circ_0046264, hsa_circ_0094088, and hsa_circ_0096041 were down-regulated and might be tumor inhibitors in osteosarcoma. In GSE28423 dataset and GSE99671 dataset, we screened out 68 differentially expressed miRNAs and 346 differentially expressed mRNAs, respectively. The heatmaps of differentially expressed circRNAs, miRNAs and mRNAs were shown in **Figure 2**.

Identification of circRNA–miRNA Interactions

Using CircInteractome database, we identified 313 targeted miRNAs for these 9 circRNAs, and there were 951 pairs of circRNA-miRNA interactions among them. Using these 313 predicted miRNAs to intersect with the 68 differentially expressed

[5]http://snf-515788.vm.okeanos.grnet.gr/

[6]https://metascape.org/gp/index.html

[7]https://clue.io/

[8]http://stitch.embl.de/

TABLE 1 | Basic information of the three microarray datasets from GEO and an RNAseq dataset from TCGA.

RNA	Dataset	Platform	Sample size (Normal/Tumor)
circRNA	GSE140256	GPL27741	3/3
miRNA	GSE28423	GPL8227	4/19
mRNA	GSE99671	GPL20148	18/18
mRNA	TCGA	–	0/85

TABLE 2 | Basic characteristics of nine circRNAs.

CircRNA	Position	Strand	Genomic length	Best transcript	Gene symbol	Regulation
hsa_circ_0000253	chr10:97999787-97999925	−	138	NM_013314	BLNK	Up
hsa_circ_0010220	chr1:17907047-18024370	+	117323	NM_018125	ARHGEF10L	Up
hsa_circ_0049271	chr19:10610070-10610756	−	686	NM_203500	KEAP1	Down
hsa_circ_0000006	chr1:1601102-1666274	−	65172	NM_001110781	SLC35E2B	Down
hsa_circ_0078767	chr6:170615843-170639638	+	23795	NM_032448	FAM120B	Down
hsa_circ_0046264	chr17:79813017-79817263	−	4246	NM_000918	P4HB	Down
hsa_circ_0094088	chr10:7318853-7407477	−	88624	NM_001029880	SFMBT2	Down
hsa_circ_0096041	chr11:61615630-61616333	+	703	ENST00000278840.4	FADS2	Down
hsa_circ_0020378	chr10:128594022-128926028	+	332006	NM_001380	DOCK1	Up

miRNAs acquired from GSE28423 dataset, we finally obtained 19 shared miRNAs. One circRNA had no corresponding targeted miRNA after intersections. Thus, we obtained 54 pairs of interactions among 8 circRNAs and 19 miRNAs.

Identification of miRNA–mRNA Interactions

We uploaded the 19 shared miRNAs obtained in the previous step to mirDIP database (Tokar et al., 2018), and the score class for the confidence was set at the top 1% for more accurate prediction. Then we acquired 6827 mRNAs targeted by these 19 miRNAs. And there were 13431 pairs of interactions among them. These 6827 mRNAs were then intersected with the differentially expressed mRNAs obtained from the differential analysis of GSE99671 dataset, and finally we identified 67 shared mRNAs and 110 pairs of miRNA-mRNA interactions.

Function Enrichment Analysis

Using DIANA-miRPath database, we found that these shared miRNAs were involved in TGF-beta signaling pathway, proteoglycans in cancer, epidermal growth factor receptor signaling pathway, fibroblast growth factor receptor signaling pathway and other pathways related to tumorigenesis, as shown in **Figure 3**.

Using Metascape database, we found that these shared mRNAs were predominantly enriched in cancer and immune-related functional activities and pathways (**Figure 4**). For Gene Ontology (GO) terms, mRNAs were enriched in extracellular structure organization, regulation of leukocyte apoptotic process, regulation of lymphocyte chemotaxis, negative regulation of cellular component organization, angiogenesis, myeloid leukocyte activation, homeostasis of number of cells and gliogenesis. For Canonical Pathways, mRNAs were enriched in TAP63 pathway, MYC repress pathway and CMYB pathway. For Reactome Gene Sets, mRNAs were enriched in Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs).

Construction of a circRNA-miRNA-mRNA Network

To present the relationship among circRNAs, miRNAs, and mRNAs, we constructed a circRNA-miRNA-mRNA regulatory network based on the interactions among these transcripts and visualized it using Cytoscape software. The result was showed in **Figure 5**.

Survival Analysis in TCGA Profiles

To explore the prognostic value of 67 shared mRNAs in the network, we performed Kaplan-Meier survival analysis and Cox regression analysis based on profiles from TCGA database. Ultimately, two mRNAs (STC2 and RASGRP2) with significantly prognostic value were screened out, as shown in **Table 3** and **Figure 6**. High expression of STC2 and RASGRP2 were correlated with poor prognosis of osteosarcoma. Stanniocalcin 2 (STC2) and Ras guanyl nucleotide releasing protein 2 (RASGRP2) were reported to be involved in many human malignancies (Tamura et al., 2009; Yokobori et al., 2010; Mele et al., 2018; Zhang et al., 2019). In order to show the regulatory mechanisms of these two mRNAs more clearly, we established a subnetwork contained these two mRNAs and associated miRNAs (hsa-miR-940 and hsa-miR-223) and circRNAs (hsa_circ_0010220, hsa_circ_0000006, hsa_circ_0078767, hsa_circ_0046264, hsa_circ_0094088, and hsa_circ_0020378) (**Figure 7**).

Identification of Potential Drugs and Construction of Drug-Protein Interaction Network

We uploaded 67 shared mRNAs to CLUE database. Based on the criteria of connectivity score < -85, we obtained 26 kinds of candidate compounds (drugs).

In order to further screen the compounds, we analyzed the interaction relationships among these candidate compounds and mRNAs. We uploaded the 26 candidate compounds and 67 mRNAs to STITCH database, and constructed a drug-protein network to visualize their interactions (**Figure 8**). Finally, only three compounds, quinacridine, thalidomide and zonisamide, were screened out. The information of these drugs was shown in **Table 4**.

DISCUSSION

CircRNAs can act as miRNA sponges and bind to miRNA, relieving the inhibitory effect of miRNAs on their target

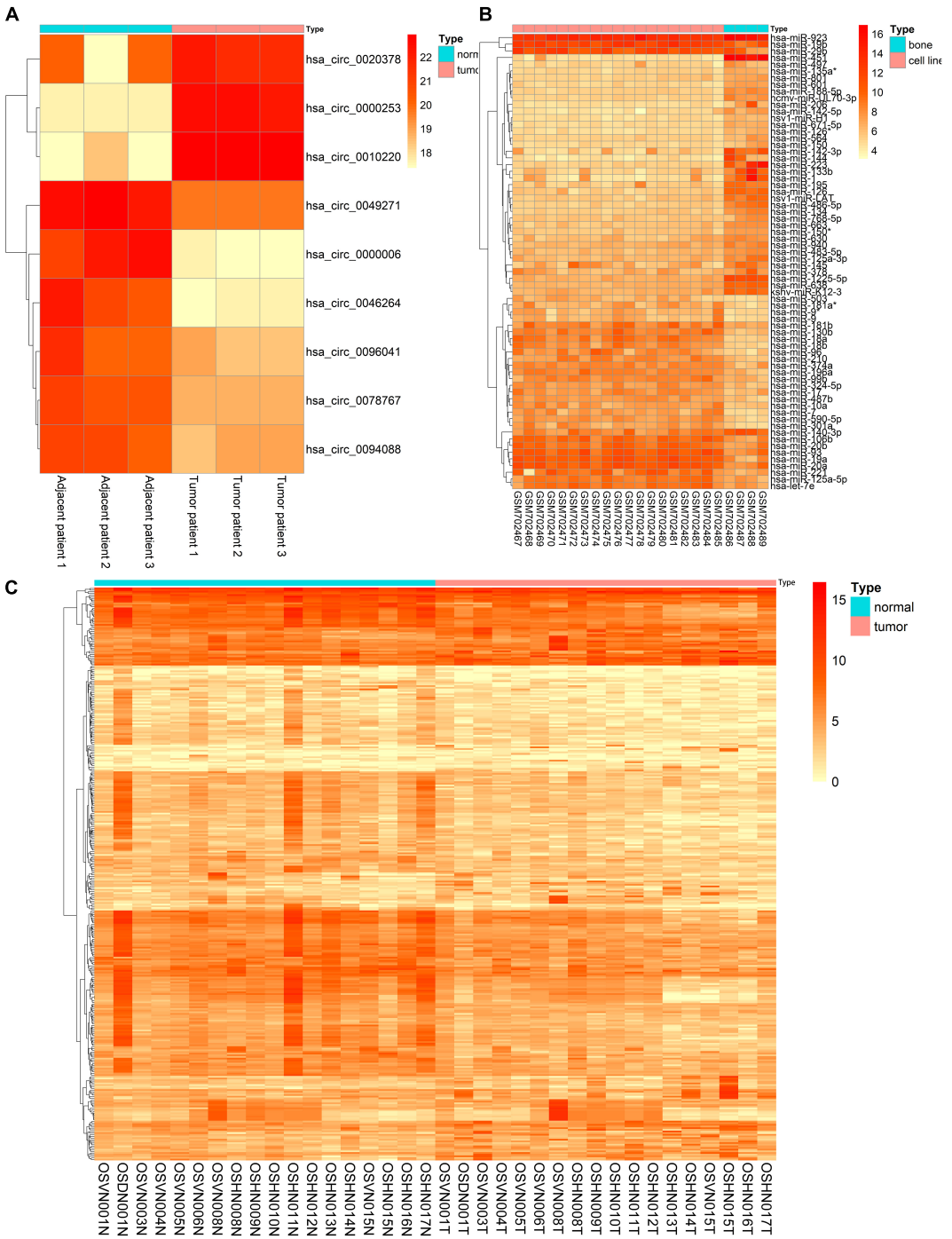


FIGURE 2 | The heatmaps of differentially expressed circRNAs **(A)**, miRNAs **(B)**, and mRNAs **(C)**.

mRNAs and regulating the expression of target mRNAs, which is known as the competitive endogenous RNA (ceRNA) mechanism (Qi et al., 2015). By interacting with cancer-associated miRNAs, circRNAs play an important regulatory role in cancer (Meng et al., 2017).

To investigate the potential role of circRNAs in osteosarcoma, we constructed a circRNA-miRNA-mRNA regulatory network using bioinformatics predictions combined with differential expression data. This network contained the interactions of 8 circRNAs, 19 miRNAs, and 67 mRNAs. Among these circRNAs, hsa_circ_0078767 was reported to suppress non-small-cell lung cancer by modulating RASSF1A expression via sponging miR-330-3p (Chen T. et al., 2019). While hsa_circ_0046264 was reported to have different effects in

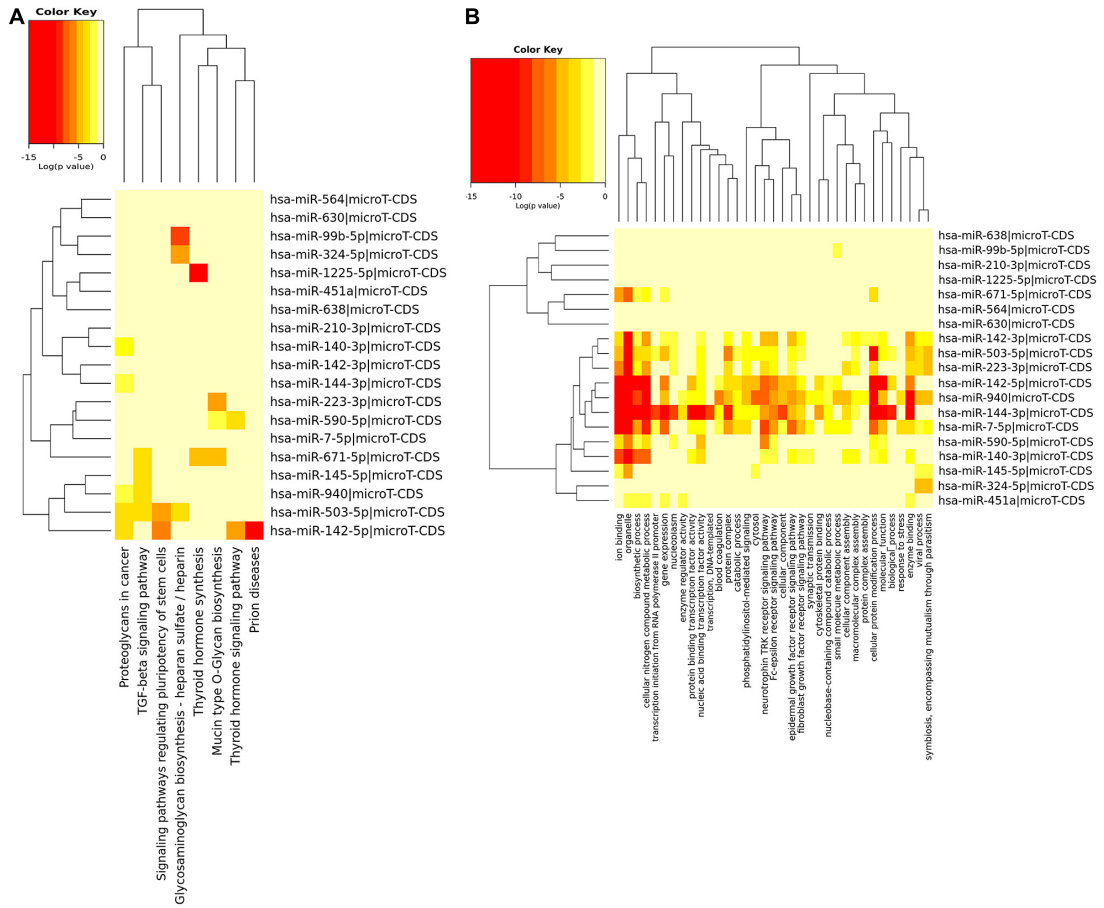


FIGURE 3 | KEGG pathway enrichment analysis **(A)** and GO enrichment analysis **(B)** of 19 shared miRNAs based on DIANA-miRPath.

different tumors. Hsa_circ_0046264 could up-regulate BRCA2 by targeting miR-1245 to induce apoptosis and inhibit cell proliferation and invasion in lung cancer cells (Yang et al., 2018). However, Liu et al. proposed the opposite conclusion that hsa_circ_0046264 was remarkably up-regulated in lung cancer patients, enhancing tumor growth, invasion, metastasis, and chemotherapy resistance (Liu et al., 2020). Hsa_circ_0049271 was down-regulated in non-small cell lung cancer, which was similar to the down-regulation we found in osteosarcoma (Li et al., 2021). The other 6 circRNAs had not been studied yet, which need further investigation, and they have the potential to be reliable biomarkers and therapeutic targets.

Function enrichment analysis showed that these miRNAs and mRNAs were mainly involved in immune and inflammatory response, angiogenesis, and some common biological processes involved in tumorigenesis. Among them, MYC mediated transcriptional amplification through super-enhancers is an essential hallmark of cancer (Kress et al., 2015). MYC was demonstrated to be related to progression and poor prognosis in osteosarcoma. Besides, MYC could be suppressed by super-enhancer inhibitors to effectively inhibit growth, migration, and invasion of osteosarcoma cells (Chen et al., 2018). Therefore, targeting MYC/super-enhancer axis represents a promising treatment strategy for patients with osteosarcoma. According to animal model systems, bone tumors' growth and metastasis depend on new blood vessel development, which is also known as angiogenesis (Gorlick et al., 2003). Vascular endothelial growth factor (VEGF) is a pivotal tumor-derived angiogenic factor with multiple biological functions, which constitute the most vital signaling pathways in tumor angiogenesis (Quan and Choong, 2006). The expression of VEGF has been considered as an important prognostic biomarker evaluating the angiogenesis in osteosarcoma (DuBois and Demetri, 2007). P63 is a member of the p53 family, which encodes the isoforms TAp63 and ΔNp63 (Ram Kumar et al., 2014). TAp63 functions as a tumor suppressor by regulating senescence through p53-independent pathways. The ability of TAp63 to trigger senescence and inhibit tumorigenesis without regard to p53 status, which identified TAp63 as a promising target of anti-cancer treatment for malignancies with compromised p53 (Guo et al., 2009).

Among these 67 mRNAs, we found that two mRNAs (STC2 and RASGRP2) were significantly related to overall survival. STC2 and RASGRP2 were associated with tumor development. It was reported that STC2 could promote head and neck squamous cell carcinoma metastasis by modulating the PI3K/AKT/snail signaling pathway (Yang et al., 2017). High expression of

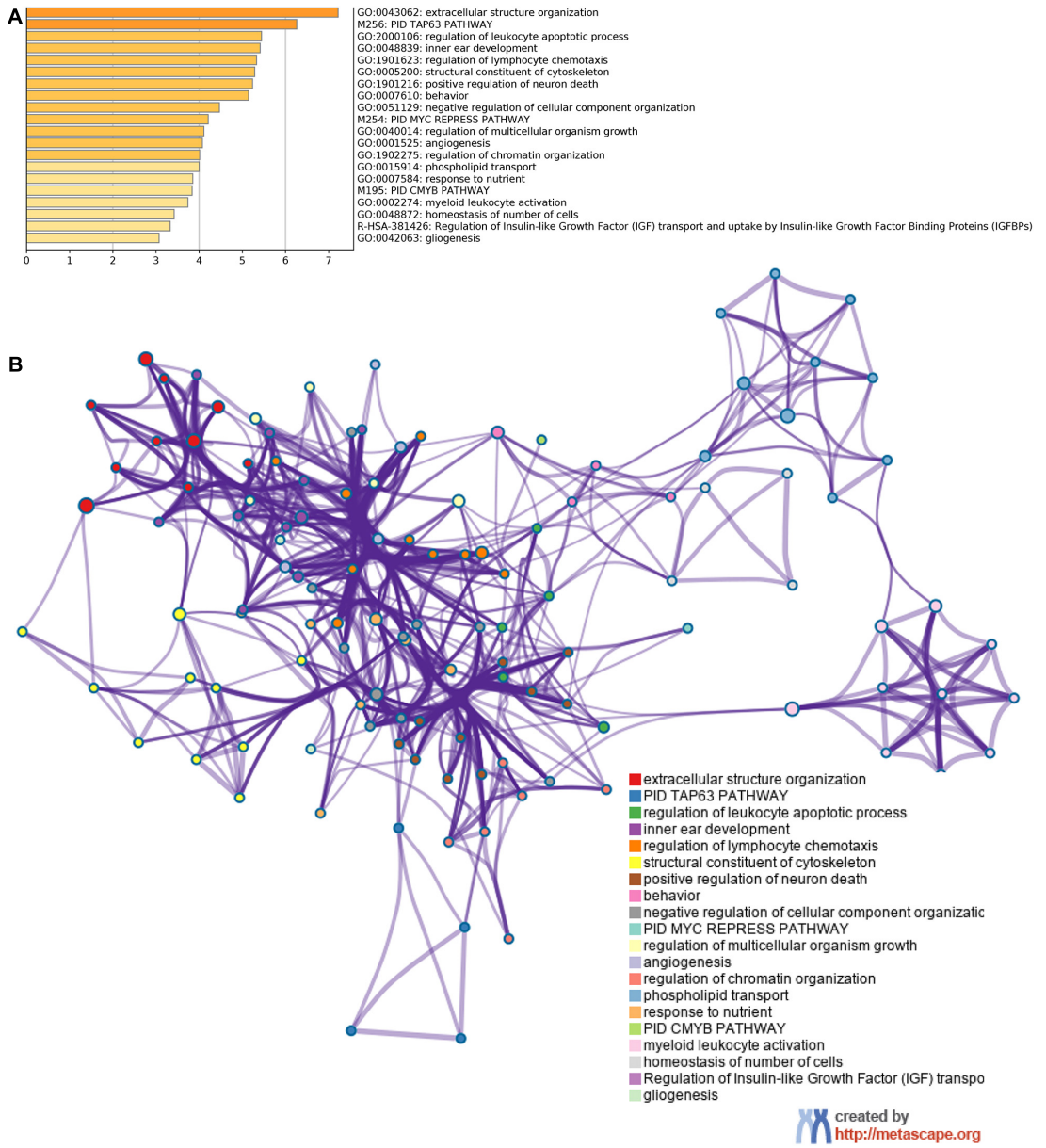


FIGURE 4 | Function enrichment analysis of 67 shared mRNAs. **(A)** The bar graph of top clusters with their representative enriched terms and **(B)** a network of enriched terms, colored by cluster-ID, where nodes that share the same cluster-ID are typically close to each other.

STC2 could also enhance the ability of migration and invasion for nasopharyngeal cancer cells after radiotherapy (He et al., 2019). In osteosarcoma, STC2 was reported to promote the proliferation, invasion and migration of osteosarcoma cells by enhancing the glycolysis (Yu et al., 2021). RASGRP2 is a guanine nucleotide exchange factor, which is well known to target to Rap1 mainly (Stone, 2011). RASGRP2 could inhibit apoptosis by activating Rap1 to inhibit tumor necrosis factor (TNF)-induced reactive oxygen species (ROS) production (Sato et al., 2019). This inhibition of apoptosis is likely to play an important role in tumor development as well. It has also been reported that calcium-sensitive RASGRP2 could promote chronic lymphocytic leukemia cell metastasis through activation of Rap1 (Mele et al., 2018). In addition, it was reported that African American enriched splice variants of PIK3CD, FGFR3, TSC2 and RASGRP2 have greater oncogenic potential compared to the corresponding European American expression variants, suggesting the oncogenic effect of RASGRP2 (Wang et al., 2017). Many studies supported the anti-tumor effects of hsa-miR-940. It was reported that hsa-miR-940 could inhibit the growth of hepatocellular carcinoma by targeting SPOCK1 (Li et al., 2019). MiR-940 was also reported to inhibit the progression of NSCLC by targeting FAM83F (Gu et al., 2018). However, no study has reported its role in osteosarcoma. But it was

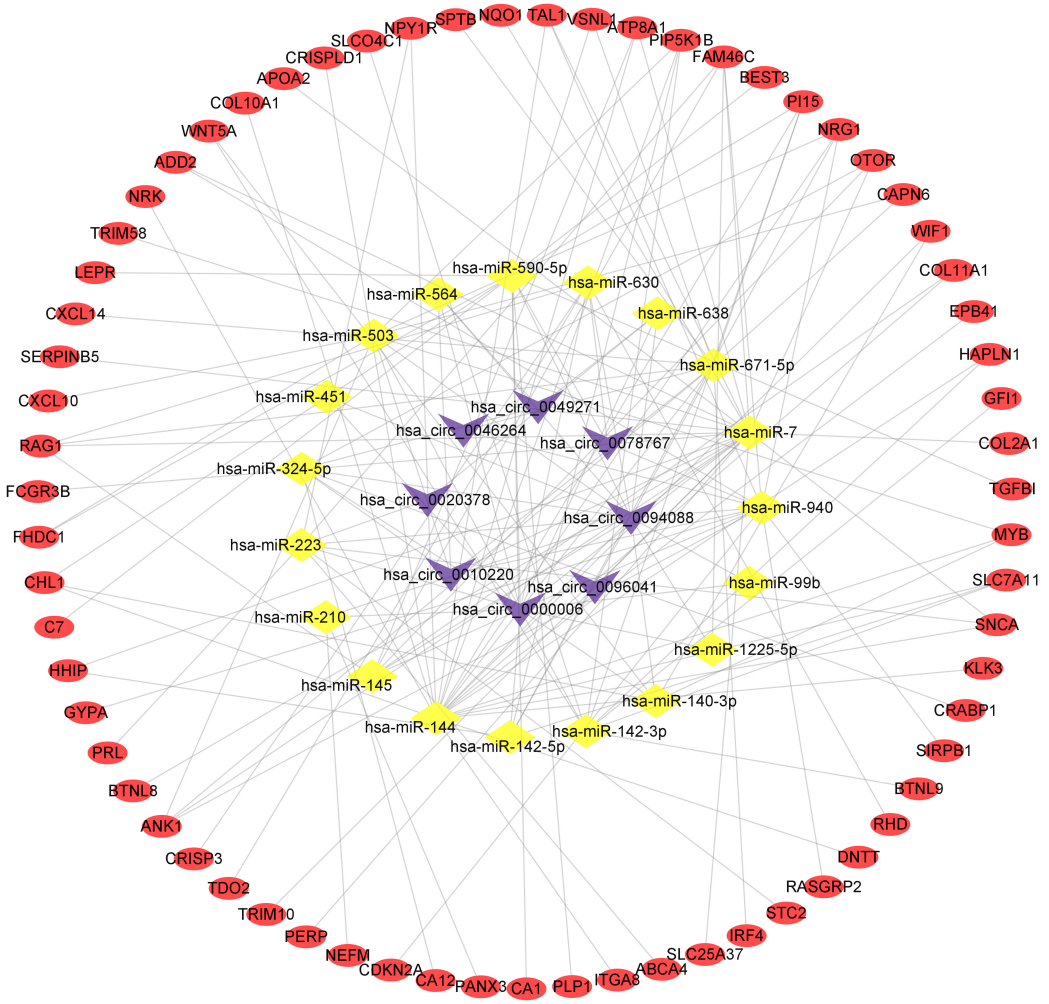


FIGURE 5 | The circRNA-miRNA-mRNA regulatory network. The purple V represents circRNAs, the yellow diamond represents miRNAs and the red oval represents mRNAs.

reported that cancer cells could secrete hsa-miR-940 to the bone microenvironment and induce an osteogenic phenotype by targeting ARHGAP1 and FAM134A (Hashimoto et al., 2018). It is well known that most osteosarcomas are osteolytic phenotypes. In our study, hsa-miR-940 was down-regulated, suggesting that it may play a tumor suppressor role, and this may also explain why osteosarcomas predominantly show an osteolytic rather than an osteogenic phenotype. Hsa-miR-223 was demonstrated to act as an oncogene in gastric cancer by targeting FBXW7/hCdc4 to regulate cell apoptosis, proliferation and invasion (Li et al., 2012). Hsa-miR-223 was also reported to work as a tumor-promotor in vulvar carcinoma by TP63 suppression (de Melo Maia et al., 2016). However, hsa-miR-223 was down-regulated in our study. We speculated mechanisms of the miRNAs vary among different malignancies. Thus, their regulatory relationships deserve further study.

To make the most of the network, we used it to explore potential compounds or drugs with reliable therapeutic effects for osteosarcoma. By using CLUE database and STITCH database, we screened out three candidate drugs, quinacridine, thalidomide and zonisamide. Quinacridine, also known as mepacrine, was originally used as an anti-malarial agent. But recently,

TABLE 3 | Survival analysis according to the Kaplan-Meier method and the Cox method for 67 shared mRNAs in TCGA profile.

mRNA	KM	Hazard ratio (HR)	HR.95L	HR.95H	*P*-value
RASGRP2	0.01941	1.05537	1.01768	1.09446	0.00368
STC2	0.00174	1.01455	1.00390	1.02531	0.00730

TABLE 4 | Potential drugs identified by CLUE and STITCH database for osteosarcoma.

Compounds	Description	Score
Quinacridine (Mepacrine)	Cytokine production inhibitor	−88.97
Thalidomide	TNF production inhibitor	−86.06
Zonisamide	Sodium channel blocker	−98.03

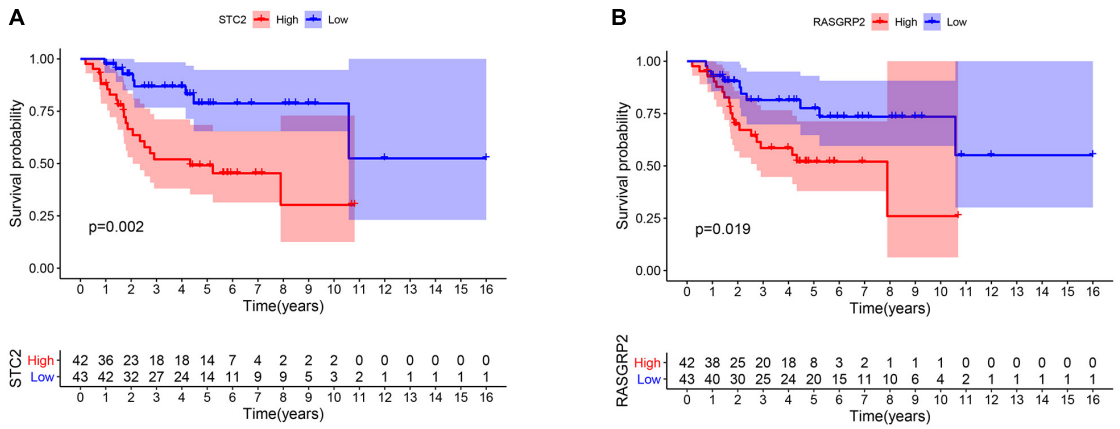


FIGURE 6 | The Kaplan-Meier survival curve of **(A)** STC2 and **(B)** RASGRP2.

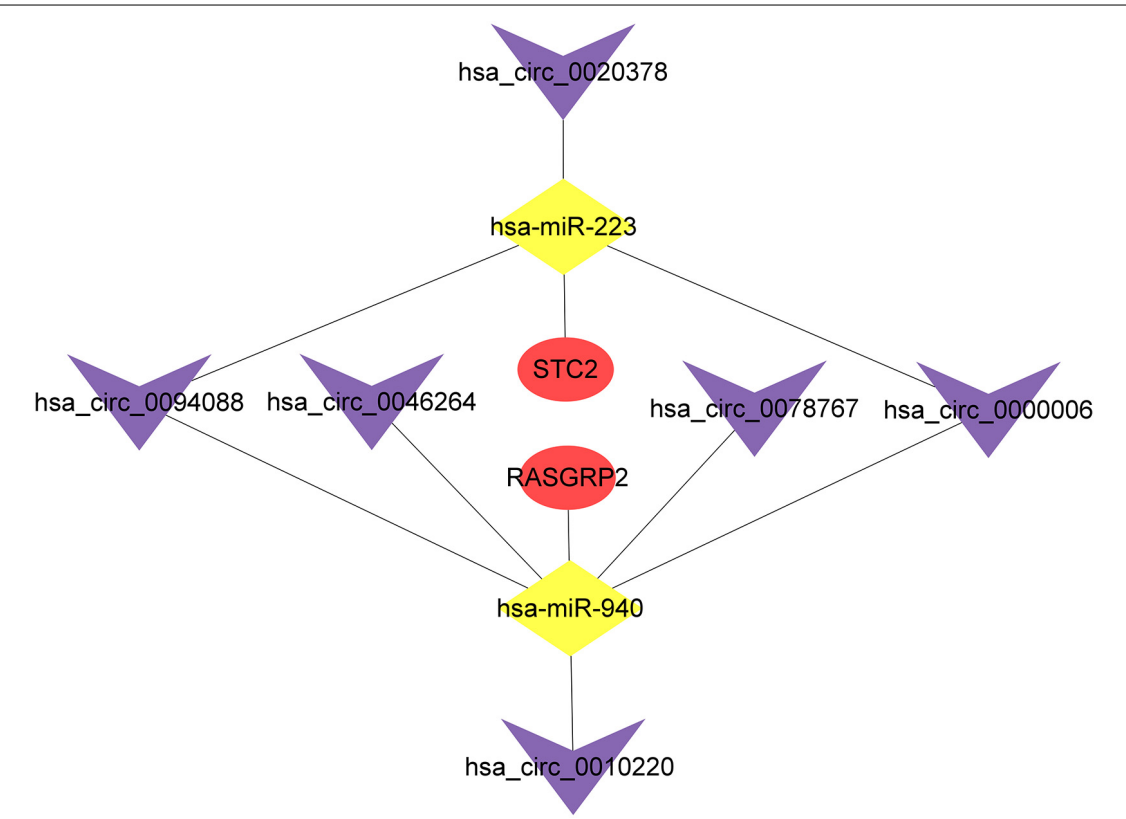


FIGURE 7 | The circRNA-miRNA-mRNA subnetwork based on mRNAs screened out with prognostic value. The purple V represents circRNAs, the yellow diamond represents miRNAs and the red oval represents mRNAs.

researchers have found that quinacridine could intercalate into DNA, impact nuclear proteins, inhibit the NFκB pathway and induce p53 expression to exhibit cytotoxicity on cancer cells (Oien et al., 2021). Pellerano et al. (2017) found that quinacridine analogs could bind CDK2/Cyclin A and inhibit its kinase activity to inhibit cancer cell proliferation, and promote accumulation of cells in S phase and G2. Hounsou et al. (2007) found that quinacridine had quadruplex binding properties and was able to target and interact with G-tetrads of the terminal part of the telomere, which made quinacridine a potentially powerful candidate for anti-cancer strategy. In addition, quinacridine was reported to enhance and restore the sensitivity of cisplatin in some cancer, which provided a new possible strategy for chemotherapy combinations (Friedman et al., 2007; Wang et al., 2010). Moreover, quinacridine has been applied in clinical practice for a long time with fewer

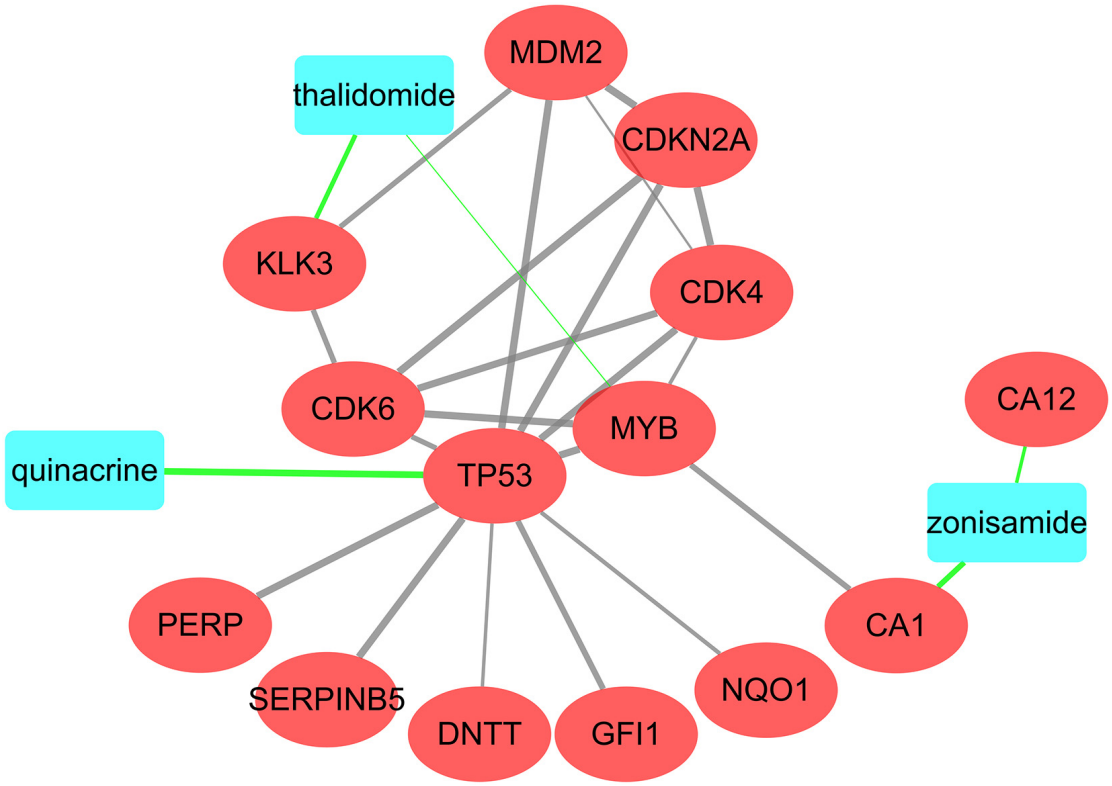


FIGURE 8 | Interactions among the three potential drugs and proteins coded by mRNAs. The blue rectangle represents the drugs and the red oval represents the proteins coded by shared mRNAs and other predicted proteins involved in the interaction.

side effect (Sokal et al., 2008). However, there is no study related to this drug in osteosarcoma. Currently, chemoresistance is a major obstacle in the treatment of osteosarcoma. Thus, quinacridine is a drug with great potential for application in osteosarcoma. Further trials are needed to evaluate its potential application. Thalidomide is a drug with anti-inflammatory and immunomodulatory properties. It was originally used to treat respiratory infections and to relieve morning sickness in pregnant women. However, it was withdrawn as it was found to be teratogenic (Sherbet, 2015). But thalidomide was reported to inhibit tumor cell proliferation, angiogenesis and induce apoptosis recently (Mercurio et al., 2017). Thalidomide has been also used to treat some malignancies, such as refractory multiple myeloma, prostate cancer and malignant glioma in the past (Singhal et al., 1999; Marx et al., 2001; Drake et al., 2003). And it was reported that the use of thalidomide and its analogs significantly improved the prognosis of patients with multiple myeloma (Holstein and McCarthy, 2017). A meta-analysis showed that thalidomide combined with transcatheter arterial chemoembolization (TACE) had better clinical outcomes and tolerable adverse events in patients with primary liver cancer compared to TACE alone (Gao et al., 2016). Thalidomide has also been verified to cause apoptosis in osteosarcoma cell lines (Zhu et al., 2016). A case report reported celecoxib combined with thalidomide in the treatment of refractory osteosarcoma, and achieved favorable outcome (Tsai et al., 2005). These results suggested thalidomide has potential to be an anti-tumor drug for osteosarcoma. However, there is still a lack of extensive clinical trials to support its use in osteosarcoma. Zonisamide is primarily used to treat epilepsy. No study has reported its effect in cancer. Further research is needed on whether it can be used in the treatment of osteosarcoma. The drug-protein network we constructed might contribute to the further study of these drugs.

To date, a previous study has constructed a circRNA-miRNA-mRNA regulatory network, but the circRNA microarray was based on 7 osteosarcoma cell lines and 1 normal bone cell, and the miRNA microarray was based on serum of osteosarcoma patients, not tumor tissues or cells (Qiu et al., 2020). The circRNA microarray in our study included 3 osteosarcoma tumor tissues and paired normal tissues, which were more comparable, and the expression data of miRNA microarray was from osteosarcoma cell lines. Therefore, our study can provide some new insights into the circRNA-miRNA-mRNA network regulatory network of osteosarcoma. Moreover, we predicted potential drugs that might be effective in the treatment of osteosarcoma. It will be helpful in providing new perspectives in the treatment of osteosarcoma. However, several limitations should be considered. First, our study was based a range of bioinformatics analysis methods and online databases. Next, when performing target and drug predictions, we have chosen the latest versions of data, the latest algorithms and a high degree of confidence level, but there may still be some inevitable random errors and selection bias.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/ and https://portal.gdc.cancer.gov/.

AUTHOR CONTRIBUTIONS

HC and YH performed the conception and design of this manuscript. YH and HZ provided useful suggestions in methodology. YH and WW performed data analysis and prepared the figures. HZ and HX drafted and revised the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank the GEO and TCGA databases for the availability of data.

REFERENCES

Chen, D., Zhao, Z., Huang, Z., Chen, D. C., Zhu, X. X., Wang, Y. Z., et al. (2018). Super enhancer inhibitors suppress MYC driven transcriptional amplification and tumor progression in osteosarcoma. *Bone Res.* 6:11. doi: 10.1038/s41413-018-0009-8

Chen, T., Yang, Z., Liu, C., Wang, L., Yang, J., Chen, L., et al. (2019). Circ_0078767 suppresses non-small-cell lung cancer by protecting RASSF1A expression via sponging miR-330-3p. *Cell Prolif.* 52:e12548. doi: 10.1111/cpr.12548

Chen, Y. T., Xie, J. Y., Sun, Q., and Mo, W. J. (2019). Novel drug candidates for treating esophageal carcinoma: a study on differentially expressed genes, using connectivity mapping and molecular docking. *Int. J. Oncol.* 54, 152–166. doi: 10.3892/ijo.2018.4618

Cheng, Z., Yu, C., Cui, S., Wang, H., Jin, H., Wang, C., et al. (2019). circTP63 functions as a ceRNA to promote lung squamous cell carcinoma progression by upregulating FOXM1. *Nat. Commun.* 10:3200. doi: 10.1038/s41467-019-11162-4

Cortini, M., Avnet, S., and Baldini, N. (2017). Mesenchymal stroma: role in osteosarcoma progression. *Cancer Lett* 405, 90–99. doi: 10.1016/j.canlet.2017.07.024

de Melo Maia, B., Rodrigues, I. S., Akagi, E. M., Soares do Amaral, N., Ling, H., Monroig, P., et al. (2016). MiR-223-5p works as an oncomiR in vulvar carcinoma by TP63 suppression. *Oncotarget* 7, 49217–49231. doi: 10.18632/oncotarget.10247

Drake, M. J., Robson, W., Mehta, P., Schofield, I., Neal, D. E., and Leung, H. Y. (2003). An open-label phase II study of low-dose thalidomide in androgen-independent prostate cancer. *Br. J. Cancer* 88, 822–827. doi: 10.1038/sj.bjc.6600817

DuBois, S., and Demetri, G. (2007). Markers of angiogenesis and clinical features in patients with sarcoma. *Cancer* 109, 813–819. doi: 10.1002/cncr.22455

Dudekula, D. B., Panda, A. C., Grammatikakis, I., De, S., Abdelmohsen, K., and Gorospe, M. (2016). CircInteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs. *RNA Biol.* 13, 34–42. doi: 10.1080/15476286.2015.1128065

Friedman, J., Nottingham, L., Duggal, P., Pernas, F. G., Yan, B., Yang, X. P., et al. (2007). Deficient TP53 expression, function, and cisplatin sensitivity are restored by quinacrine in head and neck cancer. *Clin. Cancer Res.* 13(22 Pt. 1), 6568–6578. doi: 10.1158/1078-0432.Ccr-07-1591

Gao, M., Kong, Y., Wang, H., Xie, B., Yang, G., Gao, L., et al. (2016). Thalidomide treatment for patients with previously untreated multiple myeloma: a meta-analysis of randomized controlled trials. *Tumour Biol.* 37, 11081–11098. doi: 10.1007/s13277-016-4963-8

Gorlick, R., Anderson, P., Andrulis, I., Arndt, C., Beardsley, G. P., Bernstein, M., et al. (2003). Biology of childhood osteogenic sarcoma and potential targets for therapeutic development: meeting summary. *Clin. Cancer Res.* 9, 5442–5453.

Gu, G. M., Zhan, Y. Y., Abuduwaili, K., Wang, X. L., Li, X. Q., Zhu, H. G., et al. (2018). MiR-940 inhibits the progression of NSCLC by targeting FAM83F. *Eur. Rev. Med. Pharmacol. Sci.* 22, 5964–5971. doi: 10.26355/eurrev_201809_15927

Guo, X., Keyes, W. M., Papazoglu, C., Zuber, J., Li, W., Lowe, S. W., et al. (2009). TAp63 induces senescence and suppresses tumorigenesis in vivo. *Nat. Cell Biol.* 11, 1451–1457. doi: 10.1038/ncb1988

Harrison, D. J., Geller, D. S., Gill, J. D., Lewis, V. O., and Gorlick, R. (2018). Current and future therapeutic approaches for osteosarcoma. *Expert Rev. Anticancer Ther.* 18, 39–50. doi: 10.1080/14737140.2018.1413939

Hashimoto, K., Ochi, H., Sunamura, S., Kosaka, N., Mabuchi, Y., Fukuda, T., et al. (2018). Cancer-secreted hsa-miR-940 induces an osteoblastic phenotype in the bone metastatic microenvironment via targeting ARHGAP1 and FAM134A. *Proc. Natl. Acad. Sci. U.S.A.* 115, 2204–2209. doi: 10.1073/pnas.1717363115

He, H., Qie, S., Guo, Q., Chen, S., Zou, C., Lu, T., et al. (2019). Stanniocalcin 2 (STC2) expression promotes post-radiation survival, migration and invasion of nasopharyngeal carcinoma cells. *Cancer Manag. Res.* 11, 6411–6424. doi: 10.2147/cmar.S197607

Holstein, S. A., and McCarthy, P. L. (2017). Immunomodulatory drugs in multiple myeloma: mechanisms of action and clinical experience. *Drugs* 77, 505–520. doi: 10.1007/s40265-017-0689-1

Hounsou, C., Guittat, L., Monchaud, D., Jourdan, M., Saettel, N., Mergny, J. L., et al. (2007). G-quadruplex recognition by quinacridines: a SAR, NMR, and biological study. *ChemMedChem* 2, 655–666. doi: 10.1002/cmdc.200600286

Jeck, W. R., and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nat. Biotechnol.* 32, 453–461. doi: 10.1038/nbt.2890

Kress, T. R., Sabò, A., and Amati, B. (2015). MYC: connecting selective transcriptional control to global RNA production. *Nat. Rev. Cancer* 15, 593–607. doi: 10.1038/nrc3984

Li, J., Guo, Y., Liang, X., Sun, M., Wang, G., De, W., et al. (2012). MicroRNA-223 functions as an oncogene in human gastric cancer by targeting FBXW7/hCdc4. *J. Cancer Res. Clin. Oncol.* 138, 763–774. doi: 10.1007/s00432-012-1154-x

Li, L., Sun, D., Li, X., Yang, B., and Zhang, W. (2021). Identification of key circRNAs in non-small cell lung cancer. *Am. J. Med. Sci.* 361, 98–105. doi: 10.1016/j.amjms.2020.08.008

Li, P., Xiao, Z., Luo, J., Zhang, Y., and Lin, L. (2019). MiR-139-5p, miR-940 and miR-193a-5p inhibit the growth of hepatocellular carcinoma by targeting SPOCK1. *J. Cell Mol. Med.* 23, 2475–2488. doi: 10.1111/jcmm.14121

Li, R., Jiang, J., Shi, H., Qian, H., Zhang, X., and Xu, W. (2020). CircRNA: a rising star in gastric cancer. *Cell Mol. Life Sci.* 77, 1661–1680. doi: 10.1007/s00018-019-03345-5

Li, S., and Wang, X. (2020). The potential roles of exosomal noncoding RNAs in osteosarcoma. *J. Cell Physiol.* 236, 3354–3365. doi: 10.1002/jcp.30101

Liu, Z. H., Yang, S. Z., Chen, X. T., Shao, M. R., Dong, S. Y., Zhou, S. Y., et al. (2020). Correlations of hsa_circ_0046264 expression with onset, pathological stage and chemotherapy resistance of lung cancer. *Eur. Rev. Med. Pharmacol. Sci.* 24, 9511–9521. doi: 10.26355/eurrev_202009_23036

Liu, Z., Yu, Y., Huang, Z., Kong, Y., Hu, X., Xiao, W., et al. (2019). CircRNA-5692 inhibits the progression of hepatocellular carcinoma by sponging miR-328-5p to enhance DAB2IP expression. *Cell Death Dis.* 10:900. doi: 10.1038/s41419-019-2089-9

Luetke, A., Meyers, P. A., Lewis, I., and Juergens, H. (2014). Osteosarcoma treatment - where do we stand? A state of the art review. *Cancer Treat Rev.* 40, 523–532. doi: 10.1016/j.ctrv.2013.11.006

Marx, G. M., Pavlakis, N., McCowatt, S., Boyle, F. M., Levi, J. A., Bell, D. R., et al. (2001). Phase II study of thalidomide in the treatment of recurrent glioblastoma multiforme. *J. Neurooncol.* 54, 31–38. doi: 10.1023/a:1012554328801

Mele, S., Devereux, S., Pepper, A. G., Infante, E., and Ridley, A. J. (2018). Calcium-RasGRP2-Rap1 signaling mediates CD38-induced migration of

chronic lymphocytic leukemia cells. *Blood Adv.* 2, 1551–1561. doi: 10.1182/bloodadvances.2017014506

Meng, S., Zhou, H., Feng, Z., Xu, Z., Tang, Y., Li, P., et al. (2017). CircRNA: functions and properties of a novel potential biomarker for cancer. *Mol. Cancer* 16:94. doi: 10.1186/s12943-017-0663-2

Mercurio, A., Adriani, G., Catalano, A., Carocci, A., Rao, L., Lentini, G., et al. (2017). A mini-review on thalidomide: chemistry, mechanisms of action, therapeutic potential and anti-angiogenic properties in multiple myeloma. *Curr. Med. Chem.* 24, 2736–2744. doi: 10.2174/0929867324666170601074646

Mirabello, L., Troisi, R. J., and Savage, S. A. (2009). Osteosarcoma incidence and survival rates from 1973 to 2004: data from the surveillance, epidemiology, and end results program. *Cancer* 115, 1531–1543. doi: 10.1002/cncr.24121

Nigro, J. M., Cho, K. R., Fearon, E. R., Kern, S. E., Ruppert, J. M., Oliner, J. D., et al. (1991). Scrambled exons. *Cell* 64, 607–613. doi: 10.1016/0092-8674(91)90244-s

Oien, D. B., Pathoulas, C. L., Ray, U., Thirusangu, P., Kalogera, E., and Shridhar, V. (2021). Repurposing quinacrine for treatment-refractory cancer. *Semin. Cancer Biol.* 68, 21–30. doi: 10.1016/j.semcancer.2019.09.021

Pellerano, M., Tcherniuk, S., Perals, C., Ngoc Van, T. N., Garcin, E., Mahuteau-Betzer, F., et al. (2017). Targeting conformational activation of CDK2 kinase. *Biotechnol. J.* 12. doi: 10.1002/biot.201600531

Qi, X., Zhang, D. H., Wu, N., Xiao, J. H., Wang, X., and Ma, W. (2015). ceRNA in cancer: possible functions and clinical implications. *J. Med. Genet.* 52, 710–718. doi: 10.1136/jmedgenet-2015-103334

Qiu, Y., Pu, C., Li, Y., and Qi, B. (2020). Construction of a circRNA-miRNA-mRNA network based on competitive endogenous RNA reveals the function of circRNAs in osteosarcoma. *Cancer Cell Int.* 20:48. doi: 10.1186/s12935-020-1134-1

Quan, G. M., and Choong, P. F. (2006). Anti-angiogenic therapy for osteosarcoma. *Cancer Metastasis Rev.* 25, 707–713. doi: 10.1007/s10555-006-9031-1

Ram Kumar, R. M., Betz, M. M., Robl, B., Born, W., and Fuchs, B. (2014). ΔNp63α enhances the oncogenic phenotype of osteosarcoma cells by inducing the expression of GLI2. *BMC Cancer* 14:559. doi: 10.1186/1471-2407-14-559

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

Sato, T., Takino, J. I., Nagamine, K., Nishio, K., and Hori, T. (2019). RASGRP2 suppresses apoptosis via inhibition of ROS production in vascular endothelial cells. *ScientificWorldJournal* 2019:4639165. doi: 10.1155/2019/4639165

Sherbet, G. V. (2015). Therapeutic potential of thalidomide and its analogues in the treatment of cancer. *Anticancer Res.* 35, 5767–5772.

Singhal, S., Mehta, J., Desikan, R., Ayers, D., Roberson, P., Eddlemon, P., et al. (1999). Antitumor activity of thalidomide in refractory multiple myeloma. *N. Engl. J. Med.* 341, 1565–1571. doi: 10.1056/nejm199911183412102

Sokal, D. C., Hieu do, T., Loan, N. D., Hubacher, D., Nanda, K., Weiner, D. H., et al. (2008). Safety of quinacrine contraceptive pellets: results from 10-year follow-up in Vietnam. *Contraception* 78, 66–72. doi: 10.1016/j.contraception.2008.02.011

Stone, J. C. (2011). Regulation and function of the RasGRP family of ras activators in blood cells. *Genes Cancer* 2, 320–334. doi: 10.1177/1947601911408082

Subramanian, A., Narayan, R., Corsello, S. M., Peck, D. D., Natoli, T. E., Lu, X., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 Profiles. *Cell* 171, 1437–1452.e17. doi: 10.1016/j.cell.2017.10.049

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* 44, D380–D384. doi: 10.1093/nar/gkv1277

Tamura, K., Furihata, M., Chung, S. Y., Uemura, M., Yoshioka, H., Iiyama, T., et al. (2009). Stanniocalcin 2 overexpression in castration-resistant prostate cancer and aggressive prostate cancer. *Cancer Sci.* 100, 914–919. doi: 10.1111/j.1349-7006.2009.01117.x

Tokar, T., Pastrello, C., Rossos, A. E. M., Abovsky, M., Hauschild, A. C., Tsay, M., et al. (2018). mirDIP 4.1-integrative database of human microRNA target predictions. *Nucleic Acids Res.* 46, D360–D370. doi: 10.1093/nar/gkx1144

Tsai, Y. C., Wu, C. T., and Hong, R. L. (2005). Response of refractory osteosarcoma to thalidomide and celecoxib. *Lancet Oncol.* 6, 997–999. doi: 10.1016/s1470-2045(05)70468-x

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., et al. (2015). DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.* 43, W460–W466. doi: 10.1093/nar/gkv403

Wang, B. D., Ceniccola, K., Hwang, S., Andrawis, R., Horvath, A., Freedman, J. A., et al. (2017). Alternative splicing promotes tumour aggressiveness and drug resistance in African American prostate cancer. *Nat. Commun.* 8:15921. doi: 10.1038/ncomms15921

Wang, Y., Bi, Q., Dong, L., Li, X., Ge, X., Zhang, X., et al. (2010). Quinacrine enhances cisplatin-induced cytotoxicity in four cancer cell lines. *Chemotherapy* 56, 127–134. doi: 10.1159/000313525

Wang, Y., Li, H., Ma, J., Fang, T., Li, X., Liu, J., et al. (2019). Integrated bioinformatics data analysis reveals prognostic significance Of SIDT1 in triple-negative breast cancer. *Onco Targets Ther.* 12, 8401–8410. doi: 10.2147/ott.S215898

Yang, L., Wang, J., Fan, Y., Yu, K., Jiao, B., and Su, X. (2018). Hsa_circ_0046264 up-regulated BRCA2 to suppress lung cancer through targeting hsa-miR-1245. *Respir. Res.* 19:115. doi: 10.1186/s12931-018-0819-7

Yang, S., Ji, Q., Chang, B., Wang, Y., Zhu, Y., Li, D., et al. (2017). STC2 promotes head and neck squamous cell carcinoma metastasis through modulating the PI3K/AKT/Snail signaling. *Oncotarget* 8, 5976–5991. doi: 10.18632/oncotarget.13355

Yokobori, T., Mimori, K., Ishii, H., Iwatsuki, M., Tanaka, F., Kamohara, Y., et al. (2010). Clinical significance of stanniocalcin 2 as a prognostic marker in gastric cancer. *Ann. Surg. Oncol.* 17, 2601–2607. doi: 10.1245/s10434-010-1086-0

Yu, B., Zhang, F., Liu, L., Liang, Y., Tang, X., Peng, Y., et al. (2021). The novel prognostic risk factor STC2 can regulate the occurrence and progression of osteosarcoma via the glycolytic pathway. *Biochem. Biophys. Res. Commun.* 554, 25–32. doi: 10.1016/j.bbrc.2021.03.067

Zhang, C., Chen, S., Ma, X., Yang, Q., Su, F., Shu, X., et al. (2019). Upregulation of STC2 in colorectal cancer and its clinicopathological significance. *Onco Targets Ther.* 12, 1249–1258. doi: 10.2147/ott.S191609

Zhang, X. O., Wang, H. B., Zhang, Y., Lu, X., Chen, L. L., and Yang, L. (2014). Complementary sequence-mediated exon circularization. *Cell* 159, 134–147. doi: 10.1016/j.cell.2014.09.001

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. doi: 10.1038/s41467-019-09234-6

Zhu, J., Yang, Y., Liu, S., Xu, H., Wu, Y., Zhang, G., et al. (2016). Anticancer effect of thalidomide in vitro on human osteosarcoma cells. *Oncol. Rep.* 36, 3545–3551. doi: 10.3892/or.2016.5158

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fgene.2021.671866

DNA Methylation, Deamination, and Translesion Synthesis Combine to Generate Footprint Mutations in Cancer Driver Genes in B-Cell Derived Lymphomas and Other Cancers

Igor B. Rogozin[1], Abiel Roche-Lima[2], Kathrin Tyryshkin[3], Kelvin Carrasquillo-Carrión[4], Artem G. Lada[5], Lennard Y. Poliakov[6], Elena Schwartz[7], Andreu Saura[6], Vyacheslav Yurchenko[6,8], David N. Cooper[9], Anna R. Panchenko[3] and Youri I. Pavlov[10,11,12]**

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States, [2] Center for Collaborative Research in Health Disparities – RCMI Program, University of Puerto Rico, San Juan, Puerto Rico, [3] Department of Pathology and Molecular Medicine, School of Medicine, Queen's University, Kingston, ON, Canada, [4] Integrated Informatics Services Core – RCMI, University of Puerto Rico, San Juan, Puerto Rico, [5] Department Microbiology and Molecular Genetics, University of California, Davis, Davis, CA, United States, [6] Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czechia, [7] Coordinating Center for Clinical Trials, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States, [8] Martsinovsky Institute of Medical Parasitology, Tropical and Vector Borne Diseases, Sechenov First Moscow State Medical University, Moscow, Russia, [9] Institute of Medical Genetics, Cardiff University, Cardiff, United Kingdom, [10] Eppley Institute for Research in Cancer and Allied Diseases, Omaha, NE, United States, [11] Department of Microbiology and Pathology, Biochemistry and Molecular Biology, Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, United States, [12] Department of Genetics and Biotechnology, Saint-Petersburg State University, Saint-Petersburg, Russia*

OPEN ACCESS

Edited by:
Yuriy L. Orlov, The Digital Health Institute, I.M. Sechenov First Moscow State Medical University, Russia

Reviewed by:
Richard Chahwan, University of Zurich, Switzerland
Robert W. Maul, National Institute on Aging, National Institutes of Health (NIH), United States
Alexei Fedorov, University of Toledo, United States

***Correspondence:**
Youri I. Pavlov
ypavlov@unmc.edu
Vyacheslav Yurchenko
vyacheslav.yurchenko@osu.cz

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

Received: *24 February 2021*
Accepted: *21 April 2021*
Published: *19 May 2021*

Citation:
Rogozin IB, Roche-Lima A, Tyryshkin K, Carrasquillo-Carrión K, Lada AG, Poliakov LY, Schwartz E, Saura A, Yurchenko V, Cooper DN, Panchenko AR and Pavlov YI (2021) DNA Methylation, Deamination, and Translesion Synthesis Combine to Generate Footprint Mutations in Cancer Driver Genes in B-Cell Derived Lymphomas and Other Cancers. Front. Genet. 12:671866. doi: 10.3389/fgene.2021.671866

Cancer genomes harbor numerous genomic alterations and many cancers accumulate thousands of nucleotide sequence variations. A prominent fraction of these mutations arises as a consequence of the off-target activity of DNA/RNA editing cytosine deaminases followed by the replication/repair of edited sites by DNA polymerases (pol), as deduced from the analysis of the DNA sequence context of mutations in different tumor tissues. We have used the weight matrix (sequence profile) approach to analyze mutagenesis due to Activation Induced Deaminase (AID) and two error-prone DNA polymerases. Control experiments using shuffled weight matrices and somatic mutations in immunoglobulin genes confirmed the power of this method. Analysis of somatic mutations in various cancers suggested that AID and DNA polymerases η and θ contribute to mutagenesis in contexts that almost universally correlate with the context of mutations in A:T and G:C sites during the affinity maturation of immunoglobulin genes. Previously, we demonstrated that AID contributes to mutagenesis in (de)methylated genomic DNA in various cancers. Our current analysis of methylation data from malignant lymphomas suggests that driver genes are subject to different (de)methylation processes than non-driver genes and, in addition to AID, the activity of pols η and θ

contributes to the establishment of methylation-dependent mutation profiles. This may reflect the functional importance of interplay between mutagenesis in cancer and (de)methylation processes in different groups of genes. The resulting changes in CpG methylation levels and chromatin modifications are likely to cause changes in the expression levels of driver genes that may affect cancer initiation and/or progression.

Keywords: tumor cells, frequency matrices, database, computational biology, somatic hypermutation, immunoglobulin genes, gene expression

INTRODUCTION

Epigenetic reprogramming in cancer genomes creates a distinct DNA methylation landscape encompassing clustered sites of hypermethylation at regulatory regions and protein-coding genes separated by long intergenic tracks of hypomethylated regions. Such changes in DNA methylation landscape are displayed by most cancer types, and hence have the potential to serve as a universal cancer biomarker (Sina et al., 2018; Oliver et al., 2021). Previous research has focused on the biological consequences of DNA methylation changes in genomes, whereas its impact on the structure and flexibility of DNA, and its vulnerability to modifications/repair/replication in cancer, have remained largely unexplored.

Other prominent features of cancer initiation and progression are genomic alterations. Cancer genomes harbor numerous genomic alterations, including hundreds/thousands of nucleotide sequence variations (Stratton et al., 2009; Roberts and Gordenin, 2014; Rogozin et al., 2018c). A prominent fraction of these mutations arises as a consequence of the off-target activity of enzymes participating in somatic hypermutation (SHM) in immunoglobulin (Ig) genes: DNA/RNA editing cytosine deaminases of the Activation Induced Deaminase (AID)/APOBEC family and the replication/repair of edited sites by DNA polymerases (pols), as deduced by the analysis of the DNA sequence context of mutations in different cancer tissues (Alexandrov et al., 2013; Roberts and Gordenin, 2014; Swanton et al., 2015; Granadillo Rodriguez et al., 2020). Analyses of various types of cancer by means of this technique has yielded a set of 30–50 distinct mutation signatures implying many mechanisms of hypermutation in cancer cells (Alexandrov and Stratton, 2014; Goncearenco et al., 2017; Rogozin et al., 2018c; Islam and Alexandrov, 2021).

There is a well-established association between DNA methylation and genomic alteration. Early studies revealed that methylated cytosines explain mutation hotspots in bacteria (Coulondre et al., 1978). In eukaryotic genomes, CpG sites are known to be vulnerable to mutation in both cancer and normal cells (Cooper and Youssoufian, 1988; Alsøe et al., 2017; Goncearenco et al., 2017; Rogozin et al., 2018c; Brinkman et al., 2019). We recently detected a substantial excess of mutations in CpG sites that overlap with AID mutable motifs (WRC/GYW, W = A or T, R = A or G, Y = T or C, the mutable position is underlined) forming "hybrid" mutable motifs (WRCG/CGYW) whereas the opposite trend was observed in SHM (Rogozin and Diaz, 2004; Rogozin et al., 2016). This finding implies that in many cancers the SHM machinery acts aberrantly at genomic sites containing methylated cytosine. The discovery of abundant mutations in WRCG/CGYW motifs in many types of human cancer suggests that AID-mediated, CpG methylation-dependent mutagenesis is a common feature of tumorigenesis connecting methylation and hypermutation (Rogozin et al., 2016).

A prominent feature of carcinogenesis is the presence of cancer driver and passenger mutations. A driver mutation directly or indirectly confers a selective advantage upon cancer cells, whilst a passenger mutation does not (Stratton et al., 2009). In this context, it should be appreciated that there is a difference between a driver gene and a driver gene mutation: a driver gene may accumulate recurrent driver mutations but may also harbor passenger mutations. Some genes contain only recurrent passenger mutations with frequencies comparable to driver genes (hotspots related to the intrinsic properties of the processes of mutagenesis), which complicates the identification of cancer driver mutations (Rogozin et al., 2018c). In this study, we operationally define a non-driver gene as a gene that contains numerous mutations that do not cause cancer and are classified as passenger mutations according to the MutaGene (Goncearenco et al., 2017; Brown et al., 2019) and CHASMplus (Tokheim and Karchin, 2019) computational tools.

We studied the association of mutable motifs produced by AID and two error-prone DNA pols ultimately associated with cancer, and the methylation status of sets of driver and non-driver genes. Our null hypothesis was that driver and non-driver genes would have contrasting methylation and mutation profiles, which could be studied using mutable motifs (Rogozin et al., 2016). The conventional method for the analysis of mutable DNA motifs is the consensus approach (Alexandrov and Stratton, 2014), for example, 5′WRC for the AID enzyme (Pham et al., 2011; Rogozin et al., 2018c) or 5′WA for DNA pol η (Rogozin et al., 2001, 2018b). Here, we applied the weight matrix (sequence profile) approach that is frequently used in the analysis of biological processes (Rogozin et al., 2019) to the analysis of methylation profiles and mutagenesis generated by AID and error-prone DNA pols η and θ in CpG dinucleotides. Control experiments, using shuffled sites and SHM in immunoglobulin genes, suggested that the weight matrix method adds power to the study of mutagenesis. Analysis of mutations in various cancers indicated that AID and DNA pol η mutable motifs almost universally correlate with SHM in G:C sites. Analysis of mutations and motifs in A:T sites yielded a similar correlation for pol θ. Analysis of methylation data in malignant lymphomas (the MALY-DE dataset) suggested that the methylation status of driver genes differs from that of non-driver genes and this may be one reason for the differences in distribution of mutations in the two groups of genes.

MATERIALS AND METHODS

Mutable Motif Construction Using Weight Matrices

Several approaches have been developed for the analysis of a set of mutated genomic sequences (Staden, 1984; Rogozin et al., 2018b, 2019). A mononucleotide weight matrix is a simple and straightforward way to present the structure of a functional signal and to calculate weights for the signal sequence (Gelfand, 1995). Each matrix W(b,j) (nucleotide b = A, T, G, or C in a position j) includes information on a normalized frequency of A, T, G, and C bases in each of the six positions surrounding detected sites of mutation (3 bases downstream and 3 bases upstream; **Figure 1**; corresponding raw numbers are shown in the **Supplementary Figure 1**). We calculated the weight matrices for the two studied DNA polymerases and used a collection of mutations generated by classic gap-filling DNA synthesis *in vitro* by human pols η and θ (Matsuda et al., 2001; Rogozin et al., 2001; Arana et al., 2008) (**Supplementary Figures 2, 3**).

The following formula for W(b,j) was used for data analysis: $W(b,j) = \log_2 [f(b,j) / e(b)]$, where f(b,j) is the observed frequency of the nucleotide b in position j and e(b) is the expected frequency of the nucleotide b calculated as the mean nucleotide frequencies of positions −5,−4, +4, +5 for the sites of mutation in the target sequence; the resulting W(b,j) matrices are shown in **Figure 1**.

The matching score $S_{(b1,\ldots,bL)}$ of a sequence b1,...,bL is:

$$S_{(b1,\ldots,bL)} = \sum_{j=1}^{L} W(b, j)$$

The matching score between sequence b1,...,bL and a weight matrix can be further expressed as a percentage:

$$\%\text{matching score} = 100 \times \left(S_{(b1,\ldots,bL)} - S_{min}\right) / \left(S_{max} - S_{min}\right)$$

$$S_{(min)} = \sum_{j=1}^{L} \underset{b}{MIN}(W(b, j))$$

$$S_{(max)} = \sum_{j=1}^{L} \underset{b}{MAX}(W(b, j))$$

Hereafter, we use the term "weight" instead of "% matching score." We used positions −3 : +3 to estimate the weights of sites.

ICGC/TCGA Mutation Datasets

Somatic mutation data from the ICGC and TCGA cancer genome projects were extracted from the Sanger COSMIC Whole Genome Project v75.[1] The ICGC/TCGA datasets are almost exclusively passenger mutations and, as such, they are unlikely to be subject to selection in the context of promoting cellular proliferation. Indeed, they are much more likely to reflect unselected mutational spectra (Goncearenco et al., 2017; Rogozin et al., 2018c). The tissues and cancer types were defined according to the primary tumor site and the cancer project in question. This dataset is included in the MutaGene package, where it is described in detail (Goncearenco et al., 2017; Brown et al., 2019). We also used collections of mutations obtained by means of *in vitro* experiments for human pol η (Matsuda et al., 2001) and pol θ (Arana et al., 2008; **Supplementary Figures 2, 3**) to build weight matrices.

Methylation and Expression Data

For the analysis of the association between somatic mutations, mRNA expression, mutable motifs and methylation, datasets for 26 patients with malignant lymphoma[2] were used. In the analyzed datasets, the methylation data for all patients were pooled together. Each position was characterized by the methylated/unmethylated read count and the methylation ratio (the number of methylated reads divided by the total number of reads overlapping this position and multiplied by 100). Only positions with more than nine associated reads were included in the analysis. The major methodological problem inherent in the analysis of methylation across CpG's is the absence of control sets. Therefore, we compared methylation values below and above threshold values (25 and 75%). The mean weight of mutable motifs (**Figure 1**) in the positions of methylated CpG's (the group 1 with the size S1, **Figure 2**) was compared to the mean weight of the same motifs in a contrasting dataset (the group 2 with the size S2, **Figure 2**) using the *t*-test (2-tailed test) and Monte Carlo test (MC, 1-tailed test) similar to the consensus method as previously described (Rogozin et al., 2018b). Expression of mRNA was measured using the FPKM values (Howe et al., 2011). The mean and variance for each gene were calculated across 26 studied samples.

Analysis of Mutations

DNA sequences surrounding the mutated nucleotides represent the mutation context. We compared the frequency of known mutable motifs for somatic mutations with the frequency of these motifs in the vicinity of the mutated nucleotide. Specifically, for each base substitution, the 121 bp sequence centered at the mutation was extracted (the DNA neighborhood). We used only the nucleotides immediately flanking the mutations because DNA repair/replication enzymes are thought to scan a very limited region of DNA (Roberts et al., 2013; Goncearenco et al., 2017; Rogozin et al., 2018c). This approach does not exclude any specific area of the genome, but rather uses the areas within each sample where mutagenesis has occurred (considering the variability in the mutation rate across the human genome) (Roberts et al., 2013; Rogozin et al., 2018b). A schematic representation of this procedure for CpG dinucleotides is shown in **Supplementary Figure 4**). Here, the mean weight of mutable motifs (represented by weight matrices; **Figure 1**) in the positions of each somatic mutation (in C/G or A/T positions) was compared to the mean weight of mutable motifs in C/G or A/T positions without mutations in the DNA neighborhood (**Supplementary Figure 4**) using the *t*-test (2-tailed test) and Monte Carlo test (MC, 1-tailed test) similar to the consensus

[1] https://cancer.sanger.ac.uk

[2] https://dcc.icgc.org/projects/MALY-DE

A

	−3	−2	−1	0	+1	+2	+3
A	0.28	0.21	0.32	1.00	0.29	0.16	0.31
T	0.14	0.29	0.28	0.00	0.34	0.26	0.22
G	0.30	0.17	0.21	0.00	0.11	0.26	0.23
C	0.28	0.34	0.19	0.00	0.27	0.32	0.24
		Y	**W**	**A**	*H*		

B

	−3	−2	−1	0	+1	+2	+3
A	0.21	0.24	0.47	1.00	0.25	0.12	0.15
T	0.19	0.31	0.09	0.00	0.27	0.25	0.32
G	0.45	0.29	0.22	0.00	0.26	0.30	0.26
C	0.15	0.16	0.22	0.00	0.22	0.32	0.27
			A	**A**			

C

	−3	−2	−1	0	+1	+2	+3
A	0.22	0.28	0.10	0.00	0.16	0.12	0.33
T	0.36	0.35	0.12	0.00	0.20	0.38	0.23
G	0.29	0.29	0.20	1.00	0.22	0.22	0.16
C	0.13	0.09	0.58	0.00	0.42	0.29	0.28
		D	*C*	G			

D

	−3	−2	−1	0	+1	+2	+3
A	0.26	0.16	0.23	0.00	0.19	0.14	0.25
T	0.17	0.31	0.25	0.00	0.26	0.23	0.38
G	0.23	0.19	0.18	1.00	0.13	0.26	0.22
C	0.33	0.34	0.34	0.00	0.43	0.37	0.15
				G	*C*		

FIGURE 1 | Nucleotide frequency matrices for mutations at A:T sites [**(A)** DNA pol η; **(B)** pol θ] and G:C sites [**(C)** pol θ; **(D)** DNA pol η]. Known mutable motifs (consensus sequences) (Matsuda et al., 2001; Rogozin et al., 2001) are shown below each matrix in bold, whereas mutable positions are underlined. Putative (previously unobserved) parts of mutable motifs and potentially informative positions are italicized, W = A or T; Y = T or C; B = A, T or G; D = A, T, or G. Source of data: **Supplementary Figures 2, 3**.

method, as previously described (Rogozin et al., 2018b). The MC test is based on the random sampling from the group 2. In total, 10,000 groups with size S1 have been generated. The fraction of generated groups with mean weights larger or equal to the mean value of the sample 1 is the P value.

In addition to analyses of the derived mutable motifs in cancer genomes, we performed a control experiment: we randomly shuffled a dataset of sequences surrounding the mutations in the studied target sequences (**Supplementary Figures 2, 3**) keeping position 6 (the position of mutations) intact. Each sequence was shuffled separately; thus, the overall base composition and the base compositions of each sequence were the same. Weight matrices were derived from these shuffled sequences, and the sampling procedure was repeated 1,000 times.

Detection of Driver and Non-driver Genes

In this study, we used two independent methods to predict the driver status of cancer mutations: the MutaGene (Goncearenco et al., 2017; Brown et al., 2019) and Chasmplus (Tokheim and Karchin, 2019). These methods showed top

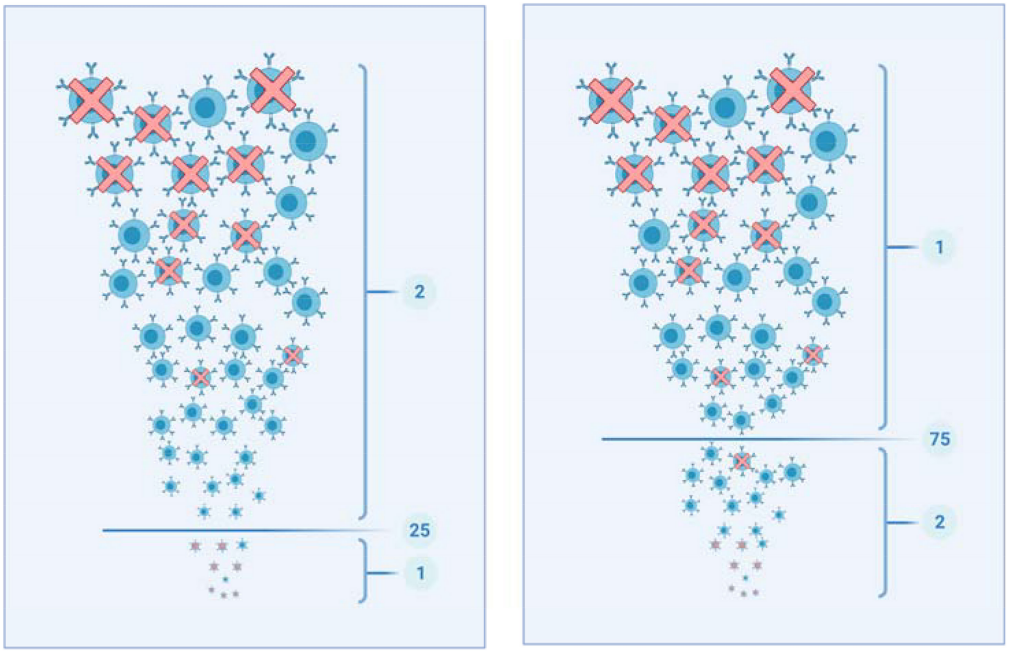


FIGURE 2 | Schematic representation of the procedure used for construction of **Table 3**. Each circle represents a methylated CpG site, with its size reflecting the methylation level. Red "X" denotes CpG sites that overlap with known mutable motifs. The left and right panels correspond to thresholds 25% and 75%. The left panel: The set "1" (the methylation levels are smaller than 25%) is compared to set "2" (the methylation levels are larger than 25%). The right panel: The set "1" (the methylation levels are larger than 75%) is compared to set "2" (the methylation levels are smaller than 75%).

performance on a recent benchmarking set (Brown et al., 2019). MutaGene is a probabilistic approach which adjusts the number of mutation recurrences in patients by means of a cancer-type specific background mutation model. The MutaGene driver mutation prediction method has not been explicitly trained on any particular set of mutations. The background models estimate the probability of obtaining a codon substitution from the underlying processes of mutagenesis. We used two MutaGene background models: one was derived from pan-cancer mutational data ("Pancancer" model in MutaGene) whereas the other was constructed directly from the MALY-DE mutational data since this cancer-specific model was not present in the MutaGene database of background models. As a result, two ranking lists of driver mutations were produced for three types of mutation: missense, nonsense and silent. Chasmplus is a machine learning method that was trained using somatic mutations from TCGA. Since no cancer-specific model was available for MALY-DE, we used pan-cancer predictions while running Chasmplus. Then we merged the predictions produced by the three different models/methods and reported only those mutations as drivers which were predicted to be "drivers" or "potential drivers" by MutaGene and had a Chasmplus score cutoff larger than 0.5. **Supplementary File 1** shows recurrent driver and passenger mutations.

Predicted driver mutations satisfy at least two of the above-mentioned criteria of driver mutations (**Supplementary File 2**). Predicted passenger mutations must satisfy all criteria of passenger mutations. Since Chasmplus does not generate predictions for nonsense and silent mutations, only predictions for missense mutations were reported. In addition, some mutations/genes were not reported by Chasmplus because it excluded them from the list of potential cancer driver genes. In this study, we defined driver genes in the following way: a driver gene must have at least one recurrent driver mutation but may also possess recurrent passenger mutations (**Supplementary Table 1**). Some genes contain only recurrent passenger mutations with frequencies comparable to driver genes. In this study, we defined a non-driver gene operationally as a gene that only contains recurrent mutations that are not associated with the process of tumorigenesis and hence are classified as passenger mutations (**Supplementary Table 2**).

RESULTS

Weight Matrices Are Powerful Descriptors of Mutable Motifs

Weight matrices constitutes a novel technique when applied to the description of preferential mutable motifs. It was shown to be a robust and precise technique to describe AID/APOBEC mutable motifs in cancer cells (Rogozin et al., 2019). The weight matrices include information on the frequency of A, T, G, and C bases in each of the ten positions surrounding the sites of mutation (5 bases downstream and 5 bases upstream). AID, DNA pol η and pol θ are involved in SHM in immunoglobulin genes (Revy et al., 2000; Matsuda et al., 2001; Pavlov et al., 2002; Zan et al., 2005; Neuberger and Rada, 2007; Arana et al., 2008; Bhattacharya et al., 2008), although this role for both polymerases

has been questioned (Dörner and Lipsky, 2001; Martomo et al., 2008).

In this study, we started from the construction of weight matrices for both DNA pols. It should be noted that we previously derived weight matrices using collections of mutations induced by AID/APOBEC deaminases in yeast genomes (Rogozin et al., 2019). For human DNA pols η and θ, such collections are not available. Therefore, we used a collection of mutations generated by human pols η and θ during classic gap-filling DNA synthesis *in vitro* (Matsuda et al., 2001; Rogozin et al., 2001; Pavlov et al., 2002) (**Supplementary Figures 2, 3**). Constructed matrices of nucleotide frequencies are shown in **Figure 1A-D** (corresponding raw numbers are shown in the **Supplementary Figure 1**). Pols η and θ exhibit known DNA context features for mutations in A:T sites. W (A or T) or A in the position -1 (**Figures 1A,B**) was the most prominent feature of A:T mutations produced by pol η and pol θ, accordingly. We cannot exclude the possibility that some other previously undetected positions may contribute to the mutable motifs, for example, a higher frequency of Y (T or C) in position -2 or a lower frequency of G may be additional features of the pol η mutable motif (**Figure 1A**).

By contrast, pols η and θ exhibit dissimilar DNA context features for mutations at G:C sites. A characteristic feature of pol θ is an elevated frequency of C at position –1 and a lower frequency of C at position –2 (**Figure 1C**). Thus, pol θ tends to produce more errors in the DCG nucleotide context (D = A or T or G). Pol η appears to have a different DNA mutational context with an excess of C in position +1 (**Figure 1D**). In general, it is hard to confidently delineate mutable motifs of either DNA polymerase using the consensus approach owing to the lack of objective inclusion criteria for position-specific context features to mutable motifs (**Figure 1**). Thus, the weight matrix approach, which utilizes information contained in all studied positions, is likely to be a more straightforward way to describe the polymerase η and θ mutable motifs than the consensus approach.

We also compared the nucleotide composition of sequences surrounding positions of mutations (**Supplementary Figure 1**) for pols η and θ using the χ^2 test. We found that these pols were significantly different with respect to the DNA sequence context of mutation sites expressed in the form of nucleotide frequency matrices (A:T sites: $\chi^2 = 155.0$, df = 40, $P = 1.9 \times 10^{-15}$; G:C sites: $\chi^2 = 82.2$, df = 40, $P = 0.00007$). Thus, DNA pol η and pol θ differ significantly in terms of the features of the DNA sequence context of mutations. This result is consistent with the different context properties of pols η and θ (**Figure 1**).

Footprints of pol η and pol θ Correlate With the Somatic Mutational Spectrum in Many Cancer Types

Previously, we demonstrated using the consensus approach that mutagenesis by AID is likely modulated by the (de)methylation and/or translesion synthesis (TLS) of CpG dinucleotides in follicular lymphomas and many other cancers (Rogozin et al., 2016). Based on analyses of mutations in CpG dinucleotides in skin cancer cells and normal cells, it was also suggested that pol η mutagenesis might also correlate with the methylation of CpG dinucleotides in cancer cells (Rogozin et al., 2018b). The weight matrix approach and the MALY-DE datasets (CpG methylation spectra and somatic mutations, see Materials and Methods) allow us to perform further analyses of the role of AID and error-prone polymerases in mutagenesis, and to see how it is affected by (de)methylation.

We examined the correlation between the nucleotide sequence context of somatic mutations in cancers and pol η and pol θ mutable motifs found after *in vitro* DNA synthesis. A correlation was inferred when the results of two statistical tests (Monte Carlo test and *t*-test) were significant at $P < 0.05$. AID has already been studied using the consensus motif WRC/GYW and weight matrices and has been shown to be one of the most ubiquitous contributors to mutations in various cancer types according to its characteristic mutable motif (the AID weight matrix) (Rogozin et al., 2019). Analysis of pol-generated mutations in G:C sites revealed that both mutation motifs are almost universally correlated with the nucleotide context of somatic mutations in G:C sites (**Figure 3**). Similar analysis of A:T site mutations also revealed correlations for pol η. A significant correlation with pol θ was documented only for a few cancer cases. This difference may reflect a more specialized role for pol θ in DNA transactions on methylated CpG's (Wood and Doublié, 2016; Brambati et al., 2020). It is also possible that pol θ is expressed in only a few cancers. Pol η probably plays a more widespread, although not particularly pronounced, role in causing mutations in cancer according to its characteristic weight matrix in various cancer types; this is consistent with our previous study where we used the consensus sequence WA (Rogozin et al., 2018b).

Control Experiments

The *in vitro* collections of mutations that were used to reconstruct weight matrices for pol η and pol θ (Matsuda et al., 2001; Arana et al., 2008) are relatively small (**Supplementary Figures 2, 3**). Thus, control experiments were important to analyze the quality of the derived weight matrices. We previously demonstrated that analyses of the association between the matrices of shuffled sites of mutation and the nucleotide context of somatic mutation in various cancer cell types is a reliable approach to estimate the impact of the accuracy of association prediction (Rogozin et al., 2019). Analysis of 16 types of cancer (**Supplementary Table 5**) suggested that the AID weight matrix is less prone to errors of prediction compared to pol η/pol θ (**Supplementary Table 5**). Only a few types of cancer have a low level of prediction errors. Fortunately, for our study of MALY-DE sets, "Blood" tissue, GCB lymphomas (from the COSMIC database) and MALY_DE malignant lymphomas have extremely low rates of false prediction (**Supplementary Table 5**). Therefore, we opted to use the derived matrices for further analysis of the MALY-DE datasets.

Analysis of somatic mutations in immunoglobulin genes can be used to estimate the prediction accuracy because the context of mutations in human immunoglobulin genes is known to correlate strongly with AID and pol η mutable motifs (Matsuda et al., 2001). Thus, these mutations can be used as a control dataset as performed previously (Rogozin et al., 2019).

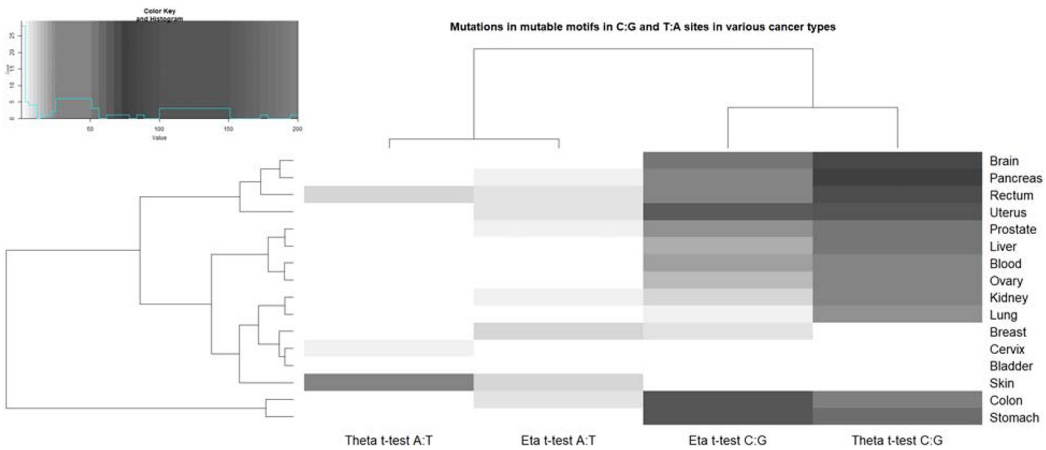


FIGURE 3 | Correlation of pol η (eta) and θ (theta) mutable motifs and the sequence context of somatic mutations. For the actual data, see **Supplementary Tables 3, 4**. The intensities of the gray color correspond to the *t*-test values (**Supplementary Tables 3, 4**). The unweighted pair group method, with arithmetic mean (UPGMA) clustering of ratio values for the pol η and θ footprints and tissues, is shown to the left and top. The upper left panel shows the distribution of the studied *t*-test values and correspondence of the *t*-test values and color intensity (the darker colors correspond to the higher correlation values). A similar plot of ratio values (the ratio being the mean weight of mutated sites divided by the mean weight of non-mutated sites) is shown in the **Supplementary Figure 5**.

TABLE 1 | Correlation between the sequence context of somatic mutations and mutable motifs in fragments of human immunoglobulin genes.

Locus	Test	Number of Mutations	AID/G:C	Pol η/G:C	Pol θ/G:C	Number of Mutations	Pol η/A:T	Pol θ/A:T
V_H26	Ratio	583	1.208	1.027	1.091	351	1.082	0.979
	t-test		**13.1***	NSE	**5.9***		**5.3***	NSE
	MC test		<0.001	0.004	<0.001		<0.001	0.699
J_H4 intron, control individuals	Ratio	177	1.341	1.05	1.029	95	1.041	1.032
	t-test		**12.3***	**2.8***	NSE		**2.4***	**2.2***
	MC test		<0.001	0.002	0.106		0.004	0.011
J_H4 intron, XP-V patients	Ratio	227	1.278	1.009	1.011	25	0.957	0.98
	t-test		**9.9***	NSE	NSE		NSE	NSE
	MC test		<0.001	0.329	0.061		0.776	0.67

"Ratio" is the mean weight of mutated sites divided by the mean weight of non-mutated sites. NSE (no significant excess) indicates the absence of a significant excess of mutations in mutable motifs, suggesting there to be no association between mutagenesis and mutable motifs. The significance of any excess was measured using the Student t and Monte Carlo (MC) tests. The asterisk (*) denotes that the corresponding P < 0.01; this is a conservative estimate of the critical overall value of the t-test having allowed for multiple testing by means of the Bonferroni correction (5 comparisons).

A significant association between the AID mutable motif and mutations was found in all three studied somatic mutation datasets (Milstein et al., 1998; Mayorov et al., 2005; **Table 1**), confirming that the AID weight matrix is a reliable descriptor of AID-induced mutagenesis. The pol η weight matrices revealed a significant association for all studied cases except xeroderma pigmentosum variant (XPV) patients where pol η is inactive (**Table 1**; Mayorov et al., 2005). Pol θ matrices yielded significant results for some studied cases (**Table 1**). This is consistent with the hypothesis that pol θ is also involved in SHM (Arana et al., 2008). The results of both control experiments suggested that the weight matrix technique approach is adequate to study the mutational spectra of DNA polymerases.

Analysis of Driver and Non-driver Genes

Analysis of driver/passenger mutations is known to be powerful approach in cancer genomics and can even be diagnostic of various cancers (Goncearenco et al., 2017; Brown et al., 2019; Tokheim and Karchin, 2019; Dietlein et al., 2020). We derived lists of recurrent driver and non-driver mutations using three computational approaches (see section "Materials and Methods"). We define driver genes as those genes, which accumulate recurrent driver mutations, but which may also possess recurrent passenger mutations (**Supplementary Tables 1**). Some genes contain only recurrent passenger mutations with frequencies comparable to driver genes; in this study, we defined a non-driver gene operationally as a gene that only contains recurrent passenger mutations (**Supplementary Table 2**).

Final lists of operationally defined driver and non-driver genes are shown in **Supplementary Tables 1, 2** (we used the ENSEMBL IDs, as recommended by the DAVID Bioinformatics Resources web site, https://david.ncifcrf.gov/). The total numbers of driver and non-driver genes are 134 and 210, respectively. We performed pathway/keyword enrichment analyses (Luque-Baena et al., 2014; Wang et al., 2014; Soldatos et al., 2015)

TABLE 2 | Correlation between mutable motifs and the sequence context of somatic mutations in driver and non-driver genes.

Group of genes	Test	Number of G:C mutations	AID/G:C	Pol η/G:C	Pol θ/G:C	Number of A:T mutations	Pol η/A:T	Pol θ/A:T
All genes	Ratio	137,775	1.021	1.005	1.091	145,768	0.992	1.011
	t-test		**23.4***	**7.2***	**23.0***		NSE	**15.8***
	MC test		<0.001	<0.001	<0.001		1	<0.001
Drivers	Ratio	4,246	1.107	1.001	1.007	3,918	0.98	1.032
	t-test		**20.0***	NSE	NSE		NSE	**7.8***
	MC test		<0.001	0.346	0.037		1	<0.001
Non-drivers	Ratio	3,553	1.079	1.059	1.057	2,793	0.995	1.045
	t-test		**14.2***	**13.8***	**11.7***		NSE	**8.9***
	MC test		<0.001	<0.001	<0.001		0.874	<0.001

"Ratio" is the mean weight of mutated sites divided by the mean weight of non-mutated sites.
NSE (no significant excess) indicates the absence of a significant excess of mutations in mutable motifs suggesting there to be no association between mutagenesis and mutable motifs. The significance of any excess was measured using the Student t and Monte Carlo (MC) tests. The asterisk (*) denotes that the corresponding P < 0.01; this is a conservative estimate of the critical overall value of the t-test having allowed for multiple testing by means of the Bonferroni correction (5 comparisons).

using the DAVID web site (Jiao et al., 2012). Results are shown in the **Supplementary Table 6**. Keywords "methylation," "nuclear chromatin," and numerous pathways/terms associated with various types of cancer are consistent with properties of GCB lymphomas (Green et al., 2015; Rogozin et al., 2016). The KEGG pathway "pathways in cancer" ($P = 0.025$) is another important descriptor of the driver gene list (**Supplementary Table 6**). In general, the driver gene set appears to be highly informative and contains many features expected for cancer-related genes (Green et al., 2015) (**Supplementary Table 6**). By contrast, analysis of non-driver genes yielded only a few significant results with no obvious functional associations with cancer (**Supplementary Table 6**).

There is a significant association of the AID mutable motif with somatic mutations in all genes, as well as in driver and non-driver genes (**Table 2**) suggesting that AID plays an important role in mutagenesis in cancer genomes; there are several pathways that can explain this process (**Figure 4**). Analysis of association between pols η and θ mutable motifs and somatic mutations detected a difference between driver and non-driver genes: mutable motifs in G:C pairs of pols η and θ correlate with somatic mutations in non-driver genes only. There was no correlation with pol η mutations at A:T pairs, whereas the pattern of somatic mutation correlated with pol θ at A:T sites both in driver and non-driver genes (**Table 2**). These observations indicate that the contribution of different pathways of generation of mutations in cancers (**Figure 4**) is distinct for AID, pols η and pol θ.

Another important feature of driver genes is a higher frequency of mutations at G:C nucleotides (4,246 and 3,918 in G:C and A:T, accordingly) compared to all other genes (137,775 – 4,246 = 133,529 and 145,786 – 3,918 = 141,868 in G:C and A:T, accordingly, **Table 2**) ($P < 0.0001$ according to the two-tailed Fisher's exact test).[3] A similar trend was observed for non-driver genes (**Table 2**, $P < 0.0001$). This may be explained by a leading role for AID/APOBEC enzyme(s) that preferentially participate in mutagenesis pathways in G:C nucleotides; AID is one such enzyme (**Figure 4**).

[3]www.graphpad.com/quickcalcs/contingency1.cfm

Patient-Specific Analysis of Somatic Mutations and Methylation

We analyzed the significance of association between AID/pol mutable motifs and the sequence context of somatic mutations for each sample (**Supplementary Table 7**). The results suggested that all studied samples have a significant association between AID/pols mutable motifs and mutation (**Supplementary Table 7**). The *t*-test values were similar to those in the merged dataset (**Supplementary Table 7** and **Table 2**). For example, *t*-test values for AID vary from 4.2 to 35.8 (**Supplementary Table 7**), this value for the merged dataset was estimated as 23.4 (**Table 2**).

We also analyzed the level of methylation in CpG sites for driver and non-driver genes for each sample separately. We derived profiles of methylation (methylation levels, positions, and chromosomes) across driver and non-driver genes separately. After that, pairwise correlation coefficients (Pearson's linear correlation coefficients CC) were estimated across all studied samples. All correlation coefficients were larger than 0.9 (the significance level < 0.001). Plots of pairwise CC values are shown in the **Supplementary Figures 6, 7**; these plots appear homogeneous (no blocks of "high" and "low" CC values that are adjacent in data matrices) (**Supplementary Figures 6, 7**).

These results suggest that studied patient-specific associations of mutable motifs with somatic mutations as well as patterns of methylation are homogeneous for driver and non-driver genes. Thus, we pooled patient-specific samples into merged datasets of somatic mutations and methylation profiles. This procedure is especially important for the analysis of small datasets that will be described below.

Analysis of DNA Methylation Patterns of Driver and Non-driver Genes Using Weight Matrices

The average methylation level of driver and non-driver genes was found to be approximately the same: ∼78% for both sets of genes (all CpG dinucleotides in driver and non-driver genes were computationally analyzed using the MALY-DE dataset). Analysis of methylation in mutable motifs was performed using the

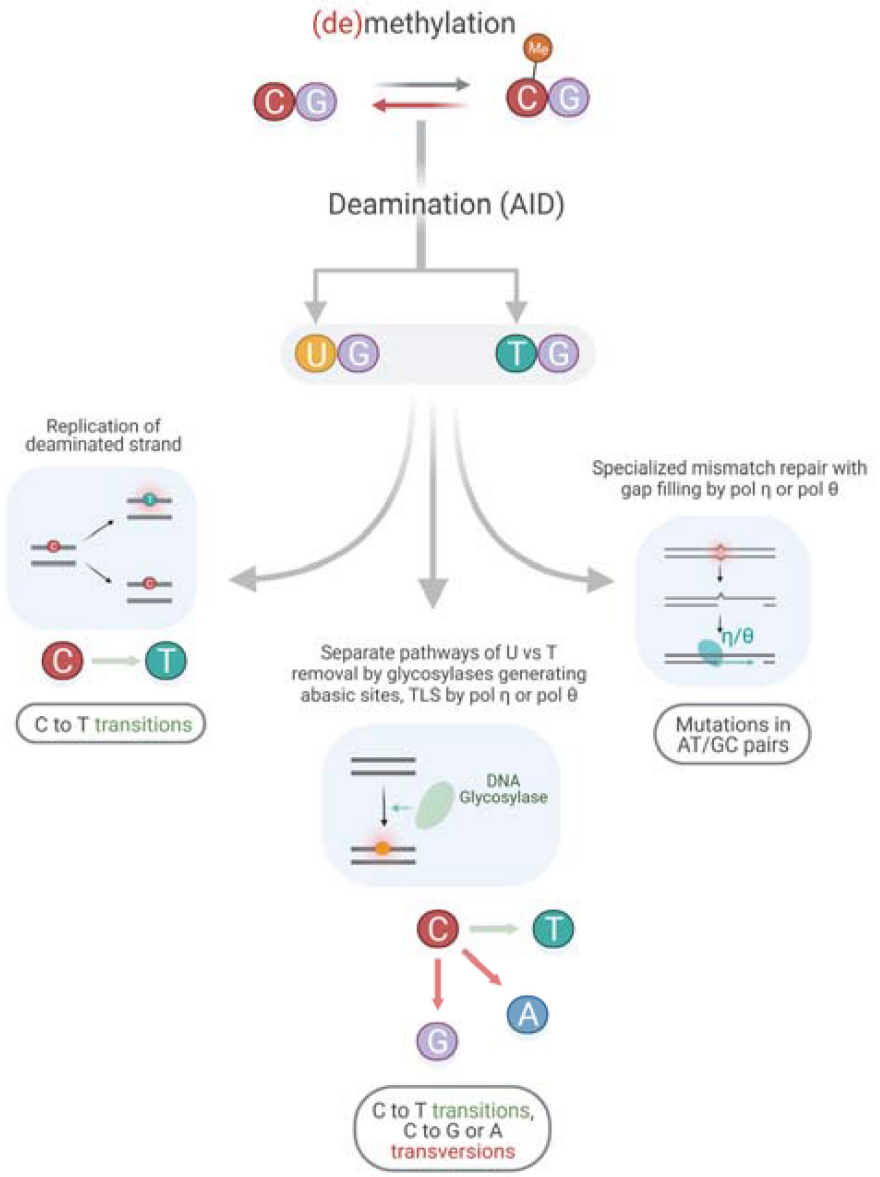


FIGURE 4 | Putative mechanism of an interplay between AID and TLS polymerases.

threshold methylation values 25 and 75%. These two values were chosen arbitrarily, values of 75 (close to the average methylation level) and higher correspond to heavily methylated CpG sites. The value 25% and smaller correspond to CpG sites that are close to the unmethylated state. Thus, values 25 and 75% reflect a dramatically different methylation status for CpG sites in the studied sets of genes (**Figure 2**).

Let us illustrate the logic of combined analysis of methylation in mutable motifs using an example from **Table 3A**. For the set of driver genes and the threshold methylation value = 25%, average weights of AID mutable motifs for subsets of CpG sites with methylation values smaller than and greater than the threshold = 25% were 57.8 and 56.4, respectively. The ratio of these values is 1.025 (57.8/56.4 = 1.025) and is shown in **Table 3A**. This difference is statistically significant, albeit not dramatically so (**Table 3**). Average weights of AID mutable motifs for non-driver genes below and above the threshold = 25% are 57.7 and 56.2, accordingly. The ratio is 1.027, and this difference is also statistically significant (**Table 3**). These results suggest that a high frequency of AID-mutable motifs is associated with lower methylation levels in driver and non-driver genes. For pol η and θ, no significant differences were detected for both driver and non-driver genes (**Table 3A**), suggesting that the global level of methylation of CpG sites of driver and non-driver genes for the threshold methylation level = 25% may not interfere with mutagenesis by pols η and θ.

TABLE 3 | Analysis of methylation in CpG sites that overlap with pols η and θ mutable motifs.

Group of genes	Number of CpG sites below and above the threshold	Tests	AID	Pol η	Pol θ
A. Levels of methylation in CpG sites that overlap with mutable motifs, with the threshold value = 25%					
Driver		Ratio	1.025	0.997	0.994
	2,867	*t*-test	**3.2***	NSE	NSE
	149,480	MC test	<0.001	0.772	0.95
Non-driver		Ratio	1.027	0.993	0.985
	5,558	*t*-test	**5.4***	NSE	NSE
	239,220	MC test	<0.001	0.989	0.989
B. Levels of methylation in CpG sites that overlap with mutable motifs, with the threshold value = 75%					
Driver		Ratio	1.004	1.009	1.021
	96,917	*t*-test	NSE	**7.9***	**20.4***
	51,290	MC test	0.433	<0.001	<0.001
Non-driver		Ratio	1.007	1.009	1.023
	155,205	*t*-test	**4.5***	**9.8***	**28.6***
	89,573	MC test	<0.001	<0.001	<0.001

*"Ratio" is the mean weight of mutable motifs in CpG sites with methylation values below (or above) the threshold divided by the mean weight of mutable motifs in CpG sites with methylation values above (or below) the threshold (25 or 75%, respectively) (a schematic representation of this analysis is shown in **Figure 2**. NSE (no significant excess) indicates the absence of any significant excess suggesting there to be no association between methylation and mutable motifs. The significance of an excess was measured using the Student t and Monte Carlo (MC) tests. The asterisk (*) denotes that the corresponding P < 0.01; this is a conservative estimate of the critical overall value of the t-test having allowed for multiple testing by means of the Bonferroni correction (3 comparisons).*

For the threshold methylation value = 75%, we observed to some extent the opposite trend. For example, the average weights of AID-mutable motifs for driver genes greater and smaller than 75% were 56.9 and 56.7, respectively. The ratio of these values is 1.004 (56.9/56.7 = 1.004) (**Table 3B**). This difference is not statistically significant (**Table 3B**). The ratio is also relatively low for the non-driver gene set although it is significant (**Table 3B**). Mutable motifs for both studied DNA polymerases appear to be associated with the methylation level for this threshold (heavily methylated CpG sites). These results suggest that the global level of methylation in driver genes for the heavily methylated positions may be affected by pol η and pol θ transactions on methylated CpG's but not AID transactions. The methylation levels of non-driver genes may be affected by all studied enzymes (**Table 3B**).

Analysis of Somatic Mutations in CpG Sites in Driver and Non-driver Genes

We analyzed the level of methylation in CpG sites that coincide with positions of somatic mutation. This dataset is much smaller compared to all methylated CpG's (the previous section). It should be noted that the studied sets are small. However, they are still amenable to statistical analysis using the threshold = 75% (**Table 4**, heavily methylated CpG sites). Unfortunately, the number of mutations for the threshold = 25% (CpG sites that are close to the unmethylated state) was too small for statistical analysis: the number of mutated sites with methylation levels below 25% is 0 and 3 for driver and non-driver genes, accordingly. Thus, we did not use the threshold value 25% but instead used the threshold value 75% only.

TABLE 4 | Levels of methylation in positions of somatic mutation in CpG sites, the threshold value = 75%.

Group of genes	Number of mutations in CpGs sites above and below the threshold	Tests	AID	Pol η	Pol θ
Driver		Ratio	1.111	1.136	1.046
	249	*t*-test	**2.9***	**7.8***	NSE
	52	MC test	0.004	<0.001	0.009
Non-driver		Ratio	1.015	1.125	1.061
	390	*t*-test	NSE	**7.3***	**3.7***
	264	MC test	0.222	<0.001	<0.001

*"Ratio" is the mean weight of mutated CpG sites above the methylation threshold divided by the mean weight of mutated sites below the threshold (a schematic representation of this analysis is shown in the **Supplementary Figure 8**). NSE (no significant excess) indicates the absence of any significant differences between these sets suggesting there to be no association between mutagenesis and motifs in the CpG sites. The significance of any excess was measured using the Student t and Monte Carlo (MC) tests. The asterisk (*) denotes that the corresponding P < 0.01; this is a conservative estimate of the critical overall value of the t-test having allowed for multiple testing by means of the Bonferroni correction (3 comparisons).*

The first result obtained is that the fraction of mutated CpG sites with methylation values below the threshold 75% is dramatically different for driver genes (52/(52+249) = 0.17, **Table 4**, the second column) and non-driver genes (0.40, **Table 4**, the second column). This difference is statistically significant ($P < 0.0001$ according to the two-tailed Fisher's exact test). Thus, CpG sites with somatic mutations in driver genes tend to have higher methylation values compared to non-driver genes.

The second interesting result is the significant correlation of AID, pol η and pol θ with mutation positions having a lower methylation level (below 75%) (**Table 4**). The correlation of the AID motif presence and mutation is more pronounced for driver genes, indicating that AID-induced mutagenesis is likely to be associated with heavily methylated CpG dinucleotides. Pol η has a role in CpG mutagenesis for both sets of genes whereas pol θ is likely to be largely involved in the mutagenesis of non-driver genes (**Table 4**). Thus, it is likely that methylation levels influence mutagenesis pathways in CpG sites through the action of all the studied enzymes, although the individual impact of studied enzymes may be different for driver and non-driver genes (for example, AID, **Table 4**). It is likely to depend on various factors including gene expression. This will be discussed in the next section.

Analysis of Expression of Driver and Non-driver Genes

We analyzed the expression levels (FPMK values) for both sets of genes (**Supplementary Tables 1, 2**). Analysis of mean and variance (**Figure 5** and **Supplementary Table 8**) suggested

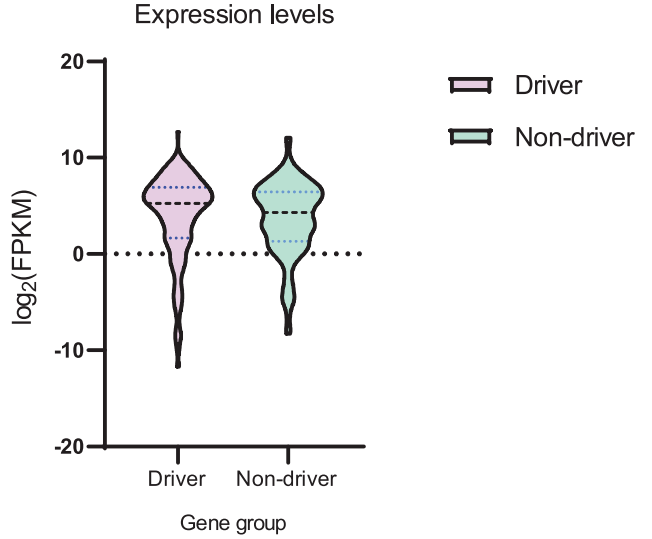


FIGURE 5 | Violin plot of mRNA expression (FPKM values) for sets of driver and non-driver genes. Log_2 transformation was used.

that mean values were not substantially different. However, the variance of expression values observed in the set of driver genes was larger as compared to the set of non-driver genes (**Supplementary Table 8**). The difference between mean values (**Supplementary Table 8**) was not statistically significant (*t*-test *P* value = 0.086), whereas the difference between variance values (**Supplementary Table 8**) was statistically significant (*F*-test *P* value = 0.007).

DISCUSSION

Some results of this study seem to be counterintuitive. For example, the AID mutable motif would appear to correlate with the context of somatic mutations in heavily methylated CpG's for driver genes only (**Table 4**). It is hard to determine the factors that are responsible for this difference. For example, variability of gene expression is significantly higher for driver genes (**Figure 5**). This may be associated with the differential regulation of expression of driver genes in different patients or methylation levels. Copy number variation of driver genes (Loohuis et al., 2014; Cheng et al., 2016) may cause problems for precise estimates of CpG methylation levels.

AID and DNA polymerases η/θ are known to participate in somatic hypermutation of immunoglobulin genes (Matsuda et al., 2001; Casali et al., 2006; Neuberger and Rada, 2007; Bhattacharya et al., 2008). In addition, it has been suggested that AID and pol η are likely to contribute to a lowering methylation levels of CpG dinucleotides in cancer cells (Rogozin et al., 2018b). Thus, we focused this study on AID and pols η/θ employing the weight matrix technique and mutation/methylation profiles. Our results suggest that AID and pols η/θ combine to generate footprint mutations in B-cell derived lymphomas and other cancers. It was reported that methylation substantially reduces the rate of APOBEC-induced mutations in CpG dinucleotides (Seplyarskiy et al., 2016). For this reason, we did not include other members of the AID/APOBEC superfamily in the current study.

The advantage of the weight matrix approach is that it is a unified computational technique that allows an accurate and objective comparison of the mutational contribution of various mutator enzymes under the same experimental conditions and for the same datasets. We confirm that while the mutational footprints of DNA polymerases η and θ are prominent in some cancers, mutable motifs characteristic of the humoral immune response somatic hypermutation machine, AID, is likely to be the most widespread feature of somatic mutational spectra attributed to any enzyme in cancer genomes (Rogozin et al., 2018b, 2019). It is important to note that the suggested technique does not depend on expert opinion as to the exact consensus sequences, and therefore objectively represents mutable motifs.

We derived matrices for A:T and G:C residues. However, the ratio of A:T to G:C mutations is variable (**Supplementary Figure 1**). For example, it known that Pol η mutates G residues at a lower frequency than A residues. However, two matrices (G:C and A:T residues, **Figure 1**) for the two motifs were used independently (**Figure 3**). We would like to develop a probabilistic model that integrates two matrices in one model. However, this approach has never been attempted before in this context and would require further investigation.

It is not possible to delineate the exact mechanism of the interplay between AID and DNA polymerases. It may be replication of the deaminated strand, separate pathways of U vs. T removal by glycosylases generating abasic sites followed by TLS by pol η or pol θ, and/or specialized mismatch repair with gap filling by pol η or pol θ (**Figure 4**) (Pilzecker and Jacobs, 2019). Unfortunately, precise mechanisms have not been clearly defined even for mutagenesis of immunoglobulin genes, with attempts to define those mechanisms having been ongoing for over 20 years.

A high rate of prediction errors for many types of cancer (**Supplementary Table 5**) is likely to be due to the small mutational spectra available for DNA polymerase η and θ (**Supplementary Figures 2, 3**). Larger sets of mutations are likely to improve the quality of prediction. We can nevertheless infer that some types of cancer, including GCB lymphomas, do not have a noticeable rate of false positives (**Supplementary Table 5**). We applied all weight matrices to study mutable motifs and methylation in the MALY-DE datasets and demonstrated that mutable motifs correlate with CpG dinucleotides and their methylation status. Another methodological problem is the small number of MALY-DE samples (26 samples) which may cause problems for the prediction of driver and passenger mutations. These problems are one of several possible explanations as to why differences between driver and non-driver genes are subtle (albeit significant) (**Tables 2-4**). However, it is likely that these differences are responsible for the major difference observed between the expression of driver and non-driver genes (**Figure 5**). The much large variance observed for driver genes may be the result of greater (de)methylation of driver gene sequences causing substantial variability of mRNA expression across patients (**Figure 5**).

Sophisticated classification approaches (prediction of mutational signatures) have been developed to extract the most prominent signatures from a complex mix of mutational spectra resulting from the action of a variety of mutagens, both exogenous and endogenous, operating during tumor evolution (Petljak and Alexandrov, 2016; Rahbari et al., 2016; Goncearenco et al., 2017; Rogozin et al., 2018c; Alexandrov et al., 2020). Both driver and passenger mutations have been used in the analysis without any attempt to analyze them separately. In this study, we analyzed driver and non-driver genes separately and detected significant differences in the relationship between mutable motifs and mutations with the methylation/demethylation status of driver and non-driver genes (**Tables 3** and **4**). It is not that easy to interpret these differences because the role of methylated CpG dinucleotides in exons is not yet fully understood (Neri et al., 2017). It has been suggested that changes in intragenic DNA methylation is important in several human diseases including syndromic and sporadic forms of various neurological disorders that involve methylation defects, including Rett syndrome, Prader–Willi and Angelman syndromes, and others, suggesting that the differential (de)methylation of genes may underpin one aspect of various neurological disorders (Dunaway et al., 2016; Rogozin et al., 2018a; Scandaglia and Barco, 2019). Such differential methylation may be caused by differences in (de)methylation processes in somatic/germ cells (Shanak and Helms, 2020). Moreover, several studies of likely deleterious mutations have observed that genes controlling methylation status, chromatin accessibility or remodeling (and hence gene expression) are enriched for genes with recurrent mutations (Geschwind and State, 2015; Sanders et al., 2015; Geisheker et al., 2017).

The difference in AID and polymerase properties (**Tables 3, 4**) for driver and non-driver genes is consistent with the participation of different mechanisms of mutagenesis and (de)methylation processes (**Figure 4**) on non-methylated and methylated DNA. The observed differences between driver and non-driver genes associated with somatic mutations in driver genes (**Tables 3, 4**) are likely to cause changes in gene expression (**Figure 5**) that then trigger cancer initiation and/or progression. This is not surprising if we consider that chromatin modification pathways (**Supplementary Table 6**) as well as the observed changes in CpG methylation levels (**Tables 3, 4**) are likely to cause changes in the expression levels of driver genes that could affect both cancer initiation and/or progression.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://dcc.icgc.org/projects/MALY-DE; https://cancer.sanger.ac.uk.

AUTHOR CONTRIBUTIONS

IBR, AR-L, KT, KC-C, AL, LP, and ES: formal analysis. All authors: investigation.

FUNDING

This work was supported by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health (IBR), RCMI grant U54 MD007600 (National Institute on Minority Health and Health Disparities) from the National Institutes of Health (AR-L), NE DHHS LB506, grant 2017-48 (YIP) and Qiagen, Inc. through a License Agreement with Cardiff University (DNC). YIP was also partially supported by the Russian Science Foundation grant 20-15-00081, and the Fred & Pamela Buffett Cancer Center Support Grant from the National Cancer Institute under award number P30 CA072720. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. ARP and KT were supported by the Department of Pathology and Molecular Medicine, Queen's University, Canada. ARP is the recipient of a Senior Canada Research Chair in Computational Biology and Biophysics and a Senior Investigator Award from the Ontario Institute of Cancer Research, Canada.

ACKNOWLEDGMENTS

ARP and KT thank Alexander Goncearenco and Jiaying You for help with data acquisition.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.671866/full#supplementary-material

REFERENCES

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421.

Alexandrov, L. B., and Stratton, M. R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* 24, 52–60. doi: 10.1016/j.gde.2013.11.014

Alsøe, L., Sarno, A., Carracedo, S., Domanska, D., Dingler, F., Lirussi, L., et al. (2017). Uracil accumulation and mutagenesis dominated by cytosine deamination in CpG dinucleotides in mice lacking UNG and SMUG1. *Sci. Rep.* 7:7199.

Arana, M. E., Seki, M., Wood, R. D., Rogozin, I. B., and Kunkel, T. A. (2008). Low-fidelity DNA synthesis by human DNA polymerase theta. *Nucleic Acids Res.* 36, 3847–3856. doi: 10.1093/nar/gkn310

Bhattacharya, P., Grigera, F., Rogozin, I. B., McCarty, T., Morse, H. C. III, and Kenter, A. L. (2008). Identification of murine B cell lines that undergo somatic hypermutation focused to A:T and G:C residues. *Eur. J. Immunol.* 38, 227–239. doi: 10.1002/eji.200737664

Brambati, A., Barry, R. M., and Sfeir, A. (2020). DNA polymerase theta (Polθ) - an error-prone polymerase necessary for genome stability. *Curr. Opin. Genet. Dev.* 60, 119–126. doi: 10.1016/j.gde.2020.02.017

Brinkman, A. B., Nik-Zainal, S., Simmer, F., Rodriguez-Gonzalez, F. G., Smid, M., Alexandrov, L. B., et al. (2019). Partially methylated domains are hypervariable in breast cancer and fuel widespread CpG island hypermethylation. *Nat. Commun.* 10:1749.

Brown, A. L., Li, M., Goncearenco, A., and Panchenko, A. R. (2019). Finding driver mutations in cancer: elucidating the role of background mutational processes. *PLoS Comput. Biol.* 15:e1006981. doi: 10.1371/journal.pcbi.1006981

Casali, P., Pal, Z., Xu, Z., and Zan, H. (2006). DNA repair in antibody somatic hypermutation. *Trends Immunol.* 27, 313–321. doi: 10.1016/j.it.2006.05.001

Cheng, F., Zhao, J., and Zhao, Z. (2016). Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Brief. Bioinform.* 17, 642–656. doi: 10.1093/bib/bbv068

Cooper, D. N., and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Hum Genet* 78, 151–155. doi: 10.1007/bf00278187

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* 274, 775–780. doi: 10.1038/274775a0

Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D., et al. (2020). Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* 52, 208–218. doi: 10.1038/s41588-019-0572-y

Dörner, T., and Lipsky, P. E. (2001). Smaller role for pol η? *Nat. Immunol.* 2, 982–984. doi: 10.1038/ni1101-982

Dunaway, K. W., Islam, M. S., Coulson, R. L., Lopez, S. J., Vogel Ciernia, A., Chu, R. G., et al. (2016). Cumulative impact of polychlorinated biphenyl and large chromosomal duplications on DNA methylation, chromatin, and expression of autism candidate genes. *Cell Rep.* 17, 3035–3048. doi: 10.1016/j.celrep.2016.11.058

Geisheker, M. R., Heymann, G., Wang, T., Coe, B. P., Turner, T. N., Stessman, H. A. F., et al. (2017). Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* 20, 1043–1051. doi: 10.1038/nn.4589

Gelfand, M. S. (1995). Prediction of function in DNA sequence analysis. *J. Comput. Biol.* 2, 87–115. doi: 10.1089/cmb.1995.2.87

Geschwind, D. H., and State, M. W. (2015). Gene hunting in autism spectrum disorder: on the path to precision medicine. *Lancet. Neurol.* 14, 1109–1120. doi: 10.1016/s1474-4422(15)00044-7

Goncearenco, A., Rager, S. L., Li, M., Sang, Q. X., Rogozin, I. B., and Panchenko, A. R. (2017). Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Res.* 45, W514–W522.

Granadillo Rodriguez, M., Flath, B., and Chelico, L. (2020). The interesting relationship between APOBEC3 deoxycytidine deaminases and cancer: a long road ahead. *Open Biol.* 10:200188. doi: 10.1098/rsob.200188

Green, M. R., Kihira, S., Liu, C. L., Nair, R. V., Salari, R., Gentles, A. J., et al. (2015). Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1116–E1125.

Howe, E. A., Sinha, R., Schlauch, D., and Quackenbush, J. (2011). RNA-Seq analysis in MeV. *Bioinformatics* 27, 3209–3210. doi: 10.1093/bioinformatics/btr490

Islam, S. M. A., and Alexandrov, L. B. (2021). Bioinformatic methods to identify mutational signatures in cancer. *Methods Mol. Biol.* 2185, 447–473. doi: 10.1007/978-1-0716-0810-4_28

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28, 1805–1806. doi: 10.1093/bioinformatics/bts251

Loohuis, L. O., Witzel, A., and Mishra, B. (2014). Improving detection of driver genes: power-law null model of copy number variation in cancer. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 1260–1263. doi: 10.1109/tcbb.2014.2351805

Luque-Baena, R. M., Urda, D., Gonzalo Claros, M., Franco, L., and Jerez, J. M. (2014). Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords. *J. Biomed. Inform.* 49, 32–44. doi: 10.1016/j.jbi.2014.01.006

Martomo, S. A., Saribasak, H., Yokoi, M., Hanaoka, F., and Gearhart, P. J. (2008). Reevaluation of the role of DNA polymerase θ in somatic hypermutation of immunoglobulin genes. *DNA Repair* 7, 1603–1608. doi: 10.1016/j.dnarep.2008.04.002

Matsuda, T., Bebenek, K., Masutani, C., Rogozin, I. B., Hanaoka, F., and Kunkel, T. A. (2001). Error rate and specificity of human and murine DNA polymerase eta. *J. Mol. Biol.* 312, 335–346.

Mayorov, V. I., Rogozin, I. B., Adkison, L. R., and Gearhart, P. J. (2005). DNA polymerase eta contributes to strand bias of mutations of A versus T in immunoglobulin genes. *J. Immunol.* 174, 7781–7786. doi: 10.4049/jimmunol.174.12.7781

Milstein, C., Neuberger, M. S., and Staden, R. (1998). Both DNA strands of antibody genes are hypermutation targets. *Proc. Natl. Acad. Sci. U.S.A.* 95, 8791–8794. doi: 10.1073/pnas.95.15.8791

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., et al. (2017). Intragenic DNA methylation prevents spurious transcription initiation. *Nature* 543, 72–77. doi: 10.1038/nature21373

Neuberger, M. S., and Rada, C. (2007). Somatic hypermutation: activation-induced deaminase for C/G followed by polymerase η for A/T. *J. Exp. Med.* 204, 7–10. doi: 10.1084/jem.20062409

Oliver, J., Garcia-Aranda, M., Chaves, P., Alba, E., Cobo-Dols, M., Onieva, J. L., et al. (2021). Emerging noninvasive methylation biomarkers of cancer prognosis and drug response prediction. *Semin. Cancer. Biol.* doi: 10.1016/j.semcancer.2021.03.012

Pavlov, Y. I., Rogozin, I. B., Galkin, A. P., Aksenova, A. Y., Hanaoka, F., Rada, C., et al. (2002). Correlation of somatic hypermutation specificity and A-T base pair substitution errors by DNA polymerase η during copying of a mouse immunoglobulin κ light chain transgene. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9954–9959. doi: 10.1073/pnas.152126799

Petljak, M., and Alexandrov, L. B. (2016). Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis* 37, 531–540. doi: 10.1093/carcin/bgw055

Pham, P., Calabrese, P., Park, S. J., and Goodman, M. F. (2011). Analysis of a single-stranded DNA-scanning process in which activation-induced deoxycytidine deaminase (AID) deaminates C to U haphazardly and inefficiently to ensure mutational diversity. *J. Biol. Chem.* 286, 24931–24942. doi: 10.1074/jbc.m111.241208

Pilzecker, B., and Jacobs, H. (2019). Mutating for good: DNA damage responses during somatic hypermutation. *Front. Immunol.* 10:438. doi: 10.3389/fimmu.2019.00438

Rahbari, R., Wuster, A., Lindsay, S. J., Hardwick, R. J., Alexandrov, L. B., Turki, S. A., et al. (2016). Timing, rates and spectra of human germline mutation. *Nat. Genet.* 48, 126–133. doi: 10.1038/ng.3469

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., et al. (2000). Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the hyper-IgM syndrome (HIGM2). *Cell* 102, 565–575. doi: 10.1016/s0092-8674(00)00079-9

Roberts, S. A., and Gordenin, D. A. (2014). Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* 14, 786–800. doi: 10.1038/nrc3816

Roberts, S. A., Lawrence, M. S., Klimczak, L. J., Grimm, S. A., Fargo, D., Stojanov, P., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* 45, 970–976. doi: 10.1038/ng.2702

Rogozin, I. B., and Diaz, M. (2004). Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J. Immunol.* 172, 3382–3384. doi: 10.4049/jimmunol.172.6.3382

Rogozin, I. B., Gertz, E. M., Baranov, P. V., Poliakov, E., and Schaffer, A. A. (2018a). Genome-wide changes in protein translation efficiency are associated with autism. *Genome Biol. Evol.* 10, 1902–1919. doi: 10.1093/gbe/evy146

Rogozin, I. B., Goncearenco, A., Lada, A. G., De, S., Yurchenko, V., Nudelman, G., et al. (2018b). DNA polymerase η mutational signatures are found in a variety of different types of cancer. *Cell Cycle* 17, 348–355. doi: 10.1080/15384101.2017.1404208

Rogozin, I. B., Lada, A. G., Goncearenco, A., Green, M. R., De, S., Nudelman, G., et al. (2016). Activation induced deaminase mutational signature overlaps with CpG methylation sites in follicular lymphoma and other cancers. *Sci. Rep.* 6:38133.

Rogozin, I. B., Pavlov, Y. I., Bebenek, K., Matsuda, T., and Kunkel, T. A. (2001). Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat. Immunol.* 2, 530–536. doi: 10.1038/88732

Rogozin, I. B., Pavlov, Y. I., Goncearenco, A., De, S., Lada, A. G., Poliakov, E., et al. (2018c). Mutational signatures and mutable motifs in cancer genomes. *Brief. Bioinform.* 19, 1085–1101.

Rogozin, I. B., Roche-Lima, A., Lada, A. G., Belinky, F., Sidorenko, I. A., Glazko, G. V., et al. (2019). Nucleotide weight matrices reveal ubiquitous mutational footprints of AID/APOBEC deaminases in human cancer genomes. *Cancers* 11:211. doi: 10.3390/cancers11020211

Sanders, S. J., He, X., Willsey, A. J., Ercan-Sencicek, A. G., Samocha, K. E., Cicek, A. E., et al. (2015). Insights into autism spectrum disorder genomic architecture and biology from 71 risk loci. *Neuron* 87, 1215–1233.

Scandaglia, M., and Barco, A. (2019). Contribution of spurious transcription to intellectual disability disorders. *J. Med. Genet.* 56, 491–498. doi: 10.1136/jmedgenet-2018-105668

Seplyarskiy, V. B., Soldatov, R. A., Popadin, K. Y., Antonarakis, S. E., Bazykin, G. A., and Nikolaev, S. I. (2016). APOBEC-induced mutations in human cancers are strongly enriched on the lagging DNA strand during replication. *Genome Res.* 26, 174–182. doi: 10.1101/gr.197046.115

Shanak, S., and Helms, V. (2020). DNA methylation and the core pluripotency network. *Dev. Biol.* 464, 145–160. doi: 10.1016/j.ydbio.2020.06.001

Sina, A. A., Carrascosa, L. G., Liang, Z., Grewal, Y. S., Wardiana, A., Shiddiky, M. J. A., et al. (2018). Epigenetically reprogrammed methylation landscape drives the DNA self-assembly and serves as a universal cancer biomarker. *Nat. Commun.* 9:4915.

Soldatos, T. G., Perdigao, N., Brown, N. P., Sabir, K. S., and O'Donoghue, S. I. (2015). How to learn about gene function: text-mining or ontologies? *Methods* 74, 3–15. doi: 10.1016/j.ymeth.2014.07.004

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12, 505–519. doi: 10.1093/nar/12.1part2.505

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature* 458, 719–724.

Swanton, C., McGranahan, N., Starrett, G. J., and Harris, R. S. (2015). APOBEC enzymes: mutagenic fuel for cancer evolution and heterogeneity. *Cancer Discov.* 5, 704–712. doi: 10.1158/2159-8290.cd-15-0344

Tokheim, C., and Karchin, R. (2019). CHASMplus reveals the scope of somatic missense mutations driving human cancers. *Cell Syst.* 9, 9–23. doi: 10.1016/j.cels.2019.05.005

Wang, J. H., Zhao, L. F., Lin, P., Su, X. R., Chen, S. J., Huang, L. Q., et al. (2014). GenCLiP 2.0: a web server for functional clustering of genes and construction of molecular networks based on free terms. *Bioinformatics* 30, 2534–2536. doi: 10.1093/bioinformatics/btu241

Wood, R. D., and Doublié, S. (2016). DNA polymerase θ (POLQ), double-strand break repair, and cancer. *DNA Repair* 44, 22–32. doi: 10.1016/j.dnarep.2016.05.003

Zan, H., Shima, N., Xu, Z., Al-Qahtani, A., Evinger Iii, A. J., Zhong, Y., et al. (2005). The translesion DNA polymerase θ plays a dominant role in immunoglobulin gene somatic hypermutation. *EMBO J.* 24, 3757–3769. doi: 10.1038/sj.emboj.7600833

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 31 May 2021
doi: 10.3389/fgene.2021.662468

Hypoxia-Induced miR-148a Downregulation Contributes to Poor Survival in Colorectal Cancer

Stepan Nersisyan[1], Alexei Galatenko[1,2], Milena Chekova[1] and Alexander Tonevitsky[1]**

[1] Faculty of Biology and Biotechnology, HSE University, Moscow, Russia, [2] Faculty of Mechanics and Mathematics, Lomonosov Moscow State University, Moscow, Russia

Hypoxia is an extensively investigated condition due to its contribution to various pathophysiological processes including cancer progression and metastasis formation. MicroRNAs (miRNAs) are well-known post-transcriptional gene expression regulators. However, their contribution to molecular response to hypoxia is highly dependent on cell/tissue types and causes of hypoxia. One of the most important examples is colorectal cancer, where no consensus on hypoxia-regulated miRNAs has been reached so far. In this work, we applied integrated mRNA and small RNA sequencing, followed by bioinformatics analysis, to study the landscape of hypoxia-induced miRNA and mRNA expression alterations in human colorectal cancer cell lines (HT-29 and Caco-2). A hypoxic microenvironment was chemically modeled using two different treatments: cobalt(II) chloride and oxyquinoline. Only one miRNA, hsa-miR-210-3p, was upregulated in all experimental conditions, while there were nine differentially expressed miRNAs under both treatments within the same cell line. Further bioinformatics analysis revealed a complex hypoxia-induced regulatory network: hypoxic downregulation of hsa-miR-148a-3p led to the upregulation of its two target genes, ITGA5 and PRNP, which was shown to be a factor contributing to tumor progression and poor survival in colorectal cancer patients.

Keywords: hypoxia, cobalt chloride, oxyquinoline, colorectal cancer, miR-148a

OPEN ACCESS

Edited by:
Tatiana V. Tatarinova, University of La Verne, United States

Reviewed by:
Evgenii Chekalin, Michigan State University, United States
Yuriy L. Orlov, I.M. Sechenov First Moscow State Medical University, Russia

****Correspondence:***
Stepan Nersisyan
snersisyan@hse.ru
Alexander Tonevitsky
atonevitsky@hse.ru

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

Received: *01 February 2021*
Accepted: *21 April 2021*
Published: *31 May 2021*

Citation:
Nersisyan S, Galatenko A, Chekova M and Tonevitsky A (2021) Hypoxia-Induced miR-148a Downregulation Contributes to Poor Survival in Colorectal Cancer. Front. Genet. 12:662468. doi: 10.3389/fgene.2021.662468

INTRODUCTION

Hypoxia is involved in the pathogenesis of colorectal and other cancers and is mainly associated with tumor growth, anti-apoptosis, recurrence, and poor survival (Brahimi-Horn et al., 2007; Vaupel and Mayer, 2007; Lange et al., 2014; Nagaraju et al., 2015). Multiple hypoxia-induced mechanisms are related to the activity of miRNAs—short non-coding RNAs whose main functional activity consists in post-transcriptional gene silencing (Cai et al., 2009). Interactions between miRNAs and their target genes were shown to play crucial roles in cell–cell communications (Turchinovich et al., 2015) and the pathogenesis of multiple diseases including, but not limited to, different types of cancer (Visone and Croce, 2009; Maltseva et al., 2014; Shkurnikov et al., 2019) and viral infections (Skalsky and Cullen, 2010; Nersisyan et al., 2020a).

Currently, there is no consensus on the regulation of miRNA expression by hypoxia: the effects heavily depend on cell types and the reason for hypoxia (either naturally or chemically induced), and only one miRNA, hsa-miR-210-3p, was found to be overexpressed under hypoxic

exposure in the majority of reports (Kulshreshtha et al., 2007; Bavelloni et al., 2017; Bhandari et al., 2019). Multiple studies revealed a link between hypoxia- and cancer-induced miRNA expression change patterns: a large fraction of cancer-associated miRNAs can also be affected by hypoxia (Kulshreshtha et al., 2007; Shen et al., 2013; Panigrahi et al., 2018).

In this work, we studied the changes in miRNA and mRNA expression landscape of colorectal cancer cell lines Caco-2 and HT-29 exposed to hypoxia. Hypoxia was modeled by two different but widely used chemical agents: cobalt(II) chloride ($CoCl_2$) and oxyquinoline, which cause long-term stabilization of hypoxia-inducible factor 1 and 2 (HIF-1 and HIF-2) (Wu and Yotnda, 2011; Osipyants et al., 2017; Muñoz-Sánchez and Chánez-Cárdenas, 2019; Savyuk et al., 2020). Such an experimental setup allowed us to precisely identify hypoxia-regulated miRNAs as well as their target genes. The role of the discovered interactions in colon cancer was further studied in patients' tumors using The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) cohort (Muzny et al., 2012)[1].

MATERIALS AND METHODS

Cell Cultures and Treatments

HT-29 cells (ATCC, Manassas, VA, United States) were grown in McCoy's 5A medium, supplemented with 10% fetal bovine serum, 2 mM glutamine, 1% non-essential amino acids, penicillin (100 U/ml), and streptomycin (100 mg/mL).

Caco-2 cells were obtained from the Russian Vertebrate Cell Culture Collection (Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia). The cells were incubated under conditions for differentiation for 21 days in Eagle's minimal essential medium with 20% fetal bovine serum, 2 mM glutamine, 0.1 mM non-essential amino acids, and 0.1% penicillin–streptomycin.

All cell culture reagents were obtained from Gibco, Waltham, MA, United States. Both cell lines were maintained in a humidified atmosphere at 37°C and 5% CO_2, changing the medium every 3 days.

A fresh stock solution of 0.3 M cobalt chloride ($CoCl_2$) was prepared in water and added to the medium to obtain the desired final concentration 300 μM for 24 h. The second treatment included a fresh portion containing oxyquinoline derivative 4896-3212 (ChemDiv Research Institute, Khimki, Russia) in dimethyl sulfoxide (DMSO; 10 mM). The final concentration of oxyquinoline in the medium was 5 μM (0.5 μl of the solution in DMSO per 1 ml medium). In the control, the cells were incubated in a medium with 0.5 μl DMSO per 1 ml medium (without $CoCl_2$ or oxyquinoline). Three biological replicates were used both for the control and the treated cells.

RNA Extraction

Cells were lysed with the QIAzol Lysis Reagent (Qiagen, Hilden, Germany) for subsequent extraction of RNA using the Qiagen miRNeasy Mini Kit (Qiagen, Hilden, Germany). Nanodrop (Thermo Fisher Scientific, Waltham, MA, United States) was used to assess quality (260/280) and quantity. Total RNA samples were also QC-checked using the Agilent High Sensitivity DNA Kit (Agilent Technologies, Santa Clara, CA, United States).

Library Preparation and Sequencing

Libraries for mRNA sequencing were prepared from total RNA samples using Illumina Stranded mRNA Library Prep Kit (Illumina, San Diego, CA, United States). Each sample was sequenced on the Illumina NextSeq 550 to generate single-end 75-nucleotide reads.

Libraries for miRNA sequencing were prepared from total RNA samples using NEBNext Multiplex Small RNA Library Prep Kit for Illumina. Each sample was sequenced on the Illumina NextSeq 550 to generate single-end 50-nucleotide reads.

RNA-Seq Data Processing

The quality of FASTQ files was assessed with FastQC v0.11.9 (Babraham Bioinformatics, Cambridge, United Kingdom); three miRNA-seq replicates (control Caco-2, $CoCl_2$-treated HT-29, and oxyquinoline-treated HT-29) had not passed the quality control (i.e., two, instead of three, replicates were used for these conditions in the downstream analysis). The adapters were trimmed with cutadapt v2.10 (Martin, 2011). The trimmed mRNA-seq reads were mapped on the reference human genome (GENCODE GRCh38.p13) with STAR v2.7.5b (Dobin et al., 2013). GENCODE genome annotation (release 34) (Frankish et al., 2019) was used to generate the count matrix. The miRNA count matrix was generated by miRDeep2 v2.0.1.2 (Friedländer et al., 2012) with the use of bowtie v1.1.1 (Langmead et al., 2009) and miRBase v22.1 (Kozomara et al., 2019).

The sequencing library sizes were normalized with the trimmed mean of *M*-values (TMM) algorithm, available in edgeR v3.30.3 package (Robinson et al., 2009), with default filtering of background noise. The same package was used to generate TMM-normalized fragments per kilobase of transcript per million mapped reads (FPKM) and reads per million mapped reads (RPM) matrices for mRNA-seq and miRNA-seq data, respectively. The obtained values were $\log_2$ -transformed. For further processing, we selected only highly expressed entries by cutting off the lower 25% of genes and 50% of miRNAs according to their median FPKM/RPM values in each experimental condition (Toung et al., 2011; Zhang X. et al., 2019). The thresholding value was higher for miRNAs since miRNA expression distribution is significantly biased toward several molecules with very high expression levels (Nersisyan et al., 2020b).

Differential Expression and Enrichment Analyses

Differential expression analysis was conducted using DESeq2 v1.28.1 (Love et al., 2014); false discovery rates (FDRs) were calculated by the Benjamini–Hochberg procedure. For mRNA-seq data we performed testing of fold changes being above 1.5 using apeglm available in DESeq2 (Zhu et al., 2019), default 0.005 threshold was set on s-values. The resulting genes were uploaded

[1]https://portal.gdc.cancer.gov/TCGA-COAD

to DAVID v6.8 (Huang et al., 2009) for enrichment analysis. For miRNA-seq data, differences with FDRs below the 0.05 threshold were considered.

Prediction of miRNA Target Genes

At the first step of miRNA target prediction, we obtained the list of miRNA–gene interactions from miRDB v6.0 (Chen and Wang, 2020). The predictions were filtered according to their target scores; threshold value was set to 80. Then, we selected negatively correlated miRNA–gene pairs from the TCGA-COAD cohort (Muzny et al., 2012, see text footnote 1). Specifically, raw matched miRNA/mRNA sequencing count matrices of tumor samples (n=426) were downloaded from GDC Data Portal[2] and converted to FPKM/RPM tables with the aforementioned TMM normalization. Then, Spearman correlation was calculated for each miRNA and predicted the target mRNA; 0.05 threshold was set on p-values. Additionally, we filtered out pairs with correlation values higher than −0.3 since they may reflect interactions that are too weak, though possibly statistically significant (Mukaka, 2012; Nersisyan et al., 2020c).

Survival Analysis

Survival analysis (logrank test and Kaplan–Meier estimation) was conducted with the Python lifelines module[3]. Thresholds for defining groups of high and low gene expression were defined using the first and the third quartile of the corresponding distribution.

[2]https://portal.gdc.cancer.gov/

[3]https://zenodo.org/record/4457577

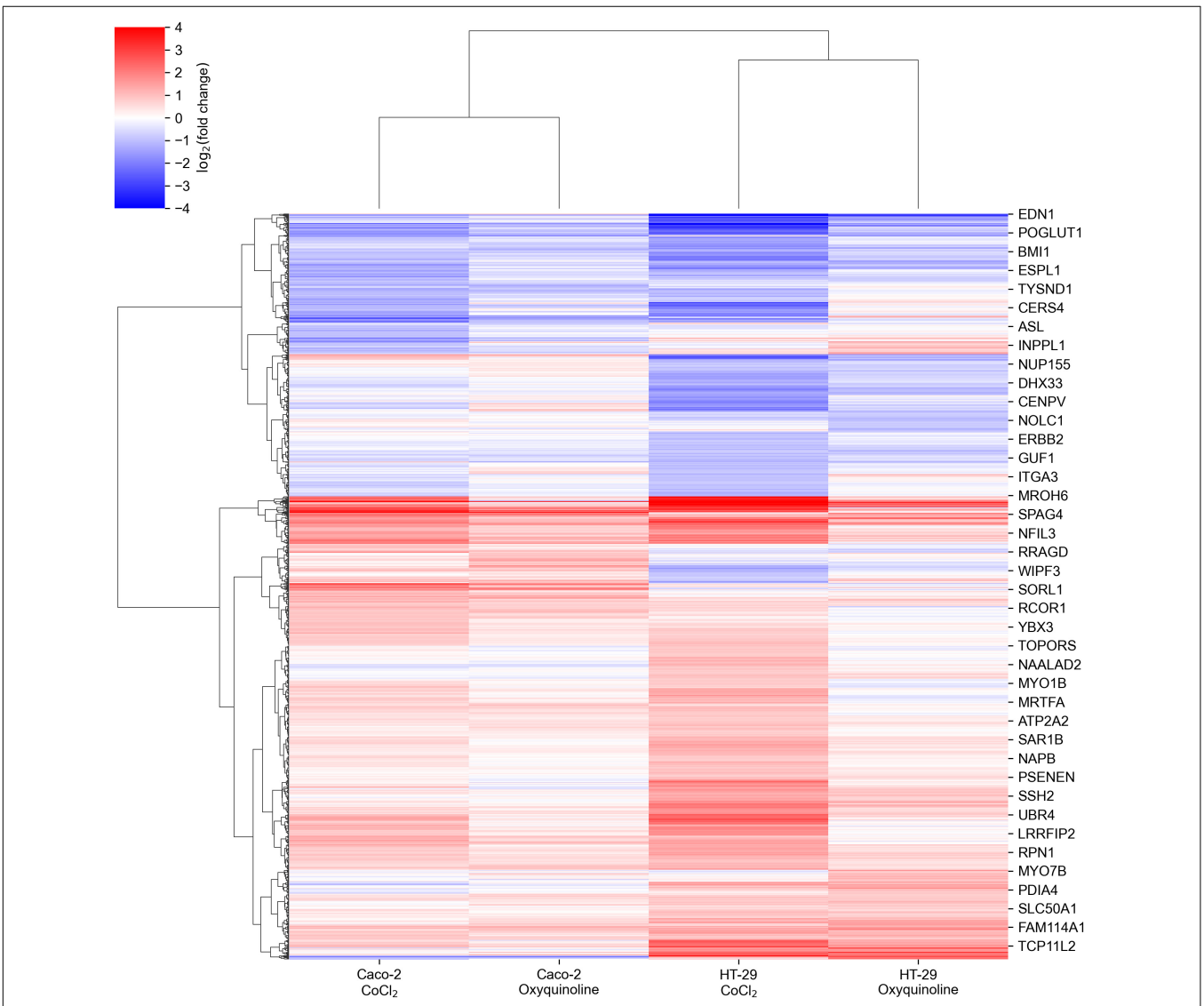


FIGURE 1 | Changes in mRNA expression levels of colorectal cell lines exposed to chemically simulated hypoxia. A gene is included in the plot in case if it is differentially expressed in at least one out of four experimental conditions.

RESULTS

Cobalt Chloride and Oxyquinoline Treatments Lead to Diverse Differential Gene Expression Landscapes

We performed mRNA sequencing of two colorectal cancer cell lines (Caco-2 and HT-29) exposed to $CoCl_2$ and oxyquinoline treatments. In general, both treatments resulted in similar patterns of differentially expressed genes within each cell line, with a notably weaker effect of oxyquinoline (**Figure 1** and **Supplementary Table 1**). Enrichment analysis of differentially expressed gene sets revealed a statistically significant activation of anaerobic glycolysis (KEGG pathway hsa00010 "glycolysis/gluconeogenesis") and HIF-1 signaling pathway (KEGG pathway hsa04066 "HIF-1 signaling pathway") in all experimental conditions, confirming the successive simulation of hypoxia. Interestingly, the level of HIF1A mRNA was increased in both treatments of HT-29 (2.6 folds for $CoCl_2$ and 1.6 folds for oxyquinoline), while the increase was insignificant for Caco-2 (1.3 folds for $CoCl_2$ and 1.06 folds for oxyquinoline). Such behavior was also reported in previous studies: while HIF-1 is mainly regulated *via* post-translational modifications, in some cell lines hypoxia may upregulate HIF1A at the transcriptional level (Catron et al., 2001; BelAiba et al., 2007).

The major difference between treatments by two chemical agents consisted in a strong upregulation of genes encoding ribosomal proteins and genes involved in the major histocompatibility (MHC) class I antigen presentation pathway in response to $CoCl_2$. This included almost all small and large

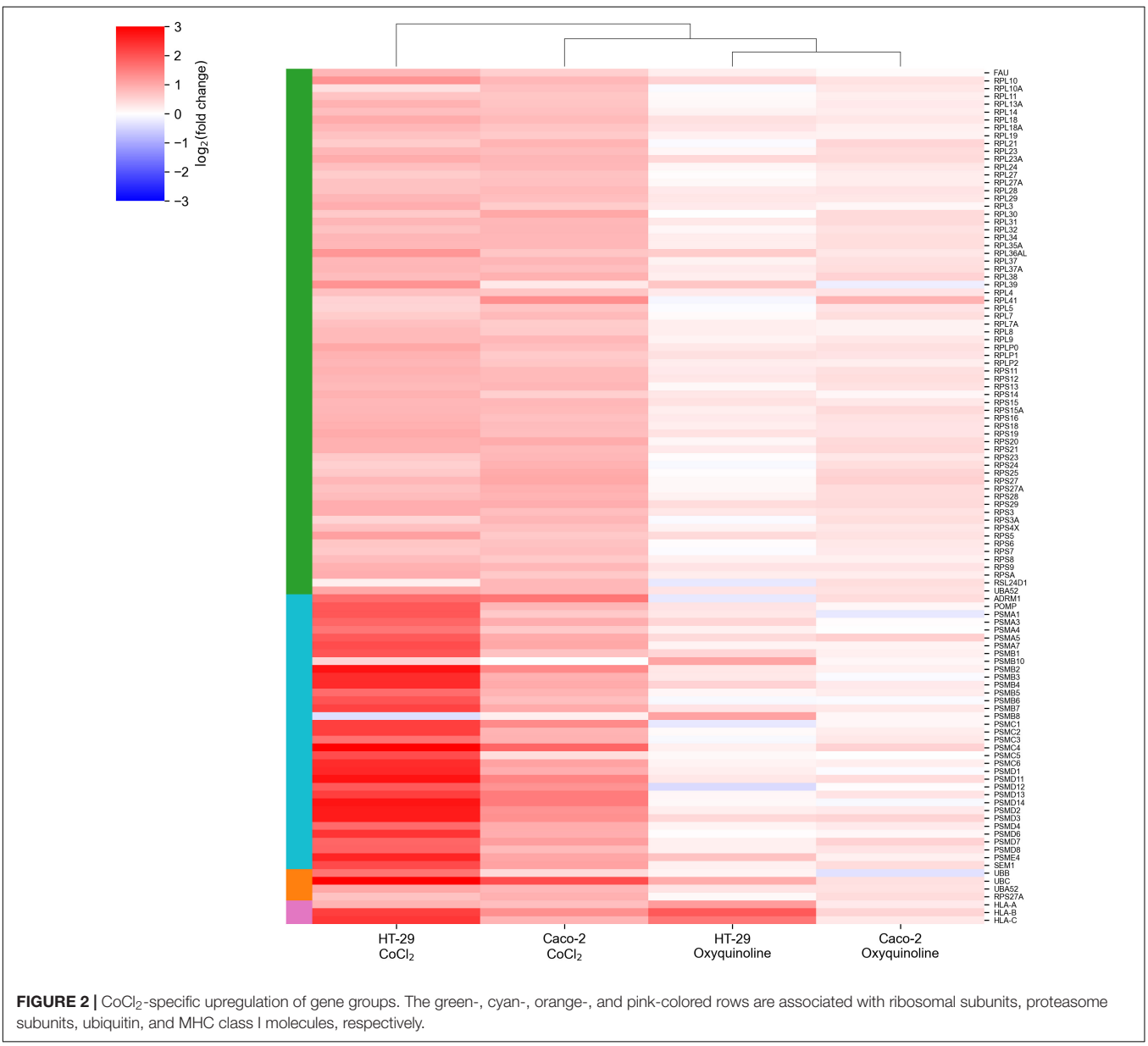


FIGURE 2 | $CoCl_2$-specific upregulation of gene groups. The green-, cyan-, orange-, and pink-colored rows are associated with ribosomal subunits, proteasome subunits, ubiquitin, and MHC class I molecules, respectively.

ribosomal subunits, subunits of proteasome 26S, ubiquitin, and HLA-A, HLA-B, and HLA-C genes that make up MHC class I (**Figure 2**). As can be seen, these effects were negligible for oxyquinoline treatment. Thus, the cell can trigger an immune response upon $CoCl_2$-induced oxidative stress *via* enhanced protein translation, proteasomal cleavage, and presentation of damaged peptides by MHC class I molecules. From the point of tumor hypoxia, this can be interpreted as activation of antitumor immunity.

A Set of miRNAs Regulated by Chemically Induced Hypoxia Is Highly Dependent on Cell Line and Chemical Agent

In addition to mRNA sequencing, we performed miRNA-seq on the same cells, which allowed us to reconstruct the whole miRNA and mRNA expression landscape under chemically simulated hypoxia. For the analysis, we selected miRNAs with a high expression level and a significant rate of differential expression in at least one experimental condition. Generally, numbers of differentially expressed miRNAs in two cell lines and two treatments followed the same pattern as in the case of mRNA sequencing, indicating a stronger molecular response to $CoCl_2$, namely, there were 22 and seven differentially expressed miRNAs for $CoCl_2$ and oxyquinoline treatments of Caco-2, while treatments of HT-29 using the same agents resulted in 16 and 7 miRNAs, respectively (**Figure 3** and **Supplementary Table 2**).

Only one miRNA was differentially expressed in all experimental conditions: hsa-miR-210-3p was consistently upregulated in both cell lines treated by both chemical agents. Nevertheless, multiple miRNAs were deregulated within the same cell line under different factors causing hypoxia: five miRNAs (hsa-miR-27b-5p, hsa-miR-148a-3p, hsa-miR-200a-5p, hsa-miR-1260a, and hsa-miR-1260b) were affected in Caco-2, and four miRNAs (hsa-let-7a-3p, hsa-miR-22-3p, hsa-miR-615-3p, and hsa-miR-4521) were identified in HT-29. Notably, these lists contained miRNA with an especially high level of expression: hsa-miR-148a-3p, which contributes to 6.9% of the whole Caco-2 miRNome, was 1.45-fold downregulated under both treatments of Caco-2. Treatment-specific miRNAs (i.e., miRNAs which are differentially expressed only under one treatment) were excluded from the downstream analysis.

Downregulation of miR-148a Contributes to Poor Survival in Colorectal Cancer Patients by Upregulation of Its Target Genes

In order to identify targets of 10 differentially expressed miRNAs, we implemented a three-step procedure. First, we made a sequence-based miRNA target prediction using miRDB software.

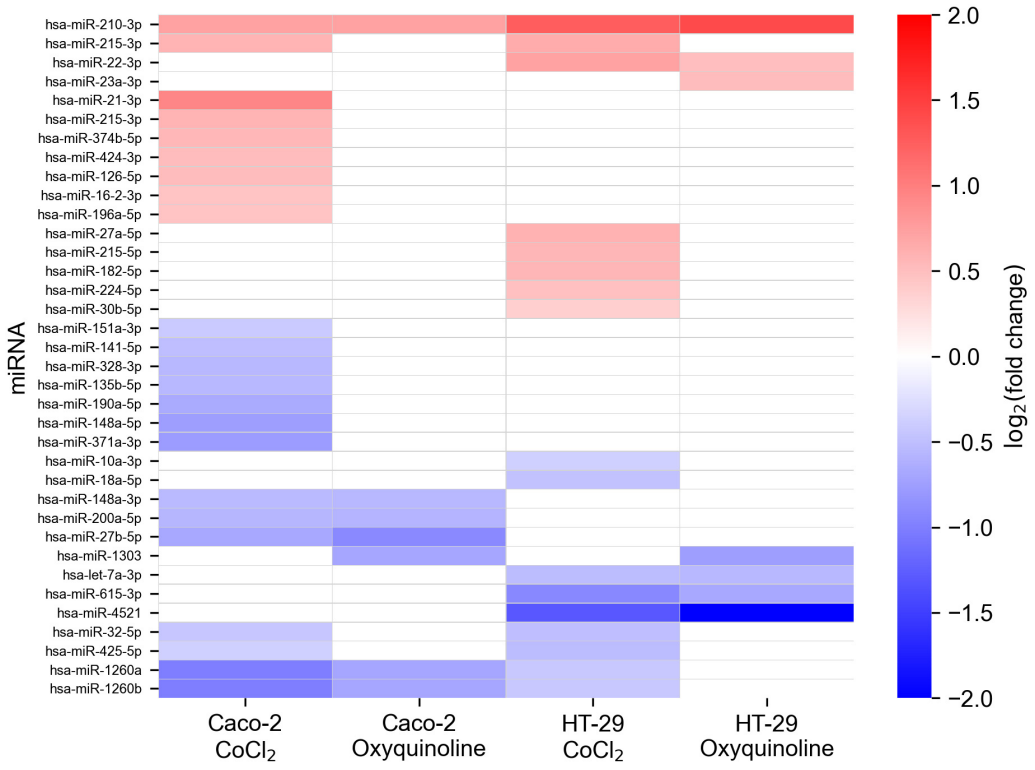


FIGURE 3 | Changes in miRNA expression levels of colorectal cell lines exposed to chemically simulated hypoxia. A miRNA is included in the plot in case if it is differentially expressed in at least one out of four experimental conditions. Empty cells correspond to miRNAs which were not differentially expressed or had not passed the thresholds on expression level (i.e., less expressed).

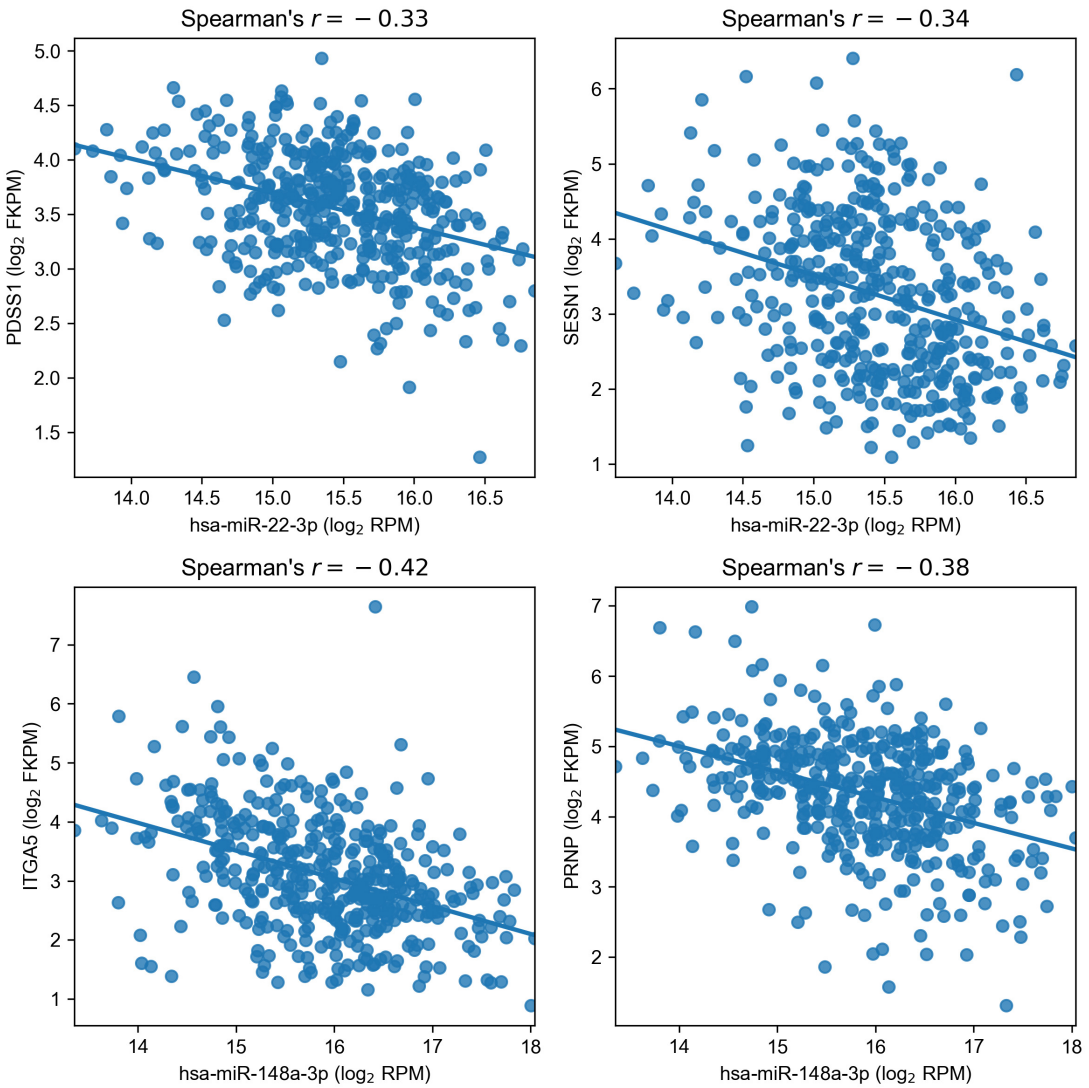


FIGURE 4 | Negative correlation of miR-22 and miR-148a with their target genes in The Cancer Genome Atlas Colon Adenocarcinoma samples.

Then, we selected miRNA-induced interactions which can lead to degradation of target mRNA in colon adenocarcinoma samples. For that, we selected negatively correlated miRNA–target mRNA pairs using the set of 426 primary tumors derived from the TCGA-COAD project. Finally, the obtained list was intersected with the list of genes significantly upregulated or downregulated in the opposite direction to the respective miRNA fold change (in both treatments of the respective cell line).

As a result, we obtained a list of four regulatory miRNA–mRNA interactions induced by two miRNAs: PDSS1 and SESN1 were targets of hsa-miR-22-3p (miR-22), and hsa-miR-148a-3p (miR-148a) regulated ITGA5 and PRNP (**Figure 4**). For additional confirmation on the existence of such interactions, we analyzed whether the target genes of miR-22 and miR-148a were deregulated only in the cell lines where the corresponding miRNA had altered the expression upon hypoxic exposure. The results fully supported our hypotheses: PDSS1 and SESN1 were not differentially expressed in Caco-2 (where miR-22 was also not differentially expressed), and PRNP was not differentially expressed in HT-29 (where miR-148a was not deregulated). Even stronger results were obtained for ITGA5: both treatments of Caco-2 induced the upregulation of three integrin alpha subunits (ITGA2, ITGA5, and ITGA6), and the expression increase for ITGA5 was much higher (18-folds for $CoCl_2$ and 15-fold for oxyquinoline) compared to that of other subunits (3.44 folds at maximum). At the same time, ITGA5 expression was not changed in the HT-29 cell line, and ITGA2 was downregulated (i.e., the differential expression of integrins cannot be explained as a consequence of a regulation by one common factor).

Surprisingly, both targets of miR-148a allowed us to predict the overall survival in colon adenocarcinoma patients with a statistical significance: logrank test p was equal to 0.0133 and 0.0119 for ITGA5 and PRNP, respectively (**Figure 5**). Moreover, overexpression of both genes led to poorer survival. In contrast,

the expression levels of miR-22 targets (PDSS1 and SESN1) were not associated with overall survival. Thus, downregulation of miR-148a can lead to colon cancer progression through upregulation of its target genes.

Hypoxia Could Induce the Downregulation of miR-148a Through TFAP2C Transcription Factor Activation

To assess whether the expression of miR-148a was decreased during hypoxia of patient tumors, we calculated Spearman correlation between the expression levels of miR-148a and two markers of hypoxia, HIF1A and SLC2A3 (Stuart Wood et al., 2007). This resulted in r=–0.22 ($p=6.80\times 10^{-6}$) for HIF1A and r=–0.3 ($p=3.48\times 10^{-10}$) for SLC2A3, supporting the hypothesis of hypoxia-induced downregulation of miR-148a.

Finally, we conducted bioinformatics analysis to predict possible transcription factors (TFs) which could directly regulate miR-148a during hypoxia. For that, we used the TransmiR database of TF–miRNA regulations supported by ChIP-seq experiments (Tong et al., 2019). Similar to what we have done with miRNA target prediction, we searched for TFs which are present in the database, deregulated under both treatments of Caco-2 in the same direction, and whose mRNA level significantly correlated with miR-148a in the TCGA-COAD samples (with the matching sign). As a result, we found transcription factor AP-2 gamma encoded by TFAP2C gene: its expression levels were increased by 10.2 and 6.6 times during $CoCl_2$ and oxyquinoline treatments, and the correlation in TCGA-COAD was equal to −0.19 ($p=2.02\times 10^{-3}$). Meanwhile, this TF was not expressed in HT-29 cell lines. Taking all these together, we propose the possible complex hypoxia-induced regulatory network (**Figure 6**): hypoxia promotes the expression of TFAP2C transcription factor, which results in the decreased expression of miR-148a. The latter induces the upregulation of two miRNA target genes (ITGA5 and PRNP), which finally contributes to tumor progression.

DISCUSSION

Two different colorectal cell lines and two chemical agents were used to assess the effect of hypoxia on cellular mRNA and miRNA expression levels. As expected, hypoxia was induced in all four experimental conditions; however, there were both similarities and dissimilarities in the landscapes of differentially expressed mRNAs and miRNAs. In particular, $CoCl_2$ induced a stronger deregulation on the transcriptomic level, promoting

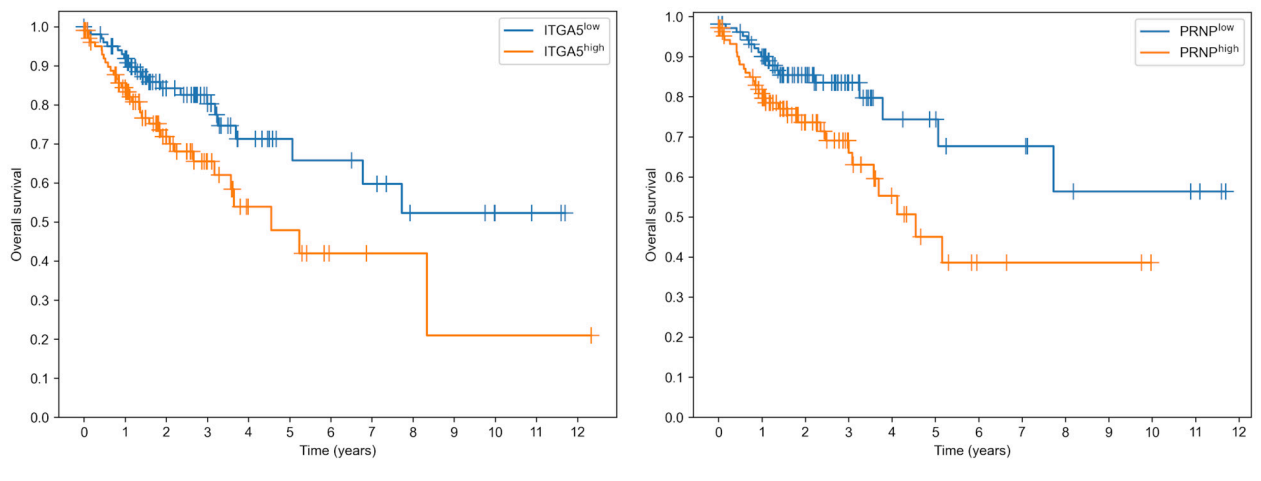


FIGURE 5 | Prognostic value of ITGA5 and PRNP genes in The Cancer Genome Atlas Colon Adenocarcinoma cohort.

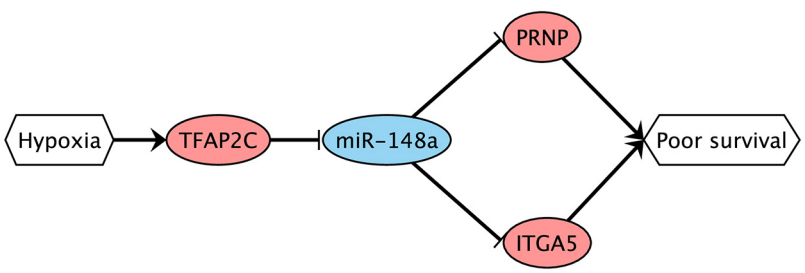


FIGURE 6 | Possible hypoxia-driven regulatory network induced by miR-148a gene silencing. Red nodes are associated with upregulation; blue nodes are associated with downregulation. The arrows and T-shaped lines signify activation and repression, respectively.

the expression of ribosomal subunits, proteasome subunits, ubiquitin, and MHC class I genes (**Figure 2**).

Consistently with previous reports, we observed the upregulation of miR-210 in both cell lines treated by both chemical agents (Kulshreshtha et al., 2007; Bavelloni et al., 2017; Bhandari et al., 2019). Aside from miR-210, we observed a differential expression of nine miRNAs within the same cell line under both exposures. One of those miRNAs, miR-148a, had a particularly high expression level contributing to approximately 7% of the whole miRNome of Caco-2. For both treatments, the expression levels of miR-148a were 1.45-folds decreased.

The multi-step bioinformatics analysis revealed two targets of miR-148a: integrin subunit α5 encoded by ITGA5 gene and the major prion protein PrP encoded by PRNP gene; both genes were upregulated in Caco-2 exposed to chemical hypoxia. Surprisingly, the elevated expression levels of both genes were associated with poor prognosis in colorectal cancer patients from the TCGA-COAD cohort. Thus, we hypothesize that direct interactions of miR-148a with ITGA5 and PRNP play an important role in tumor progression and metastasis. To the best of our knowledge, there are no studies on these interactions in colon cancer, though several groups already highlighted the role of miR-148a regulation of ITGA5 in other cancers, namely, such observations were made and experimentally verified (luciferase reporter assays) for breast (Cimino et al., 2013), gastric (Tseng et al., 2011), and non-small cell lung (Zhang J. et al., 2019) cancers. To the best of our knowledge, there are no reports containing a direct validation of interaction between miR-148a and PRNP. Aside from the mentioned target genes, it was already shown that miR-148a promotes apoptosis in colorectal cancer by silencing Bcl-2 (Zhang et al., 2011) and promotes proliferation of gastric cancer cells by targeting p27 (Guo et al., 2011).

In addition, we showed that the decreased expression of miR-148a is associated with tumor hypoxia in TCGA-COAD patients. With the use of TransmiR database of ChIP-seq experiments (Tong et al., 2019) and correlation analysis, we found a potential direct regulator of miR-148a during hypoxia: transcription factor AP-2 gamma encoded by TFAP2C gene. Further low-throughput experiments (such as overexpression/knockdown of miRNA/TF followed by reporter assay analyses) should be carried out to directly validate these hypotheses and study the proposed regulatory circuit in more detail. The proposed experiments should be also performed in more realistic *in vitro* microfluidic models (Samatov et al., 2015) and *in vivo*.

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

SN, AG, and AT contributed to the conceptualization and methodology. SN, AG, and MC contributed to the formal analysis. SN and MC contributed to the writing—original draft. SN, AG, MC, and AT contributed to the writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was performed within the framework of the Laboratory of Molecular Physiology at HSE University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.662468/full#supplementary-material

Supplementary Table 1 | The results of differential gene expression analysis.

Supplementary Table 2 | The results of differential miRNA expression analysis.

REFERENCES

Bavelloni, A., Ramazzotti, G., Poli, A., Piazzi, M., Focaccia, E., Blalock, W., et al. (2017). Mirna-210: a current overview. *Anticancer Res.* 37, 6511–6521. doi: 10.21873/anticanres.12107

BelAiba, R. S., Bonello, S., Zähringer, C., Schmidt, S., Hess, J., Kietzmann, T., et al. (2007). Hypoxia up-regulates hypoxia-inducible factor-1α transcription by involving phosphatidylinositol 3-kinase and nuclear factor κB in pulmonary artery smooth muscle cells. *Mol. Biol. Cell* 18, 4691–4697. doi: 10.1091/mbc.e07-04-0391

Bhandari, V., Hoey, C., Liu, L. Y., Lalonde, E., Ray, J., Livingstone, J., et al. (2019). Molecular landmarks of tumor hypoxia across cancer types. *Nat. Genet.* 51, 308–318. doi: 10.1038/s41588-018-0318-2

Brahimi-Horn, M. C., Chiche, J., and Pouysségur, J. (2007). Hypoxia and cancer. *J. Mol. Med.* 85, 1301–1307. doi: 10.1007/s00109-007-0281-3

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

Catron, T., Mendiola, M. A., Smith, S. M., Born, J., and Walker, M. K. (2001). Hypoxia regulates avian cardiac arnt and HIF-1α mRNA expression. *Biochem. Biophys. Res. Commun.* 282, 602–607. doi: 10.1006/bbrc.2001.4613

Chen, Y., and Wang, X. (2020). MiRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Res.* 48, D127–D131. doi: 10.1093/nar/gkz757

Cimino, D., De Pittà, C., Orso, F., Zampini, M., Casara, S., Penna, E., et al. (2013). miR148b is a major coordinator of breast cancer progression in a relapse-associated microRNA signature by targeting ITGA5, ROCK1, PIK3CA, NRAS, and CSF1. *FASEB J.* 27, 1223–1235. doi: 10.1096/fj.12-214692

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773. doi: 10.1093/nar/gky955

Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.* 40, 37–52. doi: 10.1093/nar/gkr688

Guo, S.-L., Peng, Z., Yang, X., Fan, K.-J., Ye, H., Li, Z.-H., et al. (2011). miR-148a promoted cell proliferation by targeting p27 in gastric cancer cells. *Int. J. Biol. Sci.* 7, 567–574. doi: 10.7150/ijbs.7.567

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). MiRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162. doi: 10.1093/nar/gky1141

Kulshreshtha, R., Ferracin, M., Wojcik, S. E., Garzon, R., Alder, H., Agosto-Perez, F. J., et al. (2007). A MicroRNA signature of hypoxia. *Mol. Cell. Biol.* 27, 1859–1867. doi: 10.1128/mcb.01395-06

Lange, T., Samatov, T. R., Tonevitsky, A. G., and Schumacher, U. (2014). Importance of altered glycoprotein-bound N- and O-glycans for epithelial-to-mesenchymal transition and adhesion of cancer cells. *Carbohydr. Res.* 389, 39–45. doi: 10.1016/j.carres.2014.01.010

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10:R25. doi: 10.1186/gb-2009-10-3-r25

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

Maltseva, D. V., Galatenko, V. V., Samatov, T. R., Zhikrivetskaya, S. O., Khaustova, N. A., Nechaev, I. N., et al. (2014). MiRNome of inflammatory breast cancer. *BMC Res. Notes* 7:871. doi: 10.1186/1756-0500-7-871

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17:10. doi: 10.14806/ej.17.1.200

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.* 24, 69–71.

Muñoz-Sánchez, J., and Chánez-Cárdenas, M. E. (2019). The use of cobalt chloride as a chemical hypoxia model. *J. Appl. Toxicol.* 39, 556–570. doi: 10.1002/jat.3749

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., et al. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330–337. doi: 10.1038/nature11252

Nagaraju, G. P., Bramhachari, P. V., Raghu, G., and El-Rayes, B. F. (2015). Hypoxia inducible factor-1α: its role in colorectal carcinogenesis and metastasis. *Cancer Lett.* 366, 11–18. doi: 10.1016/j.canlet.2015.06.005

Nersisyan, S., Engibaryan, N., Gorbonos, A., Kirdey, K., Makhonin, A., and Tonevitsky, A. (2020a). Potential role of cellular miRNAs in coronavirus-host interplay. *PeerJ* 8:e9994. doi: 10.7717/peerj.9994

Nersisyan, S., Shkurnikov, M., Poloznikov, A., Turchinovich, A., Burwinkel, B., Anisimov, N., et al. (2020b). A post-processing algorithm for miRNA microarray data. *Int. J. Mol. Sci.* 21:1228. doi: 10.3390/ijms21041228

Nersisyan, S., Shkurnikov, M., Turchinovich, A., Knyazev, E., and Tonevitsky, A. (2020c). Integrative analysis of miRNA and mRNA sequencing data reveals potential regulatory mechanisms of ACE2 and TMPRSS2. *PLoS One* 15:e0235987. doi: 10.1371/journal.pone.0235987

Osipyants, A. I., Smirnova, N. A., Khristichenko, A. Y., Hushpulian, D. M., Nikulin, S. V., Chubar, T. A., et al. (2017). Enzyme–substrate reporters for evaluation of substrate specificity of HIF prolyl hydroxylase isoforms. *Biochemistry* 82, 1207–1214. doi: 10.1134/S0006297917100145

Panigrahi, G. K., Ramteke, A., Birks, D., Ali, H. E. A., Venkataraman, S., Agarwal, C., et al. (2018). Exosomal microRNA profiling to identify hypoxia-related biomarkers in prostate cancer. *Oncotarget* 9, 13894–13910. doi: 10.18632/oncotarget.24532

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Samatov, T. R., Shkurnikov, M. U., Tonevitskaya, S. A., and Tonevitsky, A. G. (2015). Modelling the metastatic cascade by in vitro microfluidic platforms. *Prog. Histochem. Cytochem.* 49, 21–29. doi: 10.1016/j.proghi.2015.01.001

Savyuk, M., Krivonosov, M., Mishchenko, T., Gazaryan, I., Ivanchenko, M., Khristichenko, A., et al. (2020). Neuroprotective effect of HIF prolyl hydroxylase inhibition in an in vitro hypoxia model. *Antioxidants* 9:662. doi: 10.3390/antiox9080662

Shen, G., Li, X., Jia, Y. F., Piazza, G. A., and Xi, Y. (2013). Hypoxia-regulated microRNAs in human cancer. *Acta Pharmacol. Sin.* 34, 336–341. doi: 10.1038/aps.2012.195

Shkurnikov, M., Nikulin, S., Nersisyan, S., Poloznikov, A., Zaidi, S., Baranova, A., et al. (2019). LAMA4-regulating miR-4274 and its host gene SORCS2 play a role in IGFBP6-dependent effects on phenotype of basal-like breast cancer. *Front. Mol. Biosci.* 6:122. doi: 10.3389/fmolb.2019.00122

Skalsky, R. L., and Cullen, B. R. (2010). Viruses, microRNAs, and host interactions. *Annu. Rev. Microbiol.* 64, 123–141. doi: 10.1146/annurev.micro.112408.134243

Stuart Wood, I., Wang, B., Lorente-Cebrián, S., and Trayhurn, P. (2007). Hypoxia increases expression of selective facilitative glucose transporters (GLUT) and 2-deoxy-d-glucose uptake in human adipocytes. *Biochem. Biophys. Res. Commun.* 361, 468–473. doi: 10.1016/j.bbrc.2007.07.032

Tong, Z., Cui, Q., Wang, J., and Zhou, Y. (2019). TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Res.* 47, D253–D258. doi: 10.1093/nar/gky1023

Toung, J. M., Morley, M., Li, M., and Cheung, V. G. (2011). RNA-sequence analysis of human B-cells. *Genome Res.* 21, 991–998. doi: 10.1101/gr.116335.110

Tseng, C.-W., Lin, C.-C., Chen, C.-N., Huang, H.-C., and Juan, H.-F. (2011). Integrative network analysis reveals active microRNAs and their functions in gastric cancer. *BMC Syst. Biol.* 5:99. doi: 10.1186/1752-0509-5-99

Turchinovich, A., Tonevitsky, A. G., Cho, W. C., and Burwinkel, B. (2015). Check and mate to exosomal extracellular miRNA: new lesson from a new approach. *Front. Mol. Biosci.* 2:11. doi: 10.3389/fmolb.2015.00011

Vaupel, P., and Mayer, A. (2007). Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev.* 26, 225–239. doi: 10.1007/s10555-007-9055-1

Visone, R., and Croce, C. M. (2009). MiRNAs and cancer. *Am. J. Pathol.* 174, 1131–1138. doi: 10.2353/ajpath.2009.080794

Wu, D., and Yotnda, P. (2011). Induction and testing of hypoxia in cell culture. *J. Vis. Exp.* 54:2899. doi: 10.3791/2899

Zhang, H., Li, Y., Huang, Q., Ren, X., Hu, H., Sheng, H., et al. (2011). MiR-148a promotes apoptosis by targeting Bcl-2 in colorectal cancer. *Cell Death Differ.* 18, 1702–1710. doi: 10.1038/cdd.2011.28

Zhang, J., Zhang, Y., Shen, W., Fu, R., Ding, Z., Zhen, Y., et al. (2019). Cytological effects of honokiol treatment and its potential mechanism of action in non-small cell lung cancer. *Biomed. Pharmacother.* 117:109058. doi: 10.1016/j.biopha.2019.109058

Zhang, X., Zhang, J., Zheng, K., Zhang, H., Pei, X., Yin, Z., et al. (2019). Long noncoding RNAs sustain high expression levels of exogenous octamer-binding protein 4 by sponging regulatory microRNAs during cellular reprogramming. *J. Biol. Chem.* 294, 17863–17874. doi: 10.1074/jbc.RA119.010284

Zhu, A., Ibrahim, J. G., and Love, M. I. (2019). Heavy-Tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* 35, 2084–2092.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 02 June 2021
doi: 10.3389/fgene.2021.662843

Effect of the Expression of *ELOVL5* and *IGFBP6* Genes on the Metastatic Potential of Breast Cancer Cells

Sergey Nikulin[1], Galina Zakharova[2], Andrey Poloznikov[1,3], Maria Raigorodskaya[1,2], Daniel Wicklein[4], Udo Schumacher[4], Stepan Nersisyan[1], Jonas Bergquist[5], Georgy Bakalkin[6], Lidiia Astakhova[2,7] and Alexander Tonevitsky[1,8]**

[1] Faculty of Biology and Biotechnologies, National Research University Higher School of Economics, Moscow, Russia, [2] Scientific Research Centre Bioclinicum, Moscow, Russia, [3] School of Biomedicine, Far Eastern Federal University, Vladivostok, Russia, [4] Institute of Anatomy and Experimental Morphology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [5] Department of Chemistry – BMC, Uppsala University, Uppsala, Sweden, [6] Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden, [7] School of Life Sciences, Immanuel Kant Baltic Federal University, Kaliningrad, Russia, [8] Laboratory of Microfluidic Technologies for Biomedicine, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia

OPEN ACCESS

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

Reviewed by:
Pavel Bouchal, Masaryk University, Czechia
Anton A. Buzdin, I.M. Sechenov First Moscow State Medical University, Russia
Konstantin Kandror, Boston University, United States

***Correspondence:**
Sergey Nikulin
nikulin.c.b@gmail.com
Alexander Tonevitsky
atonevitsky@hse.ru

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

Received: *17 February 2021*
Accepted: *20 April 2021*
Published: *02 June 2021*

Citation:
Nikulin S, Zakharova G, Poloznikov A, Raigorodskaya M, Wicklein D, Schumacher U, Nersisyan S, Bergquist J, Bakalkin G, Astakhova L and Tonevitsky A (2021) Effect of the Expression of ELOVL5 and IGFBP6 Genes on the Metastatic Potential of Breast Cancer Cells. Front. Genet. 12:662843. doi: 10.3389/fgene.2021.662843

Breast cancer (BC) is the leading cause of death from malignant neoplasms among women worldwide, and metastatic BC presents the biggest problems for treatment. Previously, it was shown that lower expression of *ELOVL5* and *IGFBP6* genes is associated with a higher risk of the formation of distant metastases in BC. In this work, we studied the change in phenotypical traits, as well as in the transcriptomic and proteomic profiles of BC cells as a result of the stable knockdown of *ELOVL5* and *IGFBP6* genes. The knockdown of *ELOVL5* and *IGFBP6* genes was found to lead to a strong increase in the expression of the matrix metalloproteinase (MMP) *MMP1*. These results were in good agreement with the correlation analysis of gene expression in tumor samples from patients and were additionally confirmed by zymography. The knockdown of *ELOVL5* and *IGFBP6* genes was also discovered to change the expression of a group of genes involved in the formation of intercellular contacts. In particular, the expression of the *CDH11* gene was markedly reduced, which also complies with the correlation analysis. The spheroid formation assay showed that intercellular adhesion decreased as a result of the knockdown of the *ELOVL5* and *IGFBP6* genes. Thus, the obtained data indicate that malignant breast tumors with reduced expression of the *ELOVL5* and *IGFBP6* genes can metastasize with a higher probability due to a more efficient invasion of tumor cells.

Keywords: breast cancer, *ELOVL5*, *IGFBP6*, matrix metalloproteinases, cell–cell contacts, MDA-MB-231

INTRODUCTION

Today, breast cancer (BC) is the most common malignant neoplasm in women worldwide (Bray et al., 2018). More than 2 million new cases of this disease are registered in the world every year. Despite the decrease in mortality from BC that has been outlined in recent years, it still remains the leading cause of death among women from cancer (Bray et al., 2018). More than 600,000 women die from BC in the world annually (Bray et al., 2018).

One of the main problems in the treatment of BC is relapse after primary treatment. According to recent statistics, relapse develops in about 40% of patients (Gerber et al., 2010; Lafourcade et al., 2018). Moreover, about one-third of the cases are local relapses, and two-thirds of the cases are distant metastases (Gerber et al., 2010; Lafourcade et al., 2018). Generally, treatment of patients with distant metastases is symptomatic and is not aimed at the complete cure of the disease (Gerber et al., 2010; Redig and McAllister, 2013).

In order to predict BC relapse earlier, methods for high-throughput analysis of gene expression revealed transcriptomic prognostic gene signatures (Hyams et al., 2017; Kwa et al., 2017). Today, the most popular commercially available transcriptomic test systems for BC used in clinical practice are Oncotype DX, Prosigna, and MammaPrint (Hyams et al., 2017; Kwa et al., 2017). On the one hand, utilization of transcriptomic test systems in clinical practice makes it possible to identify a group of patients with low risk of relapse and to avoid prescription of excessive treatment for them, which significantly improves the quality of life and reduces healthcare costs. On the other hand, their use allows early identification of patients with a high risk of distant metastases and justifies utilization of more intensive treatment protocols that reduce the risk of relapse. However, it should be noted that the need to create new, more advanced test systems is evidenced by the fact that the results of various tests available on the market do not agree well with each other when applied to the same group of patients (Bartlett et al., 2016).

Previously, our research group created its own classifier to identify patients with high risk of distant BC metastases, based on measuring the expression of only two genes (Galatenko et al., 2015). A fundamentally different approach to the selection of genes included in the consideration was used (Samatov et al., 2017; Galatenko et al., 2018). Traditionally, only genes with high individual information content were used in such gene signatures (those genes whose expression levels differ significantly between groups with favorable and unfavorable prognosis). At the same time, genes whose average expression did not differ significantly between groups with different prognosis were also used to construct this classifier. It was shown that taking the expression levels of such genes together with other genes into account can significantly improve the quality of classification. According to the obtained results, the most informative pair was the *ELOVL5*–*IGFBP6* gene pair (high expression of *ELOVL5* and *IGFBP6* corresponded to favorable prognosis). Previously, these genes had not been associated with the risk of BC metastases and, individually, do not have strong predictive power (i.e., it is not possible to assess the risk of relapse accurately based on the expression of just one of these genes). However, on the basis of the analysis of large microarray dataset of BC samples (kmplot.com), it can be concluded (**Supplementary Figure 1**) that high expression of each *ELOVL5* [hazard ratio (HR) = 0.54, $p < 0.001$] and *IGFBP6* (HR = 0.76, $p < 0.001$) messenger RNAs (mRNAs) is associated with better distant metastasis-free survival (DMFS) (Győrffy et al., 2010).

Moreover, previously, *ELOVL5* and *IGFBP6* genes seemed to be unrelated to each other, and the reason for the observed synergism of the levels of expression of these two genes in the prediction of BC relapse was unclear. *ELOVL5* is one of the elongases of polyunsaturated fatty acids (PUFAs) located in the membrane of the endoplasmic reticulum (Leonard et al., 2000; Wang et al., 2008; Moon et al., 2009), and *IGFBP6* is a secreted protein that binds to insulin-like growth factors (IGFs) preventing their action on cells (Bach et al., 2013; Bach, 2015).

The spread of tumor cells throughout the body occurs during a multistep invasive-metastatic cascade, which consists of several stages (Samatov et al., 2015; Lambert et al., 2017). The aim of this work was to study the effect of the expression of the *ELOVL5* and *IGFBP6* genes on the features of BC cells associated with metastasis including the changes in the transcriptome and proteomic profiles as well as phenotypic traits.

MATERIALS AND METHODS

Analysis of Transcriptomic Databases

The Cancer Cell Line Encyclopedia (CCLE) database was analyzed to select a BC cell line suitable for knockdown of the studied genes (Barretina et al., 2012).

The following datasets (**Supplementary Table 1**) from the Gene Expression Omnibus (GEO) were used for correlation analysis: GSE102484 (Cheng et al., 2017), GSE22220 (Camps et al., 2008), GSE3494 (Miller et al., 2005), GSE58644 (Miller et al., 2005), and GSE6532 (Loi et al., 2008). We also used data obtained by the METABRIC consortium (Cerami et al., 2012; Curtis et al., 2012) and The Cancer Genome Atlas (TCGA) program (Weinstein et al., 2013).

TAC 4.0 software (Thermo Fisher Scientific) was applied to preprocess raw data from Affymetrix microarrays. To carry out correlation analysis and statistical data processing, we employed the R 3.5 programming language with the RStudio 1.1 integrated development environment. The values of the Pearson correlation coefficient R and the p-values (the significance of the difference of R from zero) were calculated using the "cor.test" function. Correction for multiple comparisons was performed with the Benjamini–Hochberg method. The correlation coefficients with $p < 0.05$ were considered significant.

Cell Culture

Human MDA-MB-231 BC cells were cultured in a complete cell culture medium consisting of Dulbecco's modified Eagle's medium (DMEM) high glucose (Gibco) supplemented with 10% vol. fetal bovine serum (Gibco), 2 mM L-glutamine (PanEco), and 1% vol. penicillin-streptomycin solution (Gibco). The cells were incubated in a cell culture incubator (37°C, 5% CO_2) MCO-18AC (Sanyo). Subcultivation was performed every 2–3 days using trypsin–ethylenediaminetetraacetic acid (EDTA) solution (PanEco). Photomicrographs of the cells were obtained using an inverted Primo Vert microscope (Carl Zeiss). Cells were counted after trypan blue (Gibco) staining using Countess automated cell counter (Invitrogen) according to the manufacturer's protocol.

To obtain three-dimensional spheroids, 96-well plates with low adhesion and a U-shaped bottom (Corning) were used. Two hundred microliters of cell suspension was added to each well of the plate. Then, the plate was incubated for 96 h in

a cell culture incubator (37°C, 5% CO_2) MCO-18AC (Sanyo). Photos of spheroids were obtained using an inverted microscope Axio Observer Z1 (Carl Zeiss). The experiment was performed independently three times. Each time a different number of cells per well was used (3,000, 5,000, and 6,000).

Stable Knockdown of ELOVL5 and IGFBP6 Genes

Two cultures of MDA-MB-231 cells with reduced expression of messenger RNA (mRNA) of the *IGFBP6* gene (**Supplementary Figure 2**) were generated earlier (Nikulin et al., 2018). In this work, only MDA-MB-231 (IGFBP6_2) cells with the most pronounced decrease in *IGFBP6* gene expression were used as the cells with the *IGFBP6* gene knockdown. Stable knockdown of *ELOVL5* gene was performed similarly using RNA interference (Schwankhaus et al., 2014; Maltseva et al., 2020). DNA oligonucleotides selected for the target sequences in the *ELOVL5* gene were ligated into the pLVX short hairpin RNA 1 (shRNA1) lentiviral vector (Clontech Laboratories) according to the manufacturer's protocol. We used two different target sequences with their own set of DNA oligonucleotides (**Supplementary Table 2**). To obtain the control MDA-MB-231 (LUC) cells, we used the same lentiviral vector pLVX shRNA1 containing shRNA to the *Photinus pyralis* firefly luciferase gene. Viral particles were obtained in the form of cell-free supernatants using transient transfection of HEK-293T cell line according to the previously described method (Weber et al., 2010, 2012). Supernatants were collected 24 h after transfection, filtered using 0.45-μm syringe filters, and stored at −80°C. Then, 5×10^4 MDA-MB-231 cells were cultured in the wells of a 24-well culture plate in 0.5 ml of cell culture medium. After 24 h, 10 μl of the supernatant containing viral particles was added to the wells, and the plate was placed in a cell culture incubator for 24 h. Then, the cell culture medium was changed, and the cells were incubated for another 24 h. After that, the selection with 1 μg/ml puromycin (Gibco) was carried out for 2 weeks.

Real-Time PCR

Real-time PCR was used to assess changes in the expression of individual genes as a result of the knockdown of the studied genes *ELOVL5* and *IGFBP6* (Nikulin et al., 2018). Cells of the studied lines were plated into six-well plates at 5×10^5 cells per well in 2.5 ml of complete culture medium and incubated in a CO_2 incubator (37°C, 5% CO_2) for 48 h. Next, the cell culture medium was removed from the wells, and the cells were washed three times with cold (4°C) Dulbecco's phosphate-buffered saline (DPBS) solution (PanEco). The cells were then lysed using QIAzol Lysis Reagent (QIAGEN). Seven hundred microliters (700 μl) of QIAzol Lysis Reagent solution (QIAGEN) was added to each well and incubated at room temperature for 5 min. Then, the contents of the wells were thoroughly mixed by pipetting and transferred into microtubes, which were stored at −80°C before RNA isolation.

RNA isolation was performed using miRNeasy Micro Kit (QIAGEN) according to the manufacturer's protocol. RNA concentration was measured with a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific). The quality of the isolated RNA (no degradation) was assessed using Experion bioanalyzer (Bio-Rad). Only the samples with RNA integrity number (RIN) ≥ 7 were used.

Reverse transcription of RNA was performed using the MMLV RT kit (Evrogen) according to the manufacturer's protocol. The obtained complementary DNA (cDNA) samples were stored at −20°C. qPCRmix-HS SYBR (Evrogen) was used for RT-PCR performed with DTprime detecting amplifier (DNA Technology).

The oligonucleotide primers used for RT-PCR were designed based on the mRNA sequences of the studied genes from the University of California Santa Cruz (UCSC) Genome Browser database (Kent et al., 2002). Primer selection was performed using Primer-BLAST software (Ye et al., 2012). The possibility of the formation of secondary structures (hairpins), homo- and heterodimers by the primers, was assessed using OligoAnalyzer 3.1 software (Owczarzy et al., 2008). *EEF1A1* and *HUWE1* were selected as reference genes (Maltseva et al., 2013). The sequences of the primers used, the lengths of the resulting amplicons, and the values of the amplification efficiencies are presented in **Supplementary Table 3**. The evaluation of the differences in the expression of the selected genes in the cells with knockdown of the *ELOVL5* and *IGFBP6* genes in comparison with the control MDA-MB-231 cells was carried out using the software REST 2009 v.2.0.13 (Pfaffl et al., 2002; Vandesompele et al., 2002). For each group, three independently obtained samples of RNA were used to assess expression levels of the selected genes.

Western Blotting

Western blotting was used to evaluate the efficiency of the knockdown of the studied genes at protein level. To assess the knockdown of the *ELOVL5* protein, cells were lysed in radioimmunoprecipitation assay (RIPA) buffer; then, the protein concentration was measured using Pierce BCA Protein Assay kit (Thermo Fisher Scientific) according to the manufacturer's instructions. Electrophoresis was performed in polyacrylamide gel (PAAG) (12%). Transfer to the polyvinylidene fluoride (PVDF) membrane was performed using Trans-Blot Turbo transfer system (Bio-Rad) according to the manufacturer's instructions. The membrane was then blocked in 3% bovine serum albumin (BSA) solution in TBST (Tris-buffered saline, 0.1% Tween 20) for 1 h and incubated with rabbit primary antibodies to ELOVL5 protein (Abcam, ab205535) overnight at 4°C. Then, the membrane was washed in TBST solution and incubated with secondary goat antibodies to rabbit immunoglobulins conjugated with peroxidase. Clarity Western ECL Substrate (Bio-Rad) was used as a substrate for peroxidase. The resulting membrane was photographed using Gel Doc XR+ gel documenting station (Bio-Rad).

Since IGFBP6 is a secreted protein, serum-free medium samples after incubation with the cells for 24 h were analyzed to assess IGFBP6 protein knockdown. A similar Western blotting protocol was used (nonfat dry milk was used to block the membrane instead of BSA) with primary antibodies to the IGFBP6 protein (Abcam, ab109765). Samples were normalized to the number of cells.

Western blotting analysis for each protein was performed independently two times.

Cell Proliferation Assay

The protocol of the used MTT cell proliferation assay has been published earlier (Nikulin et al., 2018). The proliferation rate was estimated as:

$$R_{72/24} = \frac{A_{72} - O}{A_{24} - O},$$

where $R_{72/24}$ —the ratio of the number of cells in a well after 72 h to the number of cells after 24 h from seeding; A_{24}, A_{72}— the absorption value in the wells with the studied cells after 24 and 72 h; O - mean background absorption. The experiment was carried out in six replicas. Student's t-test was used to determine the statistical significance of the observed differences.

Apoptosis Assay

To study the activation of apoptosis, Dead Cell Apoptosis Kit with annexin V Alexa Fluor^{TM} 488 and propidium iodide (PI) (Thermo Fisher Scientific) were employed according to the manufacturer's instructions. The cell suspension was centrifuged at 500 × g for 3 min, and the supernatant was collected. Then, the cells were resuspended in 100 μl of the buffer solution for annexin binding, and 5 μl of annexin V conjugate with Alexa Fluor 488 (AV) and 1 μl of propidium iodide (PI) solution with concentration of 100 μg/ml were added. After that, the cells were incubated at room temperature in a dark place for 15 min. Then, 400 μl of the buffer solution for annexin binding was added to the suspension, microtubes were transferred onto ice, and the samples were analyzed with a CytoFLEX flow cytometer (Beckman Coulter). The experiment was performed independently three times.

Analysis of raw data was carried out using FlowJo 10.6.1 software. As a result, the proportions of the entire cell population were obtained, corresponding to living cells (AV−PI−), cells at an early stage of apoptosis (AV+PI−), dead cells, including those at the late stages of apoptosis (AV+PI+), and nuclear fragments without cell membranes that can result from necrosis (AV−PI+) (Sawai and Domae, 2011). Further statistical data processing was carried out using R 3.5 programming language with RStudio 1.1 integrated development environment. Analysis of variance (ANOVA) was used to determine the statistical significance of the observed differences, followed by determination of p-values in pairwise comparisons using Tukey's test. The differences were considered significant if the $p < 0.05$.

Cell Migration Assay

Migration activity of the cells was measured by scratch assay. One hundred microliters of culture medium containing 3×10^4 cells were added to each well of a 96-well plate. After that, the plate was incubated in a CO_2 incubator (37°C, 5% CO_2) overnight. Then, mitomycin C (Kyowa) was added to each well to the final concentration of 10 μg/ml for 2 h to stop proliferation. After that, scratches were made at the center of the wells using a 200-μl pipette tip, and cell culture medium was changed. Then, the plates were placed into a cell culture incubator. Each well was microphotographed at different time points (0, 4, 8, and 10 h) using a SpectraMax i3 plate reader (Molecular Devices). The experiment was carried out in 20 replicas.

ImageJ software was used to calculate the area of the scratches. Then, the dependence of the scratch area on time was plotted for each well, and the migration rate was estimated as the slope coefficient of the resulting straight line. The wells where coefficient of determination R^2 of the fitted straight line was < 0.95 were removed from further analysis. To determine the statistical significance of the observed differences, Mann–Whitney U test was applied.

Transcriptomic Analysis

Transcriptomic analysis of the generated cell cultures was performed using Human Transcriptome Array 2.0 microarrays (Affymetrix) according to the manufacturer's procedure. For each group, three independently obtained samples of RNA were used.

Raw data were processed using TAC 4.0 software (Thermo Fisher Scientific) using the RMA algorithm. To assess the statistical significance of differences in gene expression, ANOVA FDR p-values with threshold level of 0.05 were used. Further data processing was conducted using R 3.5 programming language with the RStudio 1.1 integrated development environment.

Statistical significance of the intersection between regulated genes (the probability that the intersection is a random event) after the knockdown of *ELOVL5* and *IGFBP6* genes was determined by permutation test (Nikitin et al., 2019; Sorokin et al., 2020). To determine the distribution of the number of the genes that significantly change their expression in the same direction after the knockdown of *ELOVL5* and *IGFBP6* genes in case of completely independent changes, gene names were randomly permuted 1,000,000 times. The fold changes and p-values were conserved, and the size of the overlap between genes significantly regulated in the same direction for each generated random gene set was measured.

Analysis of the enriched biological processes among the genes with increased and decreased expression was carried out using gene ontology (GO) database (Ashburner et al., 2000; Gene and Consortium, 2019) and "topGO" package for R programming language. The results were obtained using "weight01" algorithm; p-values were calculated using Fisher's exact test.

Pathway activation levels (PALs) were calculated with Oncobox Library (Sorokin et al., 2021) with the default set of pathway databases. Comparison of PALs between *ELOVL5*/*IGBFP6* knockdown cells with control ones was done with Student's t-test; Benjamini–Hochberg procedure was used to adjust p-values.

Proteomic Analysis

For proteomic analysis, cells were lysed with 3% sodium deoxycholate (SDC) solution in bicarbonate buffer (50 mM ammonium bicarbonate in water). The lysates were incubated for 15 min at 80°C, followed by sonication. Then, the disulfide bonds in the proteins were reduced with dithiothreitol (DTT) and alkylated with iodoacetamide (IAA). The resulting protein mixture was digested with Trypsin Gold (Promega) at 37°C overnight. Then, SDC was removed from the mixture by

precipitation with trifluoroacetic acid. The resulting mixture of peptides was purified using ZipTips (Merck Millipore) according to the manufacturer's protocol. Then, the samples were dried and dissolved in 0.1% vol. formic acid solution. The resulting peptides were analyzed using a nano-high performance liquid chromatography tandem mass spectrometry (nano-HPLC-MS/MS) system coupled with a Q Exactive Orbitrap mass spectrometer (Thermo Fisher Scientific). Separation was carried out with a reversed-phase C_{18} column in gradient elution mode; the duration of the gradient was 150 min. Fragment spectra were obtained using collision-induced dissociation. For each group, three independently obtained samples of proteins were used.

Raw data were analyzed using MaxQuant 1.6 software (Tyanova et al., 2016a). The iBAQ algorithm was used to quantify the protein content (Schwanhäusser et al., 2011). Further data processing was carried out using Perseus 1.6 software (Tyanova et al., 2016b) and R 3.5 programming language with the RStudio 1.1 integrated development environment. To determine the statistical significance of the observed differences, Student's *t*-test was used.

Enrichment analysis of biological processes and pathway analysis were performed as described above for transcriptomic analysis.

Zymography

The utilized method for assessment of the activity of matrix metalloproteinases (MMPs) was described earlier (Toth and Fridman, 2001). Serum-free cell culture medium was sampled after 24 h of incubation with the cells. Then, electrophoresis was performed in polyacrylamide gel containing 0.1% of gelatin. The resulting gel was incubated at 37°C overnight. Then, the gel was stained with Coomassie blue G-250 colloidal solution (Thermo Fisher Scientific). Clear zones in the stained gel correspond to the positions of active MMPs. The gel was photographed with Gel Doc XR+ gel documenting station (Bio-Rad). Zymography was performed independently two times.

RESULTS

Stable Knockdown of ELOVL5 and IGFBP6 Genes

To select suitable cell lines for the knockdown, a two-dimensional plot of the expression of *ELOVL5* and *IGFBP6* genes in BC cell lines according to publicly available database CCLE was constructed. It can be seen from the plot (**Figure 1**) that only a few cell lines have a sufficiently high expression of both studied genes (circled in green), and they are suitable ones for knockdown. Expression of major molecular markers in this group of cell lines is presented in **Supplementary Figure 3** (Dai et al., 2017). All these cells are estrogen and progesterone receptor negative, and only one of them is HER2 positive. Among these candidates for the knockdown, there was only one cell line that is often used as a model of triple negative BC. This cell line is MDA-MB-231, and it was chosen as the original cell model in this work.

Two cultures of MDA-MB-231 cells with a stable knockdown of the *ELOVL5* gene were generated in this work (**Table 1** and **Supplementary Figure 2**). For further analysis, only MDA-MB-231 (ELOVL5_2) cells with the most pronounced decrease in *ELOVL5* gene expression were used as the cell line with the *ELOVL5* gene knockdown. Decreased expression of ELOVL5 and IGFBP6 proteins in MDA-MB-231 cell lines with stable knockdown of these genes was additionally qualitatively confirmed by Western blotting (**Figure 2** and **Supplementary Figure 4**).

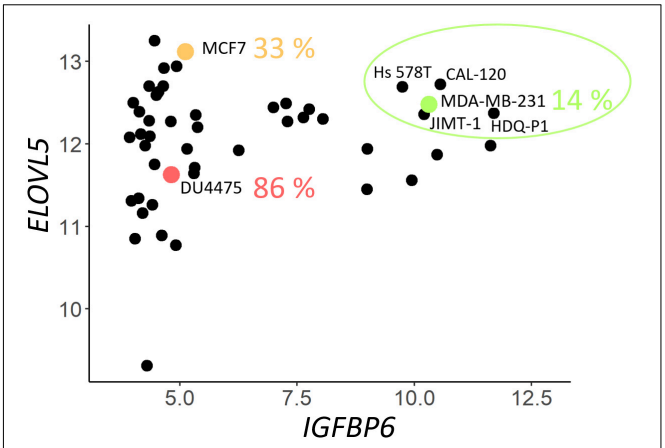


FIGURE 1 | Expression (on the Affymetrix logarithmic scale) of the ELOVL5 and IGFBP6 genes in breast cancer cell lines [according to Cancer Cell Line Encyclopedia (CCLE) data (Barretina et al., 2012)]. The proportion of cases when some of the cell lines form metastases to lungs *in vivo* (immunodeficient mice) is indicated according to Valentiner et al. (2005).

TABLE 1 | Relative expression of the *ELOVL5* gene in the cells with a stable knockdown (shRNA) of the *ELOVL5* gene compared to the control cells MDA-MB-231 (LUC).

Cell line	Relative expression	95 % confidence interval	*p*-value
MDA-MB-231 (ELOVL5_1)	*ELOVL5:* 0.532	0.382–0.753	0.026
MDA-MB-231 (ELOVL5_2)	*ELOVL5:* 0.244	0.160–0.369	0.031

Cell Proliferation

As a result of the analysis of the effect of knockdown of the *ELOVL5* and *IGFBP6* genes on the proliferation rate of MDA-MB-231 cells, it was shown (**Figure 3**) that the knockdown of the *IGFBP6* gene leads to a significant increase in the proliferation rate, while the knockdown of the *ELOVL5* gene did not statistically significantly change the proliferation rate.

Apoptosis

The analysis of the activation of apoptosis (**Figure 4** and **Supplementary Figure 5**) in MDA-MB-231 cells after knockdown of the *ELOVL5* gene showed that the number of dead cells, including the cells in the late stages of apoptosis, (AV+PI+) did not change in comparison to control cells ($p = 0.66$). At the same time, the knockdown of the *IGFBP6* gene led to a significant decrease in the proportion of dead (AV+PI+) cells in the population (by about three times,

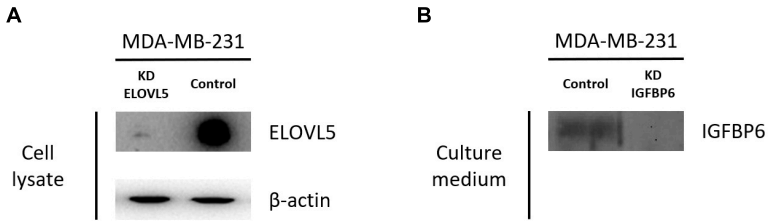


FIGURE 2 | Results of analysis of the expression of ELOVL5 and IGFBP6 proteins by Western blotting. **(A)** Relative expression of the ELOVL5 protein in the MDA-MB-231 (ELOVL5) cell line compared to the control cell line. **(B)** Relative content of the IGFBP6 protein in the conditioned cell culture medium of MDA-MB-231 (IGFBP6) cell line compared to the control cell line (samples were normalized to the number of cells).

p = 0.006). In addition, the proportion of viable cells in the population increased significantly (by about 11%, p = 0.009). Interestingly, the knockdown of the *IGFBP6* gene led to a decrease in the proportion of nuclear fragments without cell membrane (AV−PI+), which can be formed as a result of necrosis (from 2.9 to 0.6%, p = 0.02). No significant changes in the proportion of cells at an early stage of apoptosis as a result of the knockdown of the *ELOVL5* and *IGFBP6* genes were found (ANOVA, p = 0.24).

Cell Migration

As a result of the analysis of the effect of knockdown of the *ELOVL5* and *IGFBP6* genes on the migration activity of MDA-MB-231 cells by scratch assay, it was shown (**Figure 5**) that the knockdown of the *IGFBP6* gene leads to a significant decrease in migration activity (by about 27%, $p < 0.001$), while the knockdown of the *ELOVL5* gene leads to a similar, but less pronounced effect (migration activity decreases by about 15%, p = 0.029).

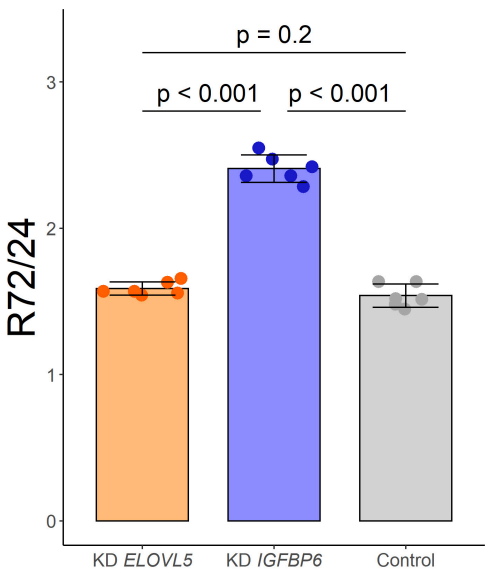


FIGURE 3 | Effect of the knockdown of *ELOVL5* and *IGFBP6* on the proliferation rate of MDA-MB-231 cells (data are mean ± SD, n = 6 wells per group). $R_{72}/24$− the ratio of the number of cells in a well after 72 h to the number of cells after 24 h from seeding.

Aggregation Into 3D Cell Spheroids

As a result of the analysis of the effect of knockdown of the *ELOVL5* and *IGFBP6* genes on the ability of MDA-MB-231 cells to form 3D spheroids, it was shown (**Figure 6** and **Supplementary Figure 6**) that the knockdown of the *IGFBP6* gene leads to the inability of cells to form 3D spheroids. The knockdown of the *ELOVL5* gene resulted in MDA-MB-231 cells forming less dense 3D spheroids with rough edges.

Transcriptomic Analysis

Our own transcriptomic analysis of the generated cell cultures demonstrated good correlation between replicates (**Supplementary Figure 7**) and showed (**Figure 7** and **Supplementary Table 4**) that the knockdown of the *ELOVL5* gene leads to a significant change in the expression of <2% of known genes, while the knockdown of the *IGFBP6* gene leads to a change in the expression of more than 16% of genes.

Among the genes with the most pronounced changes in mRNA expression, both in *ELOVL5* knockdown and *IGFBP6* knockdown cells, the *MMP1* and *MMP3* metalloproteinase mRNAs were found. After the knockdown of the *ELOVL5* gene, the content of mRNA of the *MMP1* and *MMP3* genes increased by 94 (FDR $p = 1.4 \times 10^{-12}$) and 7 (FDR $p = 4.3 \times 10^{-7}$) times, respectively, and after the knockdown of the *IGFBP6* gene by 244 (FDR $p = 4.3 \times 10^{-14}$) and 374 (FDR $p = 4.0 \times 10^{-16}$) times, respectively.

A significant change in the expression of the genes of MMPs *MMP1* and *MMP3* at mRNA level was additionally confirmed by RT-PCR. RT-PCR showed that the expression of the *MMP1* gene increased after the knockdown of the *ELOVL5* gene by about 76 times ($p < 0.001$) and after the knockdown of the *IGFBP6* gene by about 760 times (p = 0.028). It was not possible to quantify the ratio of *MMP3* gene expression levels in control cells and the cells with the knockdown of *ELOVL5* and *IGFBP6* genes using RT-PCR due to too low content of *MMP3* gene mRNA in control cells (the fluorescence intensity was below the threshold value after 40 amplification cycles). However, the registration of the PCR product was possible for cell lines with the knockdown of the *ELOVL5* and *IGFBP6* genes. Moreover, the threshold cycle value (Ct) for the cells with the knockdown of the *IGFBP6* gene

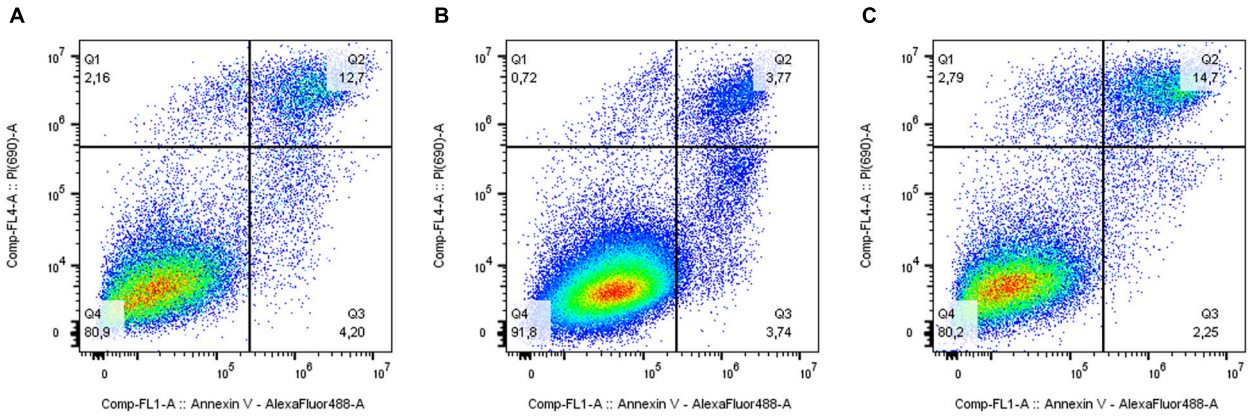


FIGURE 4 | Effect of *ELOVL5* and *IGFBP6* knockdown on the activation of apoptosis. Two-dimensional plots of the integral fluorescence intensity of the annexin V conjugate with Alexa Fluor 488 dye (horizontal axis) and the integral fluorescence intensity of propidium iodide (vertical axis) in MDA-MB-231 cells with **(A)** *ELOVL5* and **(B)** *IGFBP6* genes knockdown, as well as in **(C)** controls cells.

(31.9; standard deviation, 0.1) was significantly less than for the cells with the knockdown of the *ELOVL5* gene (37.9; standard deviation, 1.0). Thus, it can be seen from the obtained data that the results of analysis of the levels of expression of the *MMP1* and *MMP3* mRNA using real-time PCR are in good agreement with the results of transcriptomic analysis by Affymetrix chips.

Further analysis identified a group of 364 genes with statistically significant change in mRNA expression in the same direction, both in the *ELOVL5* gene knockdown and in the *IGFBP6* gene knockdown cells. It has been shown (**Supplementary Figure 8**) that the size of this overlap is too high to be a random event ($p < 0.001$). As a result of the analysis of biological processes enriched for the genes from this group, it was shown that among the genes with reduced expression, there is a significant number of genes involved in the formation of adherens junctions (**Table 2**). On the other hand, among the genes with increased expression, there are several genes involved in the formation of other types of intercellular contacts, as well as in the regulation of the formation of cell–cell contacts.

Among the most pronounced changes in the levels of expression of cell adhesion molecules, one can distinguish a strong decrease in the expression of the *CDH11* gene as a result of the knockdown of the *IGFBP6* gene (approximately 119 times, FDR $p = 9.1 \times 10^{-17}$). A similar, but significantly smaller change in the expression of the *CDH11* gene was also found after the knockdown of the *ELOVL5* gene (approximately 3.4 times, FDR $p = 1.5 \times 10^{-6}$). In addition, as a result of the knockdown of the *ELOVL5* and *IGFBP6* genes, the expression of the *CLDN1* (*ELOVL5*: 3.4 times, FDR $p = 9.2 \times 10^{-3}$; *IGFBP6*: 2.0 times, FDR $p = 4.2 \times 10^{-2}$) and *DSP* (*ELOVL5*: 1.7 times, FDR $p = 3.1 \times 10^{-3}$; *IGFBP6*: 4.9 times, FDR $p = 3.4 \times 10^{-11}$) genes consistently decreased.

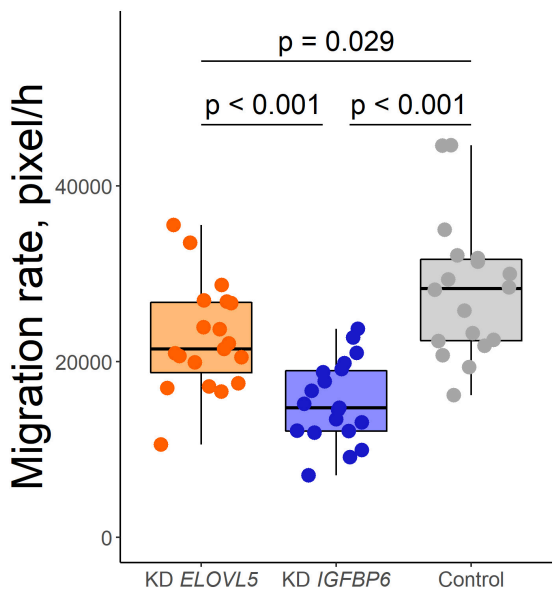


FIGURE 5 | Boxplot of cell migration activity according to the scratch wound healing assay ($n = 19$, $n = 19$, and $n = 18$ wells for KD *ELOVL5*, KD *IGFBP6*, and control cells, respectively).

Correlation Analysis

The analysis of the publicly available databases of transcriptomes of BC samples showed (**Supplementary Table 5**) that *MMP1* gene expression negatively correlates with *ELOVL5* gene expression (i.e., increases with a decrease in *ELOVL5* gene expression) in tumor samples from patients with ER+ BC in seven analyzed data sets (in total, 10 datasets of ER+ BC were analyzed) and in only one dataset of ER−BC patients (in total seven datasets of ER−BC were analyzed). In addition, *MMP1* gene expression negatively correlated with *IGFBP6* gene expression in the samples from four datasets of ER+BC and three datasets of ER−BC.

A weak negative correlation of the expression of the *MMP3* gene with the expression of the *ELOVL5* gene was observed only in one dataset of ER+BC. On the other hand, the correlation of the level of *MMP3* gene expression with the level of *IGFBP6* gene

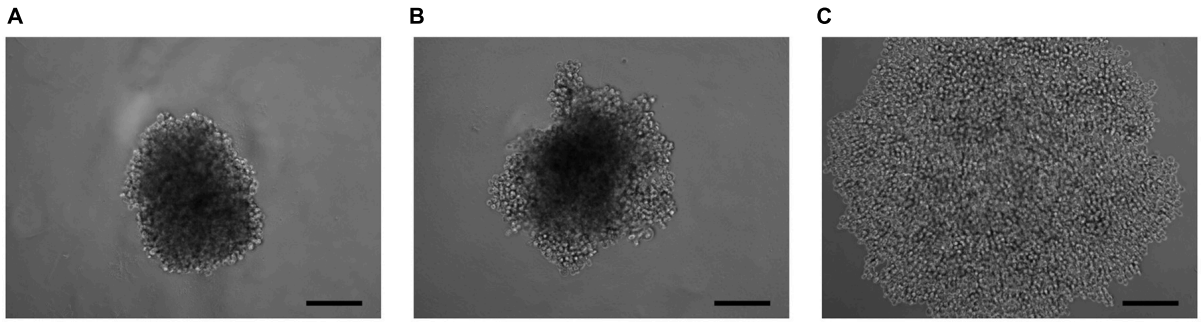


FIGURE 6 | Photo of 3D cell spheroids (5,000 cells per well at zero time point) after 96 h from seeding consisting of **(A)** control cells MDA-MB-231 (LUC) and the cells with a stable knockdown of **(B)** *ELOVL5* and **(C)** *IGFBP6* genes. The scale bar length is 200 μm.

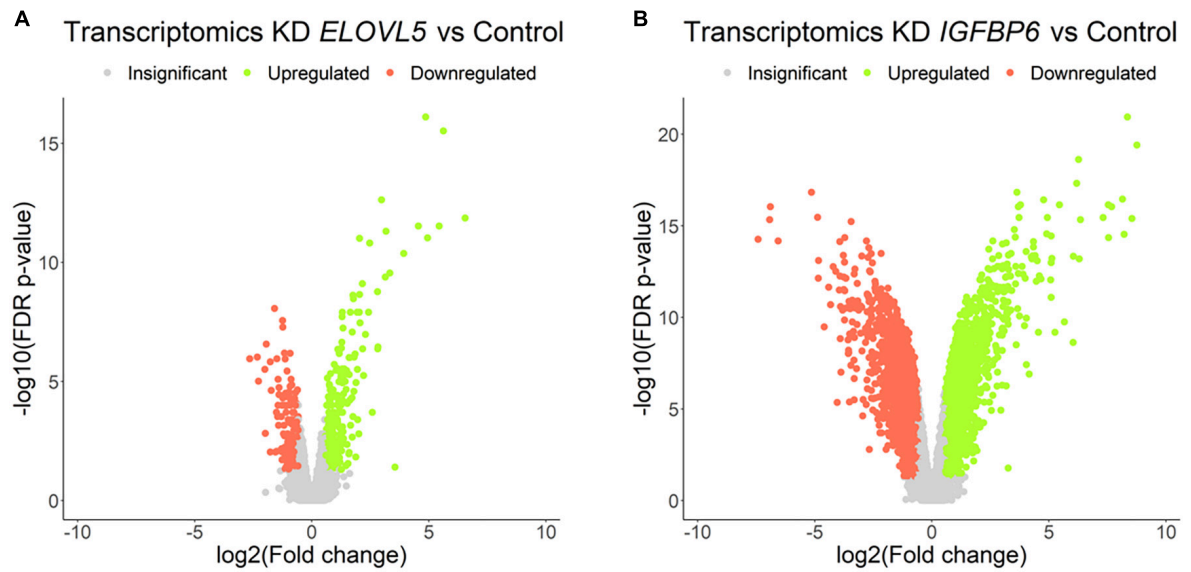


FIGURE 7 | Volcano plots for comparative transcriptome analysis of the cells with a stable knockdown of **(A)** *ELOVL5* and **(B)** *IGFBP6* genes. Thresholds: Fc = 1.5, FDR *p*-value = 0.05.

expression was positive in six datasets of ER+ tumors and in one dataset of ER− tumors.

According to the correlation analysis, the level of expression of the *CDH11* gene in the tumor tissue of patients with BC often positively correlates with the levels of expression of the *ELOVL5* and *IGFBP6* genes (*ELOVL5*: in four data sets of ER+ BC and in four data sets for ER− BC; *IGFBP6:* in seven data sets of ER+BC and three data sets of ER−BC). The only exception was the statistically significant weak negative correlation with the level of *IGFBP6* gene expression in one dataset of ER+ patients.

Proteomic Analysis

The proteomic analysis demonstrated good correlation between replicates (**Supplementary Figure 9**). As a result, it was shown (**Figure 8** and **Supplementary Table 6**) that the knockdown of the *ELOVL5* gene leads to an insignificant change in the expression of proteins in the cell (<0.2% of the total number of all measured proteins), while the knockdown of the *IGFBP6* gene leads to a significant change in expression of more than 20% of the measured proteins. The correlation analysis of the results (**Figure 9**) obtained during the study of transcriptomic and proteomic profiles revealed that in the case of the knockdown of the *IGFBP6* gene, there is a fairly high correlation of the results. In the case of the knockdown of the *ELOVL5* gene, despite the statistical significance, the correlation was low. This phenomenon can be explained by the lower sensitivity of proteomic analysis to small changes in expression as compared to a transcriptomic one.

The proteomic analysis confirmed a significant increase in the expression of the MMP1 protein after the knockdown of the *IGFBP6* gene (1,157 times, FDR $p = 3.5 \times 10^{-3}$) and a decrease in the DSP protein content (4.5-fold, FDR $p = 1.6 \times 10^{-3}$). At the same time, no decrease in the content of the OCLN protein was found (FDR $p = 0.33$). Moreover, the analysis of biological processes enriched for the proteins with altered expression after *IGFBP6* gene knockdown showed that among the genes with reduced expression, there is a significant group of genes involved

TABLE 2 | Selected enriched biological processes for the genes with a concordantly changed expression after knockdown of the *ELOVL5* and *IGFBP6* genes.

The genes with decreased expression		The genes with increased expression	
GO ID	**Biological processes**	**GO ID**	**Biological processes**
GO:0045216	Cell–cell junction organization	GO:0007156	Homophilic cell adhesion via plasma membrane adhesion molecules
GO:0034332	Adherens junction organization	GO:0016339	Calcium-dependent cell–cell adhesion via plasma membrane cell adhesion molecules
GO:0007157	Heterophilic cell–cell adhesion via plasma membrane cell adhesion molecules	GO:0007416	Synapse assembly
GO:0031102	Neuron projection regeneration	GO:2000651	Positive regulation of sodium ion transmembrane transporter activity
GO:0007155	Cell adhesion	GO:0045653	Negative regulation of megakaryocyte differentiation
GO:0003179	Heart valve morphogenesis	GO:1904837	Beta-catenin-TCF complex assembly
GO:0044331	Cell–cell adhesion mediated by cadherin	GO:0014829	Vascular smooth muscle contraction
GO:0060045	Positive regulation of cardiac muscle cell proliferation	GO:0032656	Regulation of interleukin-13 production
GO:0002921	Negative regulation of humoral immune response	GO:0022407	Regulation of cell–cell adhesion
GO:1903975	Regulation of glial cell migration	GO:0006335	DNA replication-dependent nucleosome assembly

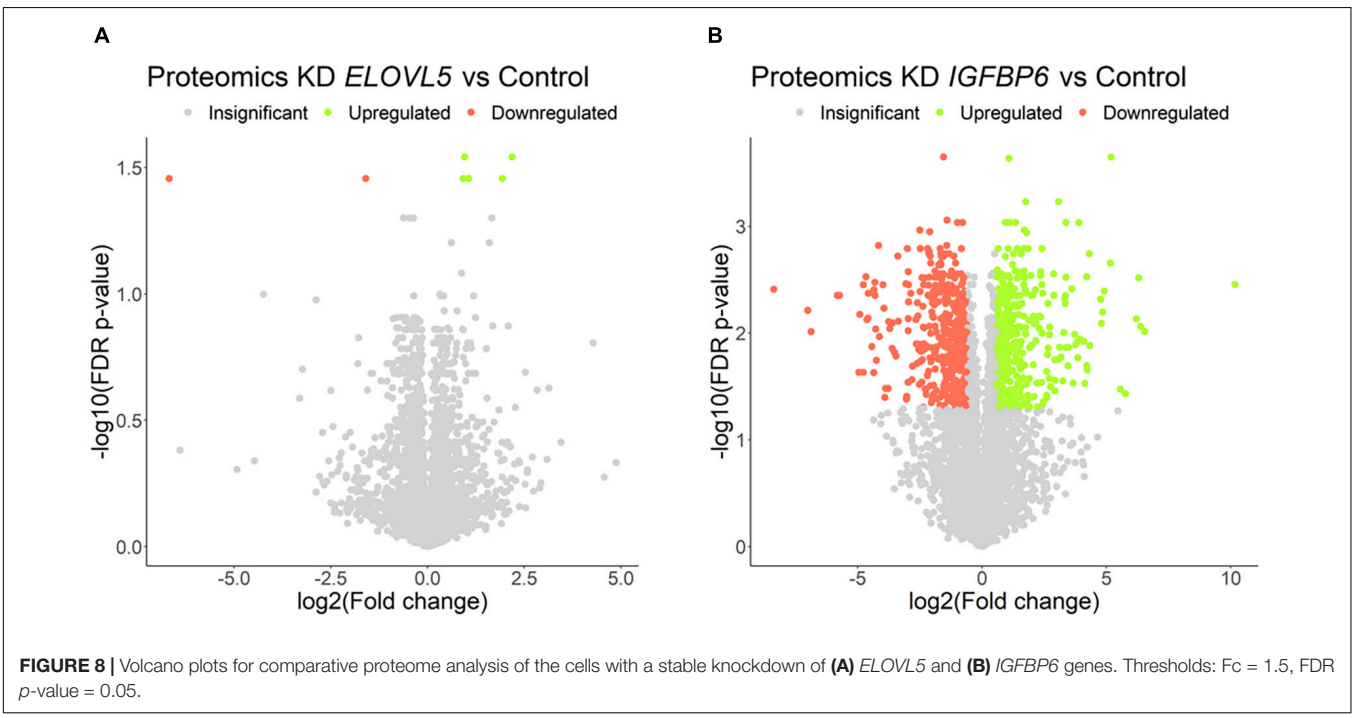


FIGURE 8 | Volcano plots for comparative proteome analysis of the cells with a stable knockdown of **(A)** *ELOVL5* and **(B)** *IGFBP6* genes. Thresholds: Fc = 1.5, FDR *p*-value = 0.05.

in cell migration and adhesion (**Table 3** and **Supplementary Table 7**). On the other hand, among the genes with increased expression, there are a lot of genes involved in cellular respiration and ribosome assembly.

Pathway Activation Analysis

Pathway activation levels were calculated for both experimental conditions (*ELOVL5* knockdown vs control, *IGFBP6* knockdown vs control) and both transcriptomics and proteomics data (**Supplementary Table 8**).

The most activated pathway after the knockdown of *ELOVL5* gene according to the microarray analysis was "Reactome basigin interactions main pathway" (FDR $p = 0.039$). Basigin (CD147) is a cell surface protein that can activate the production of MMPs by adjacent cells (Nabeshima et al., 2006). In addition, basigin is known to promote progression of various cancers (Kanekura and Chen, 2010). On the other hand, the only significantly downregulated pathway in the cells with reduced expression of *ELOVL5* was "Reactome nectin/necl trans heterodimerization main pathway" (FDR $p = 0.038$). Nectins are well known cell

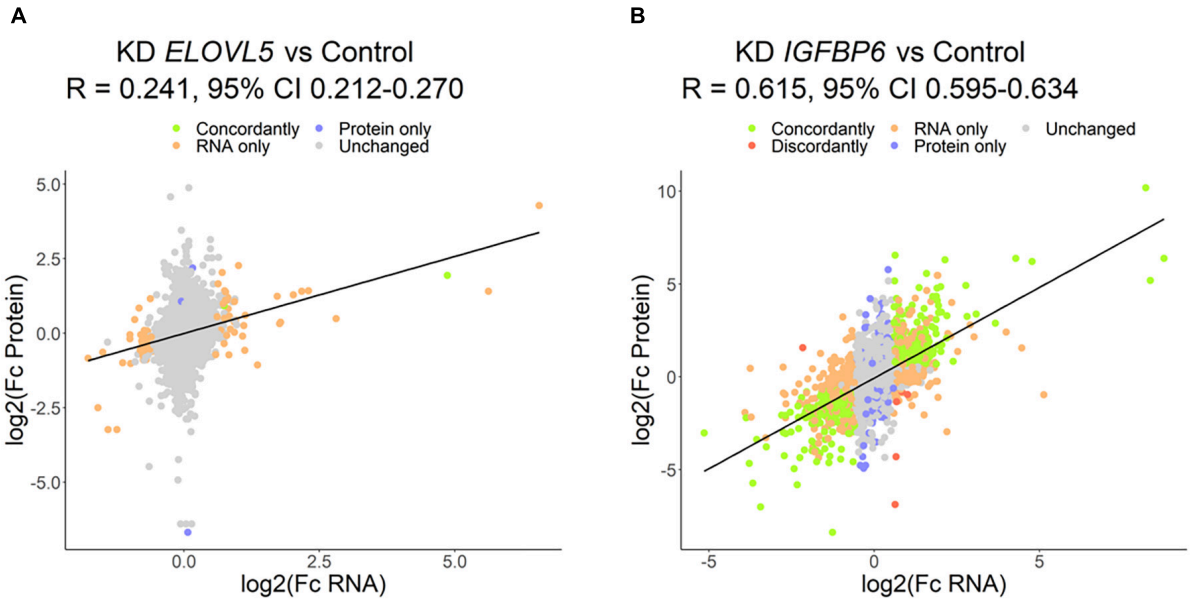


FIGURE 9 | Correlation of the results of proteomic and transcriptomic analysis of the cells with a stable knockdown of the **(A)** *ELOVL5* and **(B)** *IGFBP6* genes.

TABLE 3 | Selected enriched biological processes for the proteins with a significantly changed expression after knockdown of the *IGFBP6* gene.

The proteins with decreased expression		**The proteins with increased expression**	
GO ID	**Biological processes**	**GO ID**	**Biological processes**
GO:0044319	Wound healing, spreading of cells	GO:0000462	Maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)
GO:0022617	Extracellular matrix disassembly	GO:0031167	rRNA methylation
GO:1903779	Regulation of cardiac conduction	GO:0009303	rRNA transcription
GO:0010812	Negative regulation of cell–substrate adhesion	GO:0070125	mitochondrial translational elongation
GO:0010771	Negative regulation of cell morphogenesis involved in differentiation	GO:0045333	Cellular respiration
GO:0001666	Response to hypoxia	GO:0010501	RNA secondary structure unwinding
GO:0030239	Myofibril assembly	GO:0045943	Positive regulation of transcription by RNA polymerase I
GO:0051155	Positive regulation of striated muscle cell differentiation	GO:0030490	Maturation of SSU-rRNA
GO:0007186	G-protein-coupled receptor signaling pathway	GO:0070126	Mitochondrial translational termination
GO:1900024	Regulation of substrate adhesion-dependent cell spreading	GO:0007005	Mitochondrion organization

adhesion molecules, and downregulation of this pathway is consistent with reduced cell adhesion after the knockdown of *ELOVL5* gene (Sakisaka et al., 2007). According to the proteomics analysis, there were no downregulated pathways, while the most upregulated one was the "Biocarta Erk and PI-3 kinase are necessary for collagen binding in corneal epithelia pathway (actin filament stabilization)" (FDR $p = 0.048$), indicating importance of these signals in the progression of BC (Chu et al., 2000; Ebi et al., 2013).

Based on the transcriptomics data, the most activated pathway after the knockdown of *IGFBP6* gene was "NCI Class IB PI3K non-lipid kinase events pathway (cAMP biosynthetic process)" (FDR $p = 0.004$). It is well known that PI3K signaling is often deregulated in cancer. Specifically, class IB PI3K is important for the proliferation of pancreatic cancer cells (Edling et al., 2010). Our results suggest that class IB PI3K can be important for the proliferation of BC cells, too. The most activated pathway after the knockdown of *IGFBP6* according to proteomic analysis was the "NCI validated transcriptional targets of AP1 family members Fra1 and Fra2 main pathway" (FDR $p = 0.028$). Fra-1 and Fra-2 are well-studied transcription factors important for the progression of BC. For example, Fra-1 can directly increase the expression of *MMP1* (Belguise et al., 2005), and Fra-2 promotes the invasion of BC cells (Schröder et al., 2010). On the other hand, the most downregulated pathways in the cells with reduced expression of *IGFBP6* according to transcriptomic

analysis were integrin-linked kinase pathways ("ILK signaling pathway opsonization," FDR $p = 0.003$; "ILK signaling pathway cell adhesion," FDR $p = 0.004$; "ILK signaling pathway regulation of junction assembly at desmosomes," FDR $p = 0.004$; "ILK signaling pathway wound healing," FDR $p = 0.004$), which regulate cell adhesion, motility, and opsonization (Zheng et al., 2019). Downregulation of these pathways is consistent with observed reduced adhesion and motility of BC cells with the knockdown of *IGFBP6* gene. In addition, pathway analysis of proteomic data revealed inhibition of the "hypusine biosynthesis" pathway (FDR $p = 0.035$). Hypusine is a noncanonical amino acid containing only in two proteins: eIF5A1 and eIF5A2 (Muramatsu et al., 2016). Its modification leads to activation of the RhoA signaling pathway and increased cell motility (Muramatsu et al., 2016). Decreased cell migratory activity is consistent with the inhibition of this pathway.

Consistently with differential expression and GO terms enrichment analyses, the knockdown of *IGFBP6* led to a significantly higher number of altered pathways compared to *ELOVL5* case. Namely, 929 and 791 pathways were regulated upon *IGFBP6* knockdown for transcriptomics and proteomics data, respectively (adjusted $p < 0.05$), while only 5 and 3 pathways were identified upon *ELOVL5* knockdown. From these, three Reactome pathways were common for *ELOVL5* and *IGFBP6* knockdowns: "Reactome activation of MMPs main pathway" (upregulated upon both knockdowns, FDR $p = 0.037$ and FDR $p = 0.003$, respectively), "Reactome Basigin interactions main pathway" (upregulated upon both knockdowns, FDR $p = 0.039$ and FDR $p = 0.019$, respectively) and "Reactome Nectin/Necl trans heterodimerization main pathway" (downregulated upon both knockdowns, FDR $p = 0.038$ and FDR $p = 0.003$, respectively). While the analysis showed consistent results between transcriptomics and proteomics data upon *IGFBP6* knockdown (337 common activated pathways, $p = 1.4 \times 10^{-10}$), three pathways associated with proteomics of cells with *ELOVL5* knockdown had not intersected with other pathway sets.

Then, we analyzed alteration of the pathways that directly include *ELOVL5* and *IGFBP6* genes (**Table 4**). Specifically, *ELOVL5* was directly involved in the synthesis of very long-chain fatty acyl-CoAs and metabolism of linoleic and α-linolenic acids. According to the transcriptomic analysis, all these pathways were downregulated both upon *ELOVL5* and *IGFBP6* knockdowns; however, the majority of the differences were not statistically significant after multiple testing correction. The only exception was the "Reactome alpha linolenic acid ALA metabolism main pathway." It was significantly downregulated upon *IGFBP6* knockdown. The majority of the same pathways were not regulated according to the proteomic analysis. However, the "Reactome linoleic acid LA metabolism main pathway" was downregulated upon *ELOVL5* knockdown and upregulated upon *IGFBP6* knockdown at the protein level, but these changes were insignificant after multiple testing correction. On the other hand, only two pathways included *IGFBP6* gene, and one of them ("Reactome regulation of IGF activity by IGFBP main pathway") was activated in both knockdowns according to the proteomic analysis. However, after multiple testing correction, only the activation upon *IGFBP6* knockdown was significant. Overall, the pathway analysis indicates that the changes in the expression of one of the genes from the pair *ELOVL5–IGFBP6* can alter the pathways containing the other one; however, additional experiments are needed to prove this hypothesis.

Activity of MMPs

The increase in the expression of MMPs MMP1 (molecular weight, 43 kDa) and MMP3 (molecular weight, 45 kDa), detected by transcriptomic and proteomic analysis, was further confirmed by zymography assessment of the activity of MMPs (**Figure 10** and Supplementary **Figure 10**). It was shown that upon the knockdown of the *ELOVL5* gene, one pale band appears on the zymogram of the culture medium, which corresponds to the presence of active matrix metalloproteinase with a molecular weight of about 43 kDa (presumably MMP1). Upon the knockdown of the IGFBP6 gene, two bands appear on the zymogram of the culture medium, corresponding to the presence of active MMPs with molecular weights of about 43 and 45 kDa (presumably MMP1 and MMP3).

DISCUSSION

In this work, to study the changes in model tumor cells associated with the changes in the expression of the *ELOVL5* and *IGFBP6* genes, we decided to knock down the genes under consideration using RNA interference. We chose a stable knockdown with

TABLE 4 | The pathways containing *ELOVL5* and *IGFBP6* genes which were regulated upon knockdown of these genes.

Pathway	KD *ELOVL5*			KD *IGFBP6*		
	Direction	*p*-value	FDR *p*-value	Direction	*p*-value	FDR *p*-value
Transcriptomic analysis						
Reactome linoleic acid LA metabolism main pathway	Down	0.002	0.111	Down	0.045	0.100
Reactome alpha linolenic acid ALA metabolism main pathway	Down	0.022	0.177	Down	0.007	0.032
Reactome synthesis of very long chain fatty acyl-CoAs main pathway	Down	0.002	0.107	Down	0.030	0.077
Reactome regulation of IGF activity by IGFBP main pathway	Up	< 0.001	0.074	Up	0.056	0.115
Proteomic analysis						
Reactome linoleic acid LA metabolism main pathway	Down	0.025	0.339	Up	0.014	0.054
Reactome alpha linolenic acid ALA metabolism main pathway	Down	0.193	0.614	Down	0.476	0.703
Reactome synthesis of very long chain fatty acyl-CoAs main pathway	Down	0.236	0.659	Up	0.525	0.759
Reactome regulation of IGF activity by IGFBP main pathway	Up	0.013	0.255	Up	0.007	0.035

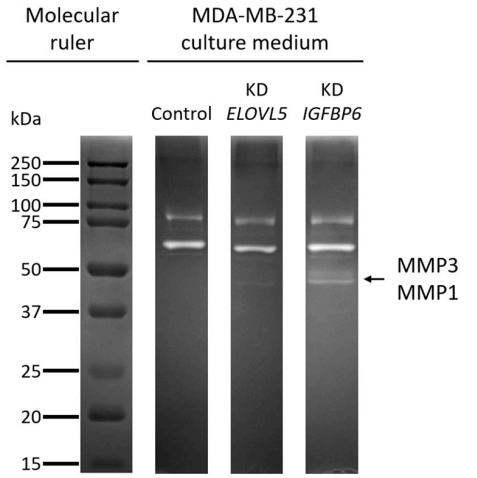


FIGURE 10 | Zymogram of cell culture medium samples obtained after incubation with control MDA-MB-231 (LUC) cells and the cells with a stable knockdown of *ELOVL5* and *IGFBP6* genes.

shRNA, as this method has been well developed to date and does not completely inhibit the expression of the selected gene. In contrast to various knockout methods, a stable shRNA-mediated knockdown partially decreases expression, which better reflects physiological changes observed *in vivo* (Boettcher and McManus, 2015). On the other hand, gene overexpression can often result in too high non-physiological levels of mRNA and protein of the selected gene, which also makes this method less suitable for this work (Prelich, 2012). Despite the fact that neither *ELOVL5* nor *IGFBP6* has substantial predictive power as single genes, we decided to perform separate knockdowns to reveal the impact of each of them on the behavior of BC cells and to find similarities and differences in their action. Therefore, it was necessary to choose a cell line with sufficiently high levels of expression of the studied genes.

According to a previously developed classification, such cells should form fewer metastases *in vivo* (Galatenko et al., 2015). This hypothesis is supported by previously published data on the ability of different BC cell lines to form lung metastases in immunodeficient mice *in vivo* (Valentiner et al., 2005). For example, MDA-MB-231 cells (green dot in **Figure 1**) formed metastases in lungs in only 14% of cases, while MCF7 cells (orange dot in **Figure 1**) with significantly lower *IGFBP6* gene expression formed metastases in 33% cases, and DU4475 cells (red dot in **Figure 1**), which, in addition, has a decreased expression of the *ELOVL5*, formed lung metastases in 86% cases (Valentiner et al., 2005).

Previously, we described generation of MDA-MB-231 cells with stable knockdown of *IGFBP6* gene (Nikulin et al., 2018). In this work, in addition, we generated MDA-MB-231 cells with stable knockdown of *ELOVL5* gene and analyzed different properties associated with metastatic potential.

Today, it is well known that the proliferation rate is an important indicator in assessing the metastatic potential of tumor cells. In particular, it was previously shown on model cell lines that an increase in the proliferation rate leads to an increase in the number of metastases *in vivo* (Hirabayashi et al., 1998; Marshall et al., 2004). These data are in agreement with clinical observations demonstrating that tumor doubling time, which depends on the rate of proliferation of tumor cells, affects relapse-free survival and overall survival (Tubiana, 1989). Our previous study revealed that the knockdown of *IGFBP6* increases the proliferation rate of MDA-MB-231 cells. This phenomenon can be explained by a classical IGF-dependent mechanism of action of the IGFBP6 protein. The knockdown of the *IGFBP6* gene leads to a decrease in the IGFBP6 protein content and to an increase in the content of free IGF-2 in the culture medium and stimulation of cell growth (Annunziata et al., 2011; Bach, 2016; Allard and Duan, 2018). In this work, we confirmed our previous finding and showed that, in contrast to the *IGFBP6* knockdown, the knockdown of *ELOVL5* has no effect on cell proliferation. An increase in the activity of cellular respiration and assembly of ribosomes found in the cells with the knockdown of the *IGFBP6* gene by means of proteomic analysis also indirectly indicates an increased proliferative activity. Moreover, we showed that the cells with *IGFBP6* mRNA knockdown are more resistant to apoptosis, probably also due to increased content of unbounded IGF-2.

Migration of tumor cells is an integral part of metastasis at almost every stage of the invasive-metastatic cascade, including local invasion. Previously, we demonstrated that knockdown of the *IGFBP6* gene leads to a sharp decrease in MDA-MB-231 cells' migration in the transwell assay (Nikulin et al., 2018). In this work, we confirmed this finding with the help of a scratch assay and also demonstrated a similar effect for the *ELOVL5* gene. Observed changes in migratory activity were consistent with the conducted pathway and enrichment analysis, based on the transcriptomic and proteomic data.

Interestingly, the decrease in migratory activity as a result of the knockdown of the *IGFBP6* gene was more pronounced in the experiment with the transwell membrane inserts when compared to the scratch assay. It should be noted that the mechanisms of cell migration in these tests are fundamentally different from each other. Thus, in the scratch assay, gradients of chemoattractants are absent, and cells move collectively, interacting with each other and with extracellular matrix (ECM) proteins (Liang et al., 2007; Jonkman et al., 2014). This collective migration better reflects the *in vivo* situation. On the other hand, when considering membrane insert, the cell must completely lose contact with other cells and significantly change their shape during the passage of the pore, while the movement occurs along the gradient of chemoattractants, since FBS is present in the lower chamber (Chen, 2005).

The data indicate that the knockdown of the *IGFBP6* mRNA has a stronger impact on the ability of single cells to migrate through narrow spaces in comparison to collective migration. These results are in good agreement with previously published data, indicating that invasion is associated with the arrest of the cell cycle, and therefore, the migration activity of rapidly proliferating cells should be lower (Kohrman and Matus, 2017). At the same time, previous modeling showed that the metastatic potential of cells with a high proliferation rate is often higher

than that of the cells with an increased ability to invade (Hecht et al., 2015).

From the classical point of view, malignant cells in the course of spreading through the body lose their contact with neighboring cells and become more mobile (Janiszewska et al., 2020). Thus, the loss of adhesion should be associated with a more aggressive phenotype. It is well known that a lot of adhesion molecules play an important role in the progression of cancer (Lange et al., 2014; Samatov et al., 2016). However, it is worth noting that cells often migrate collectively, and, in this case, they are not characterized by a complete loss of intercellular contacts (Janiszewska et al., 2020). The most striking example is the process of epithelial-to-mesenchymal transition (EMT), manifested by loss of E-cadherin and the acquisition of mobility by tumor cells (Nieto et al., 2016). The formation of spheroids by many BC cell lines is also known to depend on the expression of E-cadherin (Iglesias et al., 2013). At the same time, the level of E-cadherin expression in tumor tissue is significantly associated with the prognosis of many types of cancer, including BC (low level is associated with a poor prognosis) (Rossetti et al., 2015).

MDA-MB-231 cells express E-cadherin at a rather low level and form loose spheroids (Ivascu and Kubbies, 2007). It is known that the aggregation of cells into spheroids can also depend on other adhesion molecules, such as CDH3 (Stadler et al., 2018) and CD44 (Suarez et al., 2019). What is more, the expression of many adhesion molecules that can potentially participate in the formation of intercellular contacts in spheroids is associated with the prognosis of the disease. In particular, low expression of claudin (one of the structural components of tight junctions) in BC cells is associated with a poor prognosis (Rossetti et al., 2015). Thus, the study of intercellular adhesion can be useful in assessing the metastatic potential of cells.

The data obtained in this work indicate that, as a result of the knockdown of the *ELOVL5* and *IGFBP6* mRNAs in MDA-MB-231 cells, the expression of a number of adhesion molecules (such as *CDH11*, *CLDN1*, and *DSP*) decreases, which in turn leads to the disruption of cell–cell contacts. Furthermore, the relationship between the expression levels of the *CDH11* gene, the *ELOVL5* and *IGFBP6* genes, is the same in clinical BC samples as in our *in vitro* model. It is known that CDH11 is one of the classic type 2 cadherins, which plays an important role in the formation of intercellular contacts during osteogenesis (Piao et al., 2017). Interestingly, the increased expression of CDH11 can stimulate the invasion of some types of tumor cells (e.g., prostate cancer cells) and reduce the proliferation rate and ability to invade for other types of tumors (e.g., head and neck tumors). In this work, we ascertained that a decreased expression of the *CDH11* gene in BC cells may be associated with a more aggressive phenotype.

Matrix metalloproteinases are zinc-dependent extracellular endopeptidases involved in the remodeling of the ECM, both in normal conditions and in various pathologies, including malignant neoplasms (Gialeli et al., 2011; Cathcart et al., 2015). MMPs have different substrate specificities. In particular, MMP1 belongs to the family of collagenases, which predominantly break down various types of collagens and gelatin (denatured collagen), and MMP3 belongs to the family of stromelysins, which break down proteoglycans, laminins, fibronectin, and some types of collagens (Overall, 2002). It is also remarkable that several MMPs can regulate the availability of various growth factors to cells. For example, MMP1 can degrade the proteins IGFBP3 and IGFBP5, which bind IGFs, thereby increasing the concentration of the latter. MMP3 can also cleave IGFBP3, resulting in a similar effect.

To date, it is known that the increased expression of the MMP1 protein in tumor tissue is associated with metastatic lesions of lymph nodes in BC, and a decrease in MMP1 gene expression reduces the metastatic potential of BC cells both *in vitro* and *in vivo* in animal experiments (Liu et al., 2012; Wang et al., 2018). High MMP3 expression is also associated with a poor prognosis for BC (Mehner et al., 2015). Thus, the increase in MMP1 and MMP3 expression observed after the knockdown of the *ELOVL5* and *IGFBP6* genes is consistent with the hypothesis that ELOVL5 and IGFBP6 are associated with tumor metastatic potential. Moreover, the expression of the *MMP1* mRNA and the *ELOVL5–IGFBP6* pair of mRNAs is interrelated in patient tumors, and the direction of the change in expression is the same with our *in vitro* model. However, the conducted correlation analysis showed that the regulation of *MMP3* gene expression *in vivo* in patients' tumors may differ significantly from the pattern we observed *in vitro*.

Overall, the knockdown of *ELOVL5* had a number of seemingly unexpected consequences. For example, the decreased expression of the enzyme involved in fatty acids (FAs) elongation influenced cell migration, cell–cell interactions, and MMP's synthesis. However, this is not an entirely unexpected result, as previously, it was shown that omega-3 and omega-6 PUFAs, the products of ELOVL5 activity, affect the proliferation, migration, and invasion of cancer cells *in vitro* (Chamras et al., 2002; Yun et al., 2014; Gonzalez-Reyes et al., 2017; Huang et al., 2017) and that dietary omega-3 FAs reduce the risk of BC development, as well as the risk of its relapse (Abdelmagid et al., 2016; Playdon et al., 2017; Romieu et al., 2017; Shapira, 2017). Still, there is no simple explanation for these results, since the effect of PUFAs on cellular processes is multifaceted.

First of all, PUFAs are incorporated into membrane phospholipids and influence their fluidity and selective permeability and functioning of membrane receptors (Wiktorowska-Owczarek et al., 2015). PUFAs also can modulate the activity of different transcriptional factors (Jump et al., 1996). Furthermore, docosahexaenoic acid (DHA, omega-3 PUFA) and arachidonic acid (AA, omega-6 PUFA) have a wide range of bioactive metabolites acting as local hormones or signaling molecules and regulate cell proliferation, adhesion, migration, angiogenesis, vascular permeability, and inflammatory responses [role of DHA metabolites resolvins, protectins, and maresins is reviewed in Kuda (2017) and the role of AA metabolites eicosanoids, prostaglandins, and leukotrienes is reviewed in Tallima and El Ridi (2018)]. There is also evidence of the direct inhibition of MMPs activity by PUFAs (Nicolai et al., 2017), although information on this issue is controversial (Liuzzi et al., 2007). The inhibition effect of omega-6 PUFAs on *MMPs* expression was shown *in vivo* in a coronary heart disease-induced rat model (Lu et al., 2018), but the mechanism is unclear.

The association between IGFBP6 and cancer appears to be more obvious, as a large number of studies on the role of the IGF/IGF1R signaling pathway in oncogenesis have been carried out to date (Trajkovic-Arsic et al., 2013; Brouwer-Visser and Huang, 2015; Salisbury and Tomblin, 2015; Vigneri et al., 2015; Tracz et al., 2016). At the same time, significant differences in the primary structure of seven IGFBPs, in their posttranslational modifications and in their tissue specificity, indicate differences in their functions. Differences in IGFBPs structures also indicate that their action is not limited to the inhibition of IGFs. This is confirmed by the fact that, for some IGFBPs, the IGF- and IGF1R-independent action on cells has been demonstrated (Firth and Baxter, 2002).

For many IGFBPs, their role in various pathological processes was demonstrated, including cancer [the *IGFBP6* gene is differentially expressed in nasopharyngeal carcinoma (Chen et al., 2016); the plasma protein level of IGFBP6 changes with ovarian cancer (Gunawardana et al., 2009; Wang et al., 2013); *IGFBP6* mRNA and protein levels are significantly lower in colorectal cancer (CRC) tissues and low *IGFBP6* expression correlated with poor overall survival (Zhao et al., 2020)]. Knockdown of *IGFBP6* in HT-29, Caco-2, SW620, and HCT116 cells influenced proliferation, migration, and invasion (Zhao et al., 2020).

It was shown that IGFBP6 acts on different cancer cell lines both by the inhibition of IGFs and by IGF-independent mechanisms in an autocrine and/or paracrine fashion. Information about IGFBP6 and its effects on cellular processes is reviewed in Bach (2016). Interestingly, that IGFBP6 contains a nuclear localization signal, which targets it to the nucleus, where it regulates gene expression (Poreba and Durzynska, 2020). IGFBP6 showed its ability to bind the *EGR1* promoter and induce its activity in stably transfected nasopharyngeal cancer (NPC) cell lines overexpressing IGFBP6 (Kuo et al., 2010). The increased expression of IGFPB6 inhibited the proliferation, invasion, and metastatic activity of the NPC cells, suggesting that IGFBP6 acts as a tumor suppressor (Kuo et al., 2010). In our study, expression of *TOE1* (target of EGR1, member 1) after *IGFBP6* knockdown increased 2.7 times (FDR $p = 6.2 \times 10^{-6}$), suggesting this mechanism can play an important role in BC, too.

CONCLUSION

In conclusion, it can be assumed that low expression of the *ELOVL5* and *IGFBP6* genes leads to the stimulation of BC cell invasion at the first stage of the invasive metastatic cascade due to the increased proliferation rate, more efficient decomposition of the ECM by MMPs, and the weakening of cellular junctions. Increased resistance to apoptosis may also play an important role in the spreading of tumor cells throughout the body. Further research will help shed light on the detailed molecular mechanisms responsible for the observed changes in tumor cell properties resulting from a decreased expression of the *ELOVL5* and *IGFBP6* genes.

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 below: https://www.ebi.ac.uk/pride/archive/, PXD023892 and https://www.ncbi.nlm.nih.gov/geo/, GSE165854.

AUTHOR CONTRIBUTIONS

AT supervised the study. SNi, GZ, AP, US, and AT designed the study. SNi, GZ, MR, DW, SNe, JB, GB, and LA performed the experiments and analyzed the data. SNi and GZ wrote the manuscript. AT and AP revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The article was prepared within the framework of the HSE University Basic Research Program.

ACKNOWLEDGMENTS

JB was supported by Swedish Research Council no. 2015-4870. GB was supported by the Vetenskapsrådet-Swedish Science Research Council, 2019-01771-3. The authors thank Vladimir Galatenko and Maxim Shkurnikov for the discussion of the obtained results. The authors also thank Ganna Shevchenko and the MS facility for Proteomics at the Uppsala University for assistance with the proteomics analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.662843/full#supplementary-material

REFERENCES

Abdelmagid, S. A., MacKinnon, J. L., Janssen, S. M., and Ma, D. W. L. (2016). Role of n-3 polyunsaturated fatty acids and exercise in breast cancer prevention: identifying common targets. *Nutr. Metab. Insights* 9:NMI.S39043. doi: 10.4137/NMI.S39043

Allard, J. B., and Duan, C. (2018). IGF-binding proteins: why do they exist and why are there so many? *Front. Endocrinol. (Lausanne)* 9:117. doi: 10.3389/fendo.2018.00117

Annunziata, M., Granata, R., and Ghigo, E. (2011). The IGF system. *Acta Diabetol.* 48, 1–9. doi: 10.1007/s00592-010-0227-z

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

Bach, L. A. (2015). Recent insights into the actions of IGFBP-6. *J. Cell Commun. Signal.* 9, 189–200. doi: 10.1007/s12079-015-0288-4

Bach, L. A. (2016). Current ideas on the biology of IGFBP-6: more than an IGF-II inhibitor? *Growth Horm. IGF Res.* 30–31, 81–86.

Bach, L. A., Fu, P., and Yang, Z. (2013). Insulin-like growth factor-binding protein-6 and cancer. *Clin. Sci.* 124, 215–229. doi: 10.1042/CS20120343

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607. doi: 10.1038/nature11003

Bartlett, J. M. S., Bayani, J., Marshall, A., Dunn, J. A., Campbell, A., Cunningham, C., et al. (2016). Comparing breast cancer multiparameter tests in the OPTIMA prelim trial: no test is more equal than the others. *J. Natl. Cancer Inst.* 108:djw050. doi: 10.1093/jnci/djw050

Belguise, K., Kersual, N., Galtier, F., and Chalbos, D. (2005). FRA-1 expression level regulates proliferation and invasiveness of breast cancer cells. *Oncogene* 24, 1434–1444. doi: 10.1038/sj.onc.1208312

Boettcher, M., and McManus, M. T. (2015). Choosing the right tool for the job: RNAi, TALEN, or CRISPR. *Mol. Cell* 58, 575–585. doi: 10.1016/j.molcel.2015.04.028

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Brouwer-Visser, J., and Huang, G. S. (2015). IGF2 signaling and regulation in cancer. *Cytokine Growth Factor Rev.* 26, 371–377. doi: 10.1016/j.cytogfr.2015.01.002

Camps, C., Buffa, F. M., Colella, S., Moore, J., Sotiriou, C., Sheldon, H., et al. (2008). Hsa-miR-210 is induced by hypoxia and is an independent prognostic factor in breast cancer. *Clin. Cancer Res.* 14, 1340–1348. doi: 10.1158/1078-0432.CCR-07-1755

Cathcart, J., Pulkoski-Gross, A., and Cao, J. (2015). Targeting matrix metalloproteinases in cancer: bringing new life to old ideas. *Genes Dis.* 2, 26–34. doi: 10.1016/j.gendis.2014.12.002

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404. doi: 10.1158/2159-8290.CD-12-0095

Chamras, H., Ardashian, A., Heber, D., and Glaspy, J. A. (2002). Fatty acid modulation of MCF-7 human breast cancer cell proliferation, apoptosis and differentiation. *J. Nutr. Biochem* 13, 711–716. doi: 10.1016/S0955-2863(02)00230-9

Chen, H.-C. (2005). "Boyden chamber assay," in *Cell Migration*, ed. J. L. Guan (Totowa, NJ: Humana Press), 15–22. doi: 10.1385/1-59259-860-9:015

Chen, Q., Qin, S., Liu, Y., Hong, M., Qian, C.-N., Keller, E. T., et al. (2016). IGFBP6 is a novel nasopharyngeal carcinoma prognostic biomarker. *Oncotarget* 7, 68140–68150. doi: 10.18632/oncotarget.11886

Cheng, S. H.-C., Huang, T.-T., Cheng, Y.-H., Tan, T. B. K., Horng, C.-F., Wang, Y. A., et al. (2017). Validation of the 18-gene classifier as a prognostic biomarker of distant metastasis in breast cancer. *PLoS One* 12:e0184372. doi: 10.1371/journal.pone.0184372

Chu, C. L., Reenstra, W. R., Orlow, D. L., and Svoboda, K. K. (2000). Erk and PI-3 kinase are necessary for collagen binding and actin reorganization in corneal epithelia. *Invest. Ophthalmol. Vis. Sci.* 41, 3374–3382.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Dai, X., Cheng, H., Bai, Z., and Li, J. (2017). Breast cancer cell line classification and its relevance with breast tumor subtyping. *J. Cancer* 8, 3131–3141. doi: 10.7150/jca.18457

Ebi, H., Costa, C., Faber, A. C., Nishtala, M., Kotani, H., Juric, D., et al. (2013). PI3K regulates MEK/ERK signaling in breast cancer via the Rac-GEF, P-Rex1. *Proc. Natl. Acad. Sci.U.S.A.* 110, 21124–21129. doi: 10.1073/pnas.1314124110

Edling, C. E., Selvaggi, F., Buus, R., Maffucci, T., Di Sebastiano, P., Friess, H., et al. (2010). Key role of phosphoinositide 3-kinase class ib in pancreatic cancer. *Clin. Cancer Res.* 16, 4928–4937. doi: 10.1158/1078-0432.CCR-10-1210

Firth, S. M., and Baxter, R. C. (2002). Cellular actions of the insulin-like growth factor binding proteins. *Endocr. Rev.* 23, 824–854. doi: 10.1210/er.2001-0033

Galatenko, V. V., Maltseva, D. V., Galatenko, A. V., Rodin, S., and Tonevitsky, A. G. (2018). Cumulative prognostic power of laminin genes in colorectal cancer. *BMC Med. Genomics* 11:9. doi: 10.1186/s12920-018-0332-3

Galatenko, V. V., Shkurnikov, M. Y., Samatov, T. R., Galatenko, A. V., Mityakina, I. A., Kaprin, A. D., et al. (2015). Highly informative marker sets consisting of genes with low individual degree of differential expression. *Sci. Rep.* 5:14967. doi: 10.1038/srep14967

Gene, T., and Consortium, O. (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Gerber, B., Freund, M., and Reimer, T. (2010). Recurrent breast cancer: treatment strategies for maintaining and prolonging good quality of life. *Dtsch. Arztebl. Int.* 107, 85–91. doi: 10.3238/arztebl.2010.0085

Gialeli, C., Theocharis, A. D., and Karamanos, N. K. (2011). Roles of matrix metalloproteinases in cancer progression and their pharmacological targeting. *FEBS J.* 278, 16–27. doi: 10.1111/j.1742-4658.2010.07919.x

Gonzalez-Reyes, C., Marcial-Medina, C., Cervantes-Anaya, N., Cortes-Reynosa, P., and Salazar, E. P. (2017). Migration and invasion induced by linoleic acid are mediated through fascin in MDA-MB-231 breast cancer cells. *Mol. Cell. Biochem.* 443, 1–10. doi: 10.1007/s11010-017-3205-8

Gunawardana, C. G., Kuk, C., Smith, C. R., Batruch, I., Soosaipillai, A., and Diamandis, E. P. (2009). Comprehensive analysis of conditioned media from ovarian cancer cell lines identifies novel candidate markers of epithelial ovarian cancer. *J. Proteome Res.* 8, 4705–4713. doi: 10.1021/pr900411g

Győrffy, B., Lanczky, A., Eklund, A. C., Denkert, C., Budczies, J., Li, Q., et al. (2010). An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res. Treat.* 123, 725–731. doi: 10.1007/s10549-009-0674-9

Hecht, I., Natan, S., Zaritsky, A., Levine, H., Tsarfaty, I., and Ben-Jacob, E. (2015). The motility-proliferation-metabolism interplay during metastatic invasion. *Sci. Rep.* 5:13538. doi: 10.1038/srep13538

Hirabayashi, K., Numa, F., Suminami, Y., Murakami, A., Murakami, T., and Kato, H. (1998). Altered proliferative and metastatic potential associated with increased expression of syndecan-1. *Tumor Biol.* 19, 454–463. doi: 10.1159/000030037

Huang, L.-H., Chung, H.-Y., and Su, H.-M. (2017). Docosahexaenoic acid reduces sterol regulatory element binding protein-1 and fatty acid synthase expression and inhibits cell proliferation by inhibiting pAkt signaling in a human breast cancer MCF-7 cell line. *BMC Cancer* 17:890. doi: 10.1186/s12885-017-3936-7

Hyams, D. M., Schuur, E., Angel Aristizabal, J., Bargallo Rocha, J. E., Cabello, C., Elizalde, R., et al. (2017). Selecting postoperative adjuvant systemic therapy for early stage breast cancer: a critical assessment of commercially available gene expression assays. *J. Surg. Oncol.* 115, 647–662. doi: 10.1002/jso.24561

Iglesias, J. M., Beloqui, I., Garcia-Garcia, F., Leis, O., Vazquez-Martin, A., Eguiara, A., et al. (2013). Mammosphere formation in breast carcinoma cell lines depends upon expression of e-cadherin. *PLoS One* 8:e77281. doi: 10.1371/journal.pone.0077281

Ivascu, A., and Kubbies, M. (2007). Diversity of cell-mediated adhesions in breast cancer spheroids. *Int. J. Oncol.* 31, 1403–1413. doi: 10.3892/ijo.31.6.1403

Janiszewska, M., Primi, M. C., and Izard, T. (2020). Cell adhesion in cancer: Beyond the migration of single cells. *J. Biol. Chem.* 295, 2495–2505. doi: 10.1074/jbc.REV119.007759

Jonkman, J. E. N., Cathcart, J. A., Xu, F., Bartolini, M. E., Amon, J. E., Stevens, K. M., et al. (2014). An introduction to the wound healing assay using live-cell microscopy. *Cell Adh. Migr.* 8, 440–451. doi: 10.4161/cam.36224

Jump, D. B., Clarke, S. D., Thelen, A., Liimatta, M., Ren, B., and Badin, M. (1996). Dietary polyunsaturated fatty acid regulation of gene transcription. *Prog. Lipid Res.* 35, 227–241. doi: 10.1016/S0163-7827(96)00007-0

Kanekura, T., and Chen, X. (2010). CD147/basigin promotes progression of malignant melanoma and other cancers. *J. Dermatol. Sci.* 57, 149–154. doi: 10.1016/j.jdermsci.2009.12.008

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996–1006. doi: 10.1101/gr.229102

Kohrman, A. Q., and Matus, D. Q. (2017). Divide or conquer: cell cycle regulation of invasive behavior. *Trends Cell Biol.* 27, 12–25. doi: 10.1016/j.tcb.2016.08.003

Kuda, O. (2017). Bioactive metabolites of docosahexaenoic acid. *Biochimie* 136, 12–20. doi: 10.1016/j.biochi.2017.01.002

Kuo, Y.-S., Tang, Y.-B., Lu, T.-Y., Wu, H.-C., and Lin, C.-T. (2010). IGFBP-6 plays a role as an oncosuppressor gene in NPC pathogenesis through regulating EGR-1 expression. *J. Pathol.* 222, 299–309. doi: 10.1002/path.2735

Kwa, M., Makris, A., and Esteva, F. J. (2017). Clinical utility of gene-expression signatures in early stage breast cancer. *Nat. Rev. Clin. Oncol.* 14, 595–610. doi: 10.1038/nrclinonc.2017.74

Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M.-C., Dossus, L., and Rondeau, V. (2018). Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. *BMC Cancer* 18:171. doi: 10.1186/s12885-018-4076-4

Lambert, A. W., Pattabiraman, D. R., and Weinberg, R. A. (2017). Emerging biological principles of metastasis. *Cell* 168, 670–691. doi: 10.1016/j.cell.2016.11.037

Lange, T., Samatov, T. R., Tonevitsky, A. G., and Schumacher, U. (2014). Importance of altered glycoprotein-bound N- and O-glycans for epithelial-to-mesenchymal transition and adhesion of cancer cells. *Carbohydr. Res.* 389, 39–45. doi: 10.1016/j.carres.2014.01.010

Leonard, A. E., Bobik, E. G., Dorado, J., Kroeger, P. E., Chuang, L. T., Thurmond, J. M., et al. (2000). Cloning of a human cDNA encoding a novel enzyme involved in the elongation of long-chain polyunsaturated fatty acids. *Biochem. J.* 350 Pt 3, 765–770. doi: 10.1042/bj3500765

Liang, C.-C., Park, A. Y., and Guan, J.-L. (2007). In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. *Nat. Protoc.* 2, 329–333. doi: 10.1038/nprot.2007.30

Liu, H., Kato, Y., Erzinger, S. A., Kiriakova, G. M., Qian, Y., Palmieri, D., et al. (2012). The role of MMP-1 in breast cancer growth and metastasis to the brain in a xenograft model. *BMC Cancer* 12:583. doi: 10.1186/1471-2407-12-583

Liuzzi, G. M., Latronico, T., Rossano, R., Viggiani, S., Fasano, A., and Riccio, P. (2007). Inhibitory effect of polyunsaturated fatty acids on MMP-9 release from microglial cells - Implications for complementary multiple sclerosis treatment. *Neurochem. Res.* 32, 2184–2193. doi: 10.1007/s11064-007-9415-9

Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., Tutt, A. M., et al. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics* 9:239. doi: 10.1186/1471-2164-9-239

Lu, N., Du, Y., Li, H., Luo, Y., Ouyang, B., Chen, Y., et al. (2018). Omega-6 fatty acids down-regulate matrix metalloproteinase expression in a coronary heart disease-induced rat model. *Int. J. Exp. Pathol.* 99, 210–217. doi: 10.1111/iep.12293

Maltseva, D. V., Khaustova, N. A., Fedotov, N. N., Matveeva, E. O., Lebedev, A. E., Shkurnikov, M. U., et al. (2013). High-throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples. *J. Clin. Bioinformatics* 3:13. doi: 10.1186/2043-9113-3-13

Maltseva, D., Raygorodskaya, M., Knyazev, E., Zgoda, V., Tikhonova, O., Zaidi, S., et al. (2020). Knockdown of the α5 laminin chain affects differentiation of colorectal cancer cells and their sensitivity to chemotherapy. *Biochimie* 174, 107–116. doi: 10.1016/j.biochi.2020.04.016

Marshall, J.-C., Caissie, A. L., Callejo, S. A., Antecka, E., and Burnier, M. N. Jr. (2004). Cell proliferation profile of five human uveal melanoma cell lines of different metastatic potential. *Pathobiology* 71, 241–245. doi: 10.1159/000080057

Mehner, C., Miller, E., Nassar, A., Bamlet, W. R., Evette, S., and Radisky, D. C. (2015). Tumor cell expression of MMP3 as a prognostic factor for poor survival in pancreatic, pulmonary, and mammary carcinoma. *Genes Cancer* 6:480. doi: 10.18632/genesandcancer.90

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., et al. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci.U.S.A.* 102, 13550–13555. doi: 10.1073/pnas.0506230102

Moon, Y., Hammer, R. E., and Horton, J. D. (2009). Deletion of ELOVL5 leads to fatty liver through activation of SREBP-1c in mice. *J. Lipid Res.* 50, 412–423. doi: 10.1194/jlr.M800383-JLR200

Muramatsu, T., Kozaki, K., Imoto, S., Yamaguchi, R., Tsuda, H., Kawano, T., et al. (2016). The hypusine cascade promotes cancer progression and metastasis through the regulation of RhoA in squamous cell carcinoma. *Oncogene* 35, 5304–5316. doi: 10.1038/onc.2016.71

Nabeshima, K., Iwasaki, H., Koga, K., Hojo, H., Suzumiya, J., and Kikuchi, M. (2006). Emmprin (basigin/CD147): matrix metalloproteinase modulator and multifunctional cell recognition molecule that plays a critical role in cancer progression. *Pathol. Int.* 56, 359–367. doi: 10.1111/j.1440-1827.2006.01972.x

Nicolai, E., Sinibaldi, F., Sannino, G., Laganà, G., Basoli, F., Licoccia, S., et al. (2017). Omega-3 and Omega-6 Fatty acids act as inhibitors of the matrix metalloproteinase-2 and matrix metalloproteinase-9 activity. *Protein J.* 36, 278–285. doi: 10.1007/s10930-017-9727-9

Nieto, M. A., Huang, R. Y.-J., Jackson, R. A., and Thiery, J. P. (2016). EMT: 2016. *Cell* 166, 21–45. doi: 10.1016/j.cell.2016.06.028

Nikitin, D., Garazha, A., Sorokin, M., Penzar, D., Tkachev, V., Markov, A., et al. (2019). Retroelement—linked transcription factor binding patterns point to quickly developing molecular pathways in human evolution. *Cells* 8:130. doi: 10.3390/cells8020130

Nikulin, S. V., Raigorodskaya, M. P., Poloznikov, A. A., Zakharova, G. S., Schumacher, U., Wicklein, D., et al. (2018). In vitro model for studying of the role of IGFBP6 gene in breast cancer metastasizing. *Bull. Exp. Biol. Med.* 164, 688–692. doi: 10.1007/s10517-018-4060-7

Overall, C. M. (2002). Molecular determinants of metalloproteinase substrate specificity: matrix metalloproteinase substrate binding domains, modules, and exosites. *Mol. Biotechnol.* 22, 51–86. doi: 10.1385/MB:22:1:051

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., McQuisten, K. A., Almabrazi, H. G., et al. (2008). IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res.* 36, W163–W169. doi: 10.1093/nar/gkn198

Pfaffl, M. W., Horgan, G. W., and Dempfle, L. (2002). Relative expression software tool (REST) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. *Nucleic Acids Res.* 30:e36. doi: 10.1093/nar/30.9.e36

Piao, S., Inglehart, R. C., Scanlon, C. S., Russo, N., Banerjee, R., and D'Silva, N. J. (2017). CDH11 inhibits proliferation and invasion in head and neck cancer. *J. Oral Pathol. Med.* 46, 89–97. doi: 10.1111/jop.12471

Playdon, M. C., Ziegler, R. G., Sampson, J. N., Stolzenberg-Solomon, R., Thompson, H. J., Irwin, M. L., et al. (2017). Nutritional metabolomics and breast cancer risk in a prospective study. *Am. J. Clin. Nutr.* 106, 637–649. doi: 10.3945/ajcn.116.150912

Poreba, E., and Durzynska, J. (2020). Nuclear localization and actions of the insulin-like growth factor 1 (IGF-1) system components: transcriptional regulation and DNA damage response. *Mutat. Res. Rev. Mutat. Res.* 784:108307. doi: 10.1016/j.mrrev.2020.108307

Prelich, G. (2012). Gene overexpression: uses, mechanisms, and interpretation. *Genetics* 190, 841–854. doi: 10.1534/genetics.111.136911

Redig, A. J., and McAllister, S. S. (2013). Breast cancer as a systemic disease: a view of metastasis. *J. Intern. Med.* 274, 113–126. doi: 10.1111/joim.12084

Romieu, I. I., Amadou, A., and Chajes, V. (2017). The role of diet, physical activity, body fatness, and breastfeeding in breast cancer in young women: epidemiological evidence. *Rev. Investig. Clin.* 69, 193–203. doi: 10.24875/RIC.17002263

Rossetti, C., Reis, B., da, C. A. A., Delgado, P., de, O., Azzalis, L. A., et al. (2015). Adhesion molecules in breast carcinoma: a challenge to the pathologist. *Rev. Assoc. Med. Bras.* 61, 81–85. doi: 10.1590/1806-9282.61.01.081

Sakisaka, T., Ikeda, W., Ogita, H., Fujita, N., and Takai, Y. (2007). The roles of nectins in cell adhesions: cooperation with other cell adhesion molecules and growth factor receptors. *Curr. Opin. Cell Biol.* 19, 593–602. doi: 10.1016/j.ceb.2007.09.007

Salisbury, T. B., and Tomblin, J. K. (2015). Insulin/insulin-like growth factors in cancer: new roles for the aryl hydrocarbon receptor, tumor resistance mechanisms, and new blocking strategies. *Front. Endocrinol. (Lausanne)* 6:12. doi: 10.3389/fendo.2015.00012

Samatov, T. R., Galatenko, V. V., Block, A., Shkurnikov, M. Y., Tonevitsky, A. G., and Schumacher, U. (2017). Novel biomarkers in cancer: the whole is greater than the sum of its parts. *Semin. Cancer Biol.* 45, 50–57. doi: 10.1016/j.semcancer.2016.09.002

Samatov, T. R., Shkurnikov, M. U., Tonevitskaya, S. A., and Tonevitsky, A. G. (2015). Modelling the metastatic cascade by in vitro microfluidic platforms. *Prog. Histochem. Cytochem.* 49, 21–29. doi: 10.1016/j.proghi.2015.01.001

Samatov, T. R., Wicklein, D., and Tonevitsky, A. G. (2016). L1CAM: cell adhesion and more. *Prog. Histochem. Cytochem.* 51, 25–32. doi: 10.1016/j.proghi.2016.05.001

Sawai, H., and Domae, N. (2011). Discrimination between primary necrosis and apoptosis by necrostatin-1 in Annexin V-positive/propidium iodide-negative cells. *Biochem. Biophys. Res. Commun.* 411, 569–573. doi: 10.1016/j.bbrc.2011.06.186

Schröder, C., Schumacher, U., Müller, V., Wirtz, R. M., Streichert, T., Richter, U., et al. (2010). The transcription factor Fra-2 promotes mammary tumour progression by changing the adhesive properties of breast cancer cells. *Eur. J. Cancer* 46, 1650–1660. doi: 10.1016/j.ejca.2010.02.008

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., et al. (2011). Global quantification of mammalian gene expression control. *Nature* 473, 337–342. doi: 10.1038/nature10098

Schwankhaus, N., Gathmann, C., Wicklein, D., Riecken, K., Schumacher, U., and Valentiner, U. (2014). Cell adhesion molecules in metastatic neuroblastoma models. *Clin. Exp. Metastasis* 31, 483–496. doi: 10.1007/s10585-014-9643-8

Shapira, N. (2017). The potential contribution of dietary factors to breast cancer prevention. *Eur. J. Cancer Prev.* 26, 385–395. doi: 10.1097/CEJ.0000000000000406

Sorokin, M., Borisov, N., Kuzmin, D., Gudkov, A., Zolotovskaia, M., Garazha, A., et al. (2021). Algorithmic annotation of functional roles for components of 3,044 human molecular pathways. *Front. Genet.* 12:617059. doi: 10.3389/fgene.2021.617059

Sorokin, M., Ignatev, K., Barbara, V., Vladimirova, U., Muraveva, A., Suntsova, M., et al. (2020). Molecular pathway activation markers are associated with efficacy of trastuzumab therapy in metastatic her2-positive breast cancer better than individual gene expression levels. *Biochemistry* 85, 758–772. doi: 10.1134/S0006297920070044

Stadler, M., Scherzer, M., Walter, S., Holzner, S., Pudelko, K., Riedl, A., et al. (2018). Exclusion from spheroid formation identifies loss of essential cell-cell adhesion molecules in colon cancer cells. *Sci. Rep.* 8:1151. doi: 10.1038/s41598-018-19384-0

Suarez, J. S., Gurler Main, H., Muralidhar, G. G., Elfituri, O., Xu, H.-L., Kajdacsy-Balla, A. A., et al. (2019). CD44 regulates formation of spheroids and controls organ-specific metastatic colonization in epithelial ovarian carcinoma. *Mol. Cancer Res.* 17, 1801–1814. doi: 10.1158/1541-7786.MCR-18-1205

Tallima, H., and El Ridi, R. (2018). Arachidonic acid: physiological roles and potential health benefits – a review. *J. Adv. Res.* 11, 33–41. doi: 10.1016/j.jare.2017.11.004

Toth, M., and Fridman, R. (2001). "Assessment of gelatinases (MMP-2 and MMP-9) by gelatin zymography," in *Metastasis Research Protocols*, eds S. A. Brooks and U. Schumacher (Totowa, NJ: Humana Press), 163–174. doi: 10.1385/1-59259-136-1:163

Tracz, A. F., Szczylik, C., Porta, C., and Czarnecka, A. M. (2016). Insulin-like growth factor-1 signaling in renal cell carcinoma. *BMC Cancer* 16:453. doi: 10.1186/s12885-016-2437-4

Trajkovic-Arsic, M., Kalideris, E., and Siveke, J. T. (2013). The role of insulin and IGF system in pancreatic cancer. *J. Mol. Endocrinol.* 50, R67–R74. doi: 10.1530/JME-12-0259

Tubiana, M. (1989). Tumor cell proliferation kinetics and tumor growth rate. *Acta Oncol. (Madr).* 28, 113–121. doi: 10.3109/02841868909111193

Tyanova, S., Temu, T., and Cox, J. (2016a). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* 11, 2301–2319. doi: 10.1038/nprot.2016.136

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., et al. (2016b). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* 13, 731–740. doi: 10.1038/nmeth.3901

Valentiner, U., Hall, D. M. S., Brooks, S. A., and Schumacher, U. (2005). HPA binding and metastasis formation of human breast cancer cell lines transplanted into severe combined immunodeficient (scid) mice. *Cancer Lett.* 219, 233–242. doi: 10.1016/j.canlet.2004.07.046

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3:research0034. doi: 10.1186/gb-2002-3-7-research0034

Vigneri, P. G., Tirrò, E., Pennisi, M. S., Massimino, M., Stella, S., Romano, C., et al. (2015). The insulin/IGF system in colorectal cancer development and resistance to therapy. *Front. Oncol.* 5:230. doi: 10.3389/fonc.2015.00230

Wang, J., Sharma, A., Ghamande, S. A., Bush, S., Ferris, D., Zhi, W., et al. (2013). Serum protein profile at remission can accurately assess therapeutic outcomes and survival for serous ovarian cancer. *PLoS One* 8:e78393. doi: 10.1371/journal.pone.0078393

Wang, Q., Lv, L., Tang, Y., Zhang, L., and Wang, L. (2018). MMP-1 is overexpressed in triple-negative breast cancer tissues and the knockdown of MMP-1 expression inhibits tumor cell malignant behaviors in vitro. *Oncol. Lett.* 17, 1732–1740. doi: 10.3892/ol.2018.9779

Wang, Y., Torres-Gonzalez, M., Tripathy, S., Botolin, D., Christian, B., and Jump, D. B. (2008). Elevated hepatic fatty acid elongase-5 activity affects multiple pathways controlling hepatic lipid and carbohydrate composition. *J. Lipid Res.* 49, 1538–1552. doi: 10.1194/jlr.M800123-JLR200

Weber, K., Mock, U., Petrowitz, B., Bartsch, U., and Fehse, B. (2010). Lentiviral gene ontology (LeGO) vectors equipped with novel drug-selectable fluorescent proteins: new building blocks for cell marking and multi-gene analysis. *Gene Ther.* 17, 511–520. doi: 10.1038/gt.2009.149

Weber, K., Thomaschewski, M., Benten, D., and Fehse, B. (2012). RGB marking with lentiviral vectors for multicolor clonal cell tracking. *Nat. Protoc.* 7, 839–849. doi: 10.1038/nprot.2012.026

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., 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

Wiktorowska-Owczarek, A., Berezińska, M., and Nowak, J. Z. (2015). PUFAs: structures, metabolism and functions. *Adv. Clin. Exp. Med.* 24, 931–941. doi: 10.17219/acem/31243

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., and Madden, T. L. (2012). Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* 13:134. doi: 10.1186/1471-2105-13-134

Yun, E.-J., Song, K.-S., Shin, S., Kim, S., Heo, J.-Y., Kweon, G.-R., et al. (2014). Docosahexaenoic acid suppresses breast cancer cell metastasis by targeting matrix-metalloproteinases. *Oncotarget* 7, 49961–49971. doi: 10.18632/oncotarget.10266

Zhao, C., Zhu, X., Wang, G., Wang, W., Ju, S., and Wang, X. (2020). Decreased expression of IGFBP6 correlates with poor survival in colorectal cancer patients. *Pathol. Res. Pract.* 216:152909. doi: 10.1016/j.prp.2020.152909

Zheng, C.-C., Hu, H.-F., Hong, P., Zhang, Q.-H., Xu, W. W., He, Q.-Y., et al. (2019). Significance of integrin-linked kinase (ILK) in tumorigenesis and its potential implication as a biomarker and therapeutic target for human cancer. *Am. J. Cancer Res.* 9, 186–197.

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 03 June 2021
doi: 10.3389/fgene.2021.607690

Schizophrenia Plays a Negative Role in the Pathological Development of Myocardial Infarction at Multiple Biological Levels

Xiaorong Yang[1†], Yao Chen[1†], Huiyao Wang[2], Xia Fu[1], Kamil Can Kural[3], Hongbao Cao[3,4] and Ying Li[5]*

[1] Department of Outpatient, West China Hospital, Sichuan University/West China School of Nursing, Sichuan University, Chengdu, China, [2] Mental Health Center of West China Hospital, Sichuan University, Chengdu, China, [3] School of Systems Biology, George Mason University (GMU), Fairfax, VA, United States, [4] Department of Psychiatry, First Hospital/First Clinical Medical College of Shanxi Medical University, Taiyuan, China, [5] The Center of Gerontology and Geriatrics, National Clinical Research Center for Geriatrics, West China Hospital, Sichuan University, Chengdu, China

OPEN ACCESS

Edited by:
Anastasia Anashkina, Engelhardt Institute of Molecular Biology, Russian Academy of Sciences (RAS), Russia

Reviewed by:
Yuqi Zhao, Jackson Laboratory, United States
Elvira Galieva, Novosibirsk State University, Russia
Yundai Chen, Department of Cardiology, Chinese PLA General Hospital, China

***Correspondence:**
Ying Li
yingli@scu.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: *17 September 2020*
Accepted: *06 May 2021*
Published: *03 June 2021*

Citation:
Yang X, Chen Y, Wang H, Fu X, Kural KC, Cao H and Li Y (2021) Schizophrenia Plays a Negative Role in the Pathological Development of Myocardial Infarction at Multiple Biological Levels. Front. Genet. 12:607690. doi: 10.3389/fgene.2021.607690

It has shown that schizophrenia (SCZ) is associated with a higher chance of myocardial infarction (MI) and increased mortality. However, the underlying mechanism is largely unknown. Here, we first constructed a literature-based genetic pathway linking SCZ and MI, and then we tested the expression levels of the genes involved in the pathway by a meta-analysis using nine gene expression datasets of MI. In addition, a literature-based data mining process was conducted to explore the connection between SCZ at different levels: small molecules, complex molecules, and functional classes. The genetic pathway revealed nine genes connecting SCZ and MI. Specifically, SCZ activates two promoters of MI (IL6 and CRP) and deactivates seven inhibitors of MI (ADIPOQ, SOD2, TXN, NGF, ADORA1, NOS1, and CTNNB1), suggesting that no protective role of SCZ in MI was detected. Meta-analysis showed that one promoter of MI (CRP) presented no significant increase, and six out of seven genetic inhibitors of MI demonstrated minor to moderately increased expression. Therefore, the elevation of CRP and inhibition of the six inhibitors of MI by SCZ could be critical pathways to promote MI. Nine other regulators of MI were influenced by SCZ, including two gene families (inflammatory cytokine and IL1 family), five small molecules (lipid peroxide, superoxide, ATP, ascorbic acid, melatonin, arachidonic acid), and two complexes (CaM kinase 2 and IL23). Our results suggested that SCZ promotes the development and progression of MI at different levels, including genes, small molecules, complex molecules, and functional classes.

Keywords: schizophrenia, myocardial infarction, genetic pathway, regression analysis, meta-analysis

INTRODUCTION

Schizophrenia (SCZ) is one of the most chronically disabled mental illnesses (McGrath et al., 2008). The early manifestations of the disease usually appear in middle and late adolescence, and the clinical onset usually begins 2–5 years later. Patients with SCZ pose unique challenges due to affect, cognition, and socio-demographic factors. Myocardial infarction (MI) and afterward heart failure

are the significant causes of death and disability in the developed countries, characterized by acute myocardial ischemia derived from coronary artery occlusion, myocardial injury, and even necrosis (Lu et al., 2015; Sakaguchi et al., 2020).

An increasing amount of literature has discussed the strong correlations between mental disorders and increased MI mortality, especially in patients with SCZ (Nielsen et al., 2015). A study using a nationwide inpatient sample examines the outcomes of Acute Myocardial Infarction (AMI) in patients with SCZ. They found that 4,648 out of 1,196,698 discharged with AMI were also diagnosed SCZ, and these patients diagnosed with both SCZ and AMI showed higher in-hospital mortality (Karthik et al., 2012). Risks of AMI were raised nearly twofold in younger people with SCZ (age under 35) (Wu et al., 2015). A study by Nielsen et al. (2015) reported that 75% of SCZ patients developed silent MI, which may be related to the psychiatric diseases covering up cardiovascular diseases (Kugathasan et al., 2018). Thus, some studies indicate that SCZ is a significant risk factor of in-hospital mortality in MI patients (Sohn et al., 2015; Wu et al., 2015). Also, high mortality following incident MI in individuals with SCZ may associate with low access to care (Kurdyak et al., 2012).

Despite the clinical outcomes that support the relationship between MI and SCZ (Karthik et al., 2012; Nielsen et al., 2015), the underlying mechanisms of the promotion effect of SCZ on MI are largely unknown. It has been suggested that the pathogenesis of a disease can be explained through a multiscale interactome network of proteins, drug targets, and biological functions (Ruiz et al., 2021), and the network-based location of each disease module determines its pathological relationship to other diseases (Menche et al., 2015). Here, we studied the potential influence of SCZ on MI at different levels (genes, small molecules, complex molecules, and functional classes), with functional pathways constructed. Moreover, a meta-analysis was conducted using MI expression data to explore the gene expression variation within the SCZ-driven MI-regulating genetic pathway. Results from this study may add new insights into the understanding of the negative roles that SCZ plays in the pathological development of MI, which is critical in the prevention and treatment of MI in SCZ patients.

MATERIALS AND METHODS

This study is organized as follows. First, we conducted a Natural Language Processing (NLP)-based literature data-mining (Daraselia et al., 2004) to construct a genetic pathway connecting SCZ and MI. Second, we performed a meta-analysis to test the gene expression variations of the pathway genes in MI patients. Lastly, we explored SCZ-driven MI regulators at other levels, including small molecules, functional gene classes, and complex molecules.

Identify SCZ-MI Genetic Pathways

Assisted by Pathway Studio ([1]version 12.3), we conducted an NLP-based large-scale literature data mining to identify common genes that were downstream targets of SCZ and up-regulators of MI. That is, each gene was identified as influenced by SCZ, and was also regulating MI, forming a SCZ→Gene→MI relationship. For each relationship identified, there were at least three independent supporting references, which were provided in the supplementary material SCZ_MI→Ref4GeneticPathway, including the title, DOI/PMID, and the sentences where the relationship was identified. The process was conducted by using MedScan (Daraselia et al., 2004), an NLP-based literature data-mining tool. The data mining covered over 24 million PubMed abstracts and 3.5 million Elsevier and 3rd part full-text papers. Each relationship/edge was built based on the fact extracted from the literature by NLP technology with at least three supporting references. A manually quality control process was enforced to remove unreliable relationships and relationships with non-specific polarities. Here, unreliable relationships refer to these with unmatched sentences, which were false positives by the NLP technique.

All the entities within the remaining network were tested using a meta-analysis with nine independent MI RNA-expression datasets. The purpose of the meta-analysis was to explore the gene expression patterns of these SCZ-drive genes, which may help to understand the literature-based relationships identified. To note, instead of using reported results from original data-related studies, we used the original data to calculate the expression levels. The process is described as follows.

Selection of Gene Expression Datasets for Meta-Analysis

The MI expression datasets were identified within the GEO database[2] (Clough and Barrett, 2016). The search was conducted using the keyword "myocardial infarction" with 12,193 items identified. Among these items, 678 studies with series data were selected. We made an outline of the metadata of the identified datasets and selected a sub-set for the meta-analysis with the following steps and criteria: (1) The dataset was array expression data (296 datasets); (2) The original data and the corresponding format file were downloadable (152 datasets; metadata summary of these datasets were presented in Supplementary data SCZ_MIMI_datasets); (3) The model organism of the study was indicated as "human" or "Homo sapiens" (143 datasets); (4) The study design was MI cases vs. healthy control (9 datasets). For step 4, we manually checked the metadata of the 143 datasets from step 3, and the qualified datasets were included for meta-analysis. The nine datasets that satisfied the above criteria were included in the meta-analysis, as shown in **Table 1**.

Meta-Analysis Models

For each gene, the meta-analysis estimated the effect size in terms of gene expression log2 fold-change (LFC). Results from using both the random-effects model and fixed-effect model were compared following the statistics estimation used by Borenstein et al. (2010). To determine the heterogeneity of the datasets, between- and within-study variance was calculated and

[1]www.pathwaystudio.com

[2]https://www.ncbi.nlm.nih.gov/geo/

TABLE 1 | The nine myocardial infarction expression datasets selected for meta-analysis.

Dataset GEO ID	#Control	#Case	Country	Study Sample Age	Organism
GSE24519	4	34	Italy	3	Homo sapiens
GSE24591	4	34	Italy	3	Homo sapiens
GSE34198	48	49	Czechia Republic	6	Homo sapiens
GSE48060	21	31	United States	6	Homo sapiens
GSE60993	7	10	South Korea	5	Homo sapiens
GSE60993	7	17	South Korea	5	Homo sapiens
GSE62646	14	84	Poland	6	Homo sapiens
GSE66360	50	49	United States	5	Homo sapiens
GSE97320	3	3	China	3	Homo sapiens

compared. When the total variance (Cochran's Q statistic) was no bigger than the expected value of the between-study variances (df), the model sets the ISq (percentage of the within- over between-study variance) to zero. In this case, the fixed-effect model, instead of the random-effects model, will be selected for the meta-analysis. The definition of Cochran's Q statistic, df, and ISq was provided in Eq. (1) to (3) (Borenstein et al., 2010). All analyses were performed using Matlab (R2017a version).

$$Q = \sum_{i=1}^{k} W_i {T_i}^2 - \frac{\left(\sum_{i=1}^{k} W_i T_i\right)^2}{\sum_{i=1}^{k} W_i}, \quad (1)$$

Where T_i is the deviation of each study, W_i is the inverse variance of each study, and k is the total number of studies.

$$df = k - 1, \quad (2)$$

Where k is the total number of studies.

$$ISq = (Q - df)/Q, \quad (3)$$

Where Q is the total variance defined by Equation (1), and df is the degree of freedom defined in Equation (2).

Analysis of Influential Factors

To estimate the possible influence of several factors (e.g., study date, country of origin, and sample size) on the gene expression in MI patients, we conducted a multiple linear regression (MLR) analysis and reported the *P*-values for each of these factors.

Identification of Additional SCZ-Driven MI Regulators

To explore SCZ-driven MI regulators at other levels, we conducted another NLP-based literature data mining assisted by the network building module of Pathway Studio[3]. The analysis was first performed to identify functional gene class, small molecule, complex molecules, and cells that induced MI's pathological development, and then these types of entities regulated by SCZ were also identified. The overlapped entities were used to construct SCZ→MI network. To ensure high confidence in the identified relationship, we used the confidence level of three (identified entities were supported by at least three references) to filter the relationships. We presented the identified entities and the corresponding relationships in SCZ_MI→Ref4SMPathway and SCZ_MI→Ref4CellPathway.

[3]https://supportcontent.elsevier.com/Support%20Hub/Pathway%20Studio/Network%20Builder%20basic%20_Interactive%20NB%20v114.pdf

RESULTS

SCZ-MI Genetic Pathway

As shown in **Figure 1**, there were nine genes driven by SCZ that were MI regulators. Specifically, SCZ activates two MI promoters (IL6 and CRP) and deactivates seven MI inhibitors (ADIPOQ, SOD2, TXN, NGF, ADORA1, NOS1, and CTNNB1). These could be the potential pathways where SCZ plays an essential role in the pathological development and progression of MI. Notably, seven out of the nine genes were MI inhibitors, indicating that SCZ is strongly associated with MI's pathological development more through MI-inhibitors' deactivation than through its promoters' activation. SCZ may play more roles in the progression deterioration than in the initiation of MI. For the details of the pathways presented in **Figure 1**, please refer to SCZ_MI→Ref4GeneticPathway. From this genetic pathway analysis, we identified no "good" effect of SCZ in MI.

Meta-Analysis Results

We conducted a meta-analysis using nine MI-expression datasets to test the expression variation of the genes involved in SCZ-driven genetic pathways for MI. We presented the major results of the meta-analysis and MLR analysis in **Table 2**. The detailed results of the meta-analysis were presented in SCZ_MI→Meta-analysis.

As shown in **Table 2**, population region (Country) was suggested as a significant influential factor for the expression of almost all the genes tested except CRP and CTNNB1. Patients from different countries usually carry racial and ethnic variations that influence gene expression patterns (Hicks et al., 2013). While the sample size only influences the expression of NGF and TXN (p-value = $3e^{-16}$ and $6e^{-6}$, respectively), and study age seems to be an influential factor for IL6 alone (p-value = $9e^{-5}$). These results suggested the complexity of the disease of MI, which could be influenced by multiple factors.

As shown in **Figure 1** and **Table 2**, only one MI promoter (IL6) was significantly up-regulated in MI patients (LFC = 0.71, p-value = 0.021). The other MI promoter (CRP) presented no significant expression change (LFC = 0.018; p-value = 0.40). Therefore, the activation of CRP could be a required course where SCZ promotes MI's pathological development.

We identified two inhibitors of MI, which got moderate elevation in their expression levels, including ADIPOQ and SOD2 (LFC = 0.44 and 0.21, respectively; p-value = 0.10 and 0.24, respectively). Therefore, inhibiting these two genes' activity could be another path that SCZ contributes to the promotion of MI. Most of the other MI inhibitors demonstrated minor elevated expression except CTNNB1 (LFC = −0.069). The increased expressions of all these inhibitors of MI were beneficial in the

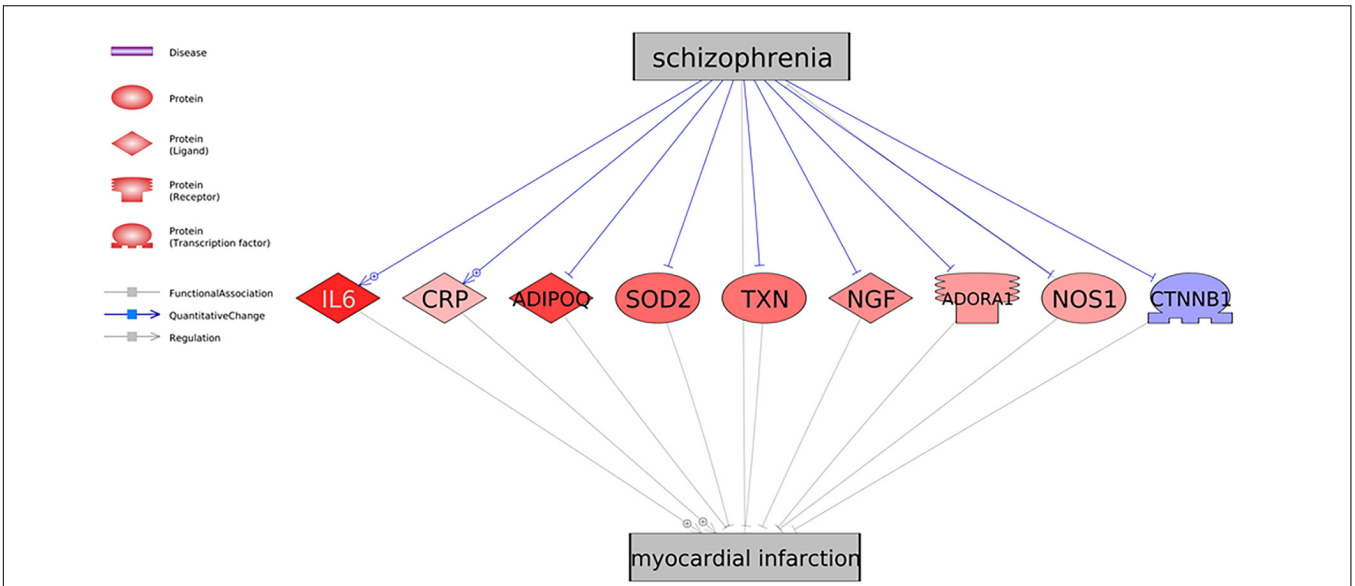


FIGURE 1 | Schizophrenia driven genetic regulators of myocardial infarction. Each relationship was supported by three or more references. The network contains four different types of proteins and three different types of relationships (edges). Genes in red represent an increased gene expression level in myocardial infarction, and blue means decreased.

TABLE 2 | Meta-analysis and Multiple Linear Regression analysis results.

Gene Name	Meta-analysis results				Multiple Linear Regression analysis results (*p*-value)		
	Random-Effects Model (yes = 1; no = 0)	# of Study	Effect size (LFC)	*p*-value	# of Sample	Country	Study Age
ADIPOQ	1	8	0.44	0.10	0.39	*0.013*	0.31
ADORA1	1	9	*0.055*	0.36	0.98	*0.008*	0.06
CRP	0	8	*0.018*	0.40	0.76	0.104	0.53
CTNNB1	0	9	−0.070	0.10	0.88	0.053	0.29
IL6	1	5	0.71	*0.02*	1.00	$3\ e^{-5}$	$9\ e^{-5}$
NGF	1	3	0.086	0.48	$3\ e^{-16}$	$3\ e^{-16}$	1.00
NOS1	0	8	*0.044*	0.20	0.83	*0.008*	0.19
SOD2	1	9	0.21	0.24	0.56	*0.087*	0.84
TXN	1	8	0.17	0.34	$6\ e^{-6}$	$8\ e^{-6}$	1.00

progression MI. Thus, by deactivating these MI inhibitors, SCZ could worsen the progression of MI.

SCZ-Driven Small Molecule Regulators of MI

To explore the connections between SCZ and MI in other levels, we first identified the small molecules that were downstream targets of SCZ and upstream regulators of MI. Three or more references supported each of these relationships (see SCZ_MIRef4SMPathways). Six small molecules satisfied our data mining criteria and formed the small molecule pathway, as shown in **Figure 2**. Among these six small molecules, SCZ activates two out of three MI promoters and deactivates all three MI inhibitors. Although SCZ could also deactivate one MI promoter (arachidonic acid), the overall conclusion from the small molecular pathway (**Figure 2**) is consistent with the genetic pathway (**Figure 1**)—SCZ plays a more negative than positive role at a small molecular level in the pathological development of MI.

SCZ Driven Regulators of MI at Gene Family and Complex Level

Besides small molecules, we also identified two gene families (inflammatory cytokine and IL1 family) and two complexes (CaM kinase 2 and IL23) that were promoters of MI and stimulated by SCZ. To note, the gene family and complex pathways shown in **Figure 3** support only the negative influence of SCZ on MI without a positive effect identified. For the details of the pathways presented in **Figure 3**, please refer to SCZ_MI→Ref4FCPathway.

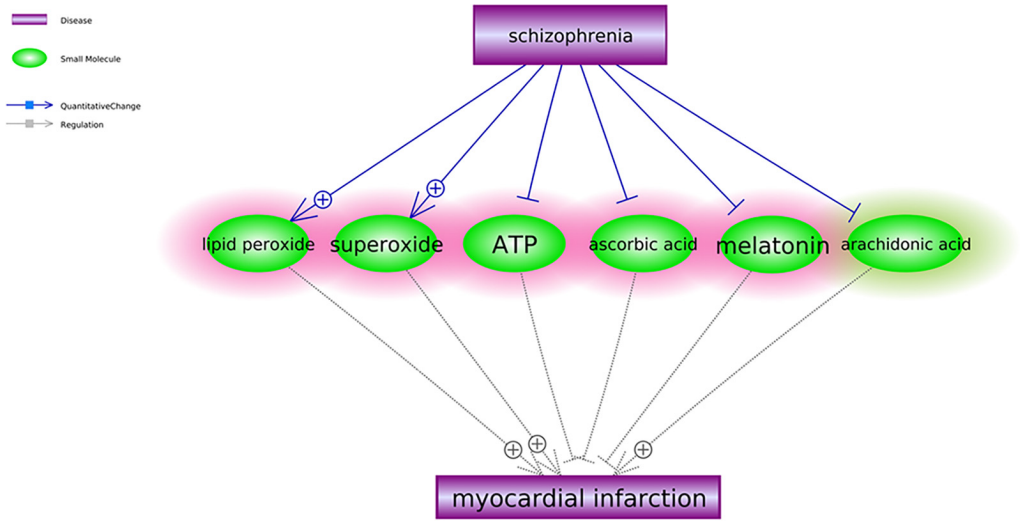


FIGURE 2 | Schizophrenia-driven small molecular regulators of myocardial infarction. Each relationship was supported by three or more references. The network contains two different types of relationships (edges): quantitative change and regulation. The small molecules highlighted in red were the regulators of myocardial infarction driven by schizophrenia to play antagonistic roles in the development and progression of myocardial infarction. The ones in green play antagonistic roles in the development and progression of myocardial infarction.

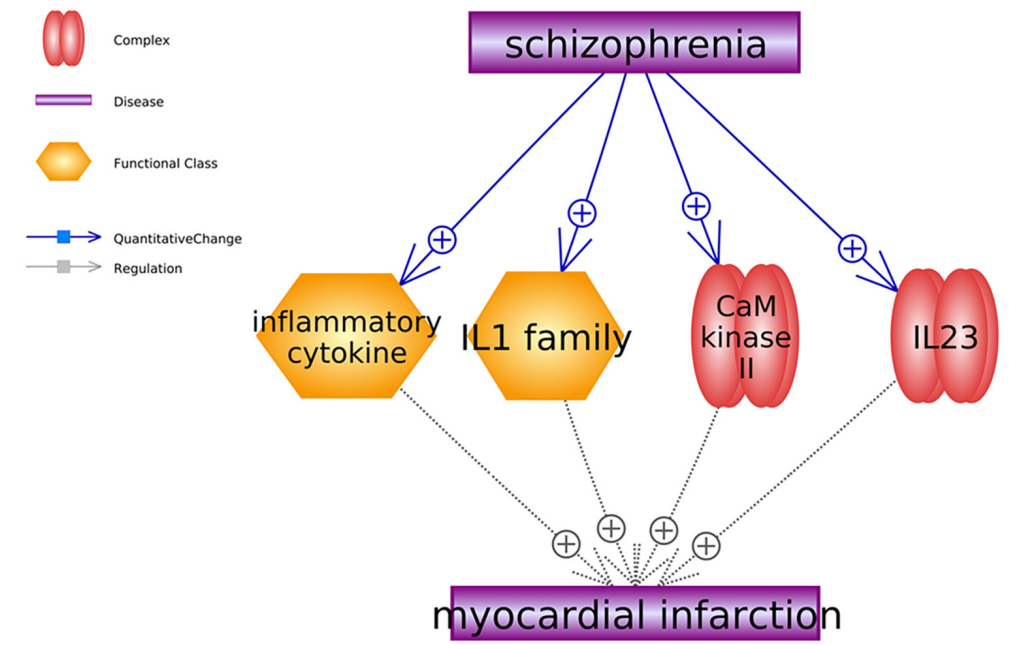


FIGURE 3 | Schizophrenia-driven cells, gene families and complex as regulators of myocardial infarction. Each relationship was supported by three or more references. The network contains two types of relationships (quantitative change and regulation).

DISCUSSION

This study explored the SCZ influenced MI regulators at multiple levels: genetic, gene family, small molecule, and complex. Corresponding pathways were constructed with a meta-analysis to test the gene expression within the genetic pathway in MI patients. The pathway built suggested a negative role of SCZ in MI's pathological development, which is consistent with previous studies (Karthik et al., 2012; Nielsen et al., 2015). However, different from the clinical research exploring the co-occurrence and common clinical features of two diseases, this study mainly focused on experiment data-based studies to uncover potential mechanisms underlying the clinical association between SCZ and MI.

Firstly, we identified the potential association between SCZ and MI at the genetic level through the connection with nine common genes (**Figure 1**). Most of the pathways presented in **Figure 1** pointed to a negative role of SCZ in the pathological development of MI. For example, the interleukin-6 (IL-6) serum concentrations of SCZ patients was confirmed by multiple studies to get significantly elevated (El Kissi et al., 2015), which was shown to be associated with clinical progression of unstable angina and increased risk of MI (Deten et al., 2003; Gori et al., 2006). Our meta-analysis result confirmed that elevated IL-6 expressions in patients with MI (LFC = 0.71, p-value = 0.021). Therefore, the SCZ-IL6-MI could be one of the pathological paths where SCZ promotes MI.

The meta-analysis also showed that the expression of CRP, a promoter of MI, was not significantly elevated in the nine MI-datasets employed in the meta-analysis (**Table 2**). However, in chronic SCZ patients, the expression levels of CRP could be significantly increased (Meyer et al., 2009), which has been shown to play an essential role in the development of heart failure after MI (Al Aseri et al., 2019). Therefore, the SCZ-CRP-MI pathway could be an essential mechanism explaining the promotion role of SCZ in the progression of MI. Besides the effect of SCZ on the MI promoters, most of the MI inhibitors demonstrated increased expression in MI (LFC $>$ 0), which indicates that SCZ may play more roles in the progression deterioration than in the initiation of MI.

However, we noted that the early stage of SCZ could play some protective role in MI progression through the up-regulation of TXN (an MI inhibitor) (**Figure 1**). The overexpression of TXN has been suggested as a therapeutical target for MI (Sag et al., 2014; Yang et al., 2016). In the early stage of SCZ, TXN expression could be elevated (Zhang et al., 2009), while in chronic SCZ patients, TXN was shown to be down-regulated, which inverses the role of SCZ back to negative in the progression of MI (Aydın et al., 2018).

ADORA1 forms an oligomeric structure with P2RY1 (Yoshioka et al., 2001) to mediate purine signaling. This activation triggers two different Ca^{+2} release pathways through Calmodulin Kinase 2 (CamKII) and inositol triphosphate (Paredes-Gamero et al., 2006). The release of calcium ions is essential for heart muscle contractions and electrical signal formation. Therefore, an expression increase in ADORA1 might disrupt the intensity of electrical signals generated by heart and muscle tissues. Increased IL-6 expression could be tied to increased release of calcium ions and inositol triphosphate, which causes a positive feedback loop (Bustamante et al., 2014).

Perhaps the most important finding is related to the formation of adherens and gap junction interactions after expression changes of the identified molecules. Deactivation of the beta-catenin transactivating complex is crucial for structural changes in heart muscle formation and maintenance. ADIPOQ, NOS1, TXN, CTNNB1, and CRP play an essential role in the beta-catenin transactivating complex's deactivation. Beta-catenin is localized at the fascia adherens junction, where it is part of the N-cadherin– actin complex. Over and underexpression of CTNNB1 (beta-catenin) is tied to cardiomyopathies due to structural changes in heart muscle (Sheikh et al., 2009). Beta-catenin is crucial in cell differentiation in the brain as well, and an abnormal Wnt gene expression and plasma protein levels are proven to be related to SCZ in the earlier studies (Hoseth et al., 2018).

We also identified six SCZ-driven small molecules that influence MI's advance, as shown in **Figure 2**. Different from the genetic pathway, we also identified one potential "good" pathway (SCZ→arachidonic acid→MI) where SCZ plays a protective role in MI development. It has been shown that arachidonic acid levels are reduced in post mortem and peripheral red blood cell membranes in SCZ (Berger et al., 2016), while the 5-lipoxygenase derivatives of arachidonic acid have been shown to play an important pathogenic role during MI (Lisovyy et al., 2009). However, the regulation of other MI inhibitors and promoters (**Figure 2**) suggested that SCZ plays a more negative than positive role at the small molecular (compound) level in the pathological development and progression of MI, which is consistent with that of the genetic pathway presented in **Figure 1**. For instance, SCZ has been shown to reduce the secretion of melatonin, which was implicated as a protector for the cardiac microvascular ischemia to improve the therapeutic outcomes of MI (Zhou et al., 2018; Saberi et al., 2019). More details of the pathway presented in **Figure 2** can be found in SCZ_MI→Ref4SMPathway.

Moreover, our study also uncovered four MI promoters (**Figure 3**), including two SCZ-driven gene families and two complexes. Notably, plasma concentrations of the Interleukin-1 family (IL-1 family) were found significantly increased in SCZ patients (Sirota et al., 2003), which may characteristically modify the process of coronary artery disease associated with Chlamydia pneumonia infection, leading to the development of MI (Momiyama et al., 2001). The expression of calmodulin kinase II (CaM kinase II) was also elevated in the tissues of patients with SCZ (Lee et al., 2010). It has been suggested that CaM kinase II inhibition could improve ventricular functions and restores normal Ca2+ homeostasis after MI (Hund et al., 2008; Fu et al., 2013), while the overexpression of CaM kinase II causes dilated cardiomyopathy and ventricular dysfunction associated with abnormal Ca2+ handling (Hund et al., 2008). Therefore, the pathways presented in **Figure 3** may add new insights into the understanding of the negative role of SCZ in MI development.

This study has several limitations that need to be addressed in future work. First, the identification of the SCZ driven MI-regulators was filtered to have support by at least three references. While this decreased the identified entities' false-positive ratio, some vital information might be lost between SCZ and MI connection, which needs further consideration. Second, the pathways and relations were constructed based on previous studies that were conducted in different backgrounds. Biology experiments are needed to validate any of the relationships identified in this study. Third, the sample size of the MI datasets employed in this study presented significant variance, which influenced the meta-analysis results. Forth, the relationships identified in this study were mainly quantitative changes at the gene/protein expression level. Other types of relation (e.g., genetic change by GWAS study) may add new insights into the understanding of the SCZ-MI relationship. Fifth, due to the

limitation of the NLP technique employed in this study, we did not separate different study types (e.g., human, animal, or cell line) when building the pathway given in **Figures 1–3**.

CONCLUSION

We identified 19 SCZ-driven MI-regulators at different biological levels, and SCZ exerts an overall negative influence on MI through the regulation of most of them (18 out of 19). Our results indicated the complexity of the connection between SCZ and MI and may add new insights into the understanding of the negative role that SCZ plays in the pathology of MI.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

XY, YC, and YL developed the study design, analyzed the data, and wrote the original manuscript. All authors read and approved the final manuscript.

FUNDING

The study was supported Key R&D Project, Department of Science and Technology of Sichuan Province (20ZDYF2800 and 2020YFG0086), National Natural Science Foundation of China (81501197), Research Project on Healthcare in Sichuan Province (Chuanganyan ZH2019-101), and Special Fund Project for Science and Technology Cooperation of Sichuan University and Zigong City, Sichuan Province (2019CDZG-25).

REFERENCES

Al Aseri, Z. A., Habib, S. S., and Marzouk, A. (2019). Predictive value of high sensitivity C-reactive protein on progression to heart failure occurring after the first myocardial infarction. *Vasc. Health Risk Manag.* 15, 221–227. doi: 10.2147/VHRM.S198452

Aydın, E. P., Genç, A., Dalkıran, M., Uyar, E. T., Deniz, I., Özer, ÖA., et al. (2018). Thioredoxin is not a marker for treatment-resistance depression but associated with cognitive function: an rTMS study. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 80, 322–328. doi: 10.1016/j.pnpbp.2017.04.025

Berger, G. E., Smesny, S., Schäfer, M. R., Milleit, B., Langbein, K., Hipler, U. C., et al. (2016). Niacin skin sensitivity is increased in adolescents at ultra-high risk for psychosis. *PLoS One* 11:e0148429. doi: 10.1371/journal.pone.0148429

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res. Synth. Methods* 1, 97–111. doi: 10.1002/jrsm.12

Bustamante, M., Fernández-Verdejo, R., Jaimovich, E., and Buvinic, S. (2014). Electrical stimulation induces IL-6 in skeletal muscle through extracellular ATP by activating Ca(2+) signals and an IL-6 autocrine loop. *Am. J. Physiol. Endocrinol. Metab.* 306, E869–E882. doi: 10.1152/ajpendo.00450.2013

Clough, E., and Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.* 1418, 93–110. doi: 10.1007/978-1-4939-3578-9_5

Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics* 20, 604–611. doi: 10.1093/bioinformatics/btg452

Deten, A., Volz, H. C., Holzl, A., Briest, W., and Zimmer, H. G. (2003). Effect of propranolol on cardiac cytokine expression after myocardial infarction in rats. *Mol. Cell Biochem.* 251, 127–137.

El Kissi, Y., Samoud, S., Mtiraoui, A., Letaief, L., Hannachi, N., Ayachi, M., et al. (2015). Increased interleukin-17 and decreased BAFF serum levels in drug-free acute schizophrenia. *Psychiatry Res.* 225, 58–63. doi: 10.1016/j.psychres.2014.10.007

Fu, Q., Chen, X., and Xiang, Y. K. (2013). Compartmentalization of β-adrenergic signals in cardiomyocytes. *Trends Cardiovasc. Med.* 23, 250–256. doi: 10.1016/j.tcm.2013.02.001

Gori, A. M., Sofi, F., Corsi, A. M., Gazzini, A., Sestini, I., Lauretani, F., et al. (2006). Predictors of vitamin B6 and folate concentrations in older persons: the InCHIANTI study. *Clin. Chem.* 52, 1318–1324. doi: 10.1373/clinchem.2005.066217

Hicks, C., Miele, L., Koganti, T., Young-Gaylor, L., Rogers, D., Vijayakumar, V., et al. (2013). Analysis of patterns of gene expression variation within and between ethnic populations in pediatric B-ALL. *Cancer Inform.* 12, 155–173. doi: 10.4137/CIN.S11831

Hoseth, E. Z., Krull, F., Dieset, I., Mørch, R. H., Hope, S., Gardsjord, E. S., et al. (2018). Exploring the Wnt signaling pathway in schizophrenia and bipolar disorder. *Transl. Psychiatry* 8:55. doi: 10.1038/s41398-018-0102-1

Hund, T. J., Decker, K. F., Kanter, E., Mohler, P. J., Boyden, P. A., Schuessler, R. B., et al. (2008). Role of activated CaMKII in abnormal calcium homeostasis and I(Na) remodeling after myocardial infarction: insights from mathematical modeling. *J. Mol. Cell Cardiol.* 45, 420–428. doi: 10.1016/j.yjmcc.2008.06.007

Karthik, M., Gagan, K., Abhishek, D., Rajesh, S., and Jawahar, M. (2012). Schizophrenia and use of revascularization procedures after acute myocardial infarction. *J. Am. Coll. Cardiol.* 59(Suppl.):E1898. doi: 10.1016/S0735-1097(12)61899-3

Kugathasan, P., Laursen, T. M., Grøntved, S., Jensen, S. E., Aagaard, J., and Nielsen, R. E. (2018). Increased long-term mortality after myocardial infarction in patients with schizophrenia. *Schizophr. Res.* 199, 103–108. doi: 10.1016/j.schres.2018.03.015

Kurdyak, P., Vigod, S., Calzavara, A., and Wodchis, W. P. (2012). High mortality and low access to care following incident acute myocardial infarction in individuals with schizophrenia. *Schizophr. Res.* 142, 52–57. doi: 10.1016/j.schres.2012.09.003

Lee, J. G., Cho, H. Y., Park, S. W., Seo, M. K., and Kim, Y. H. (2010). Effects of olanzapine on brain-derived neurotrophic factor gene promoter activity in SH-SY5Y neuroblastoma cells. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 34, 1001–1006. doi: 10.1016/j.pnpbp.2010.05.013

Lisovyy, O. O., Dosenko, V. E., Nagibin, V. S., Tumanovska, L. V., Korol, M. O., Surova, O. V., et al. (2009). Cardioprotective effect of 5-lipoxygenase gene (ALOX5) silencing in ischemia-reperfusion. *Acta Biochim. Pol.* 56, 687–694.

Lu, L., Liu, M., Sun, R., Zheng, Y., and Zhang, P. (2015). Myocardial infarction: symptoms and treatments. *Cell Biochem. Biophys.* 72, 865–867. doi: 10.1007/s12013-015-0553-4

McGrath, J., Saha, S., Chant, D., and Welham, J. (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiol. Rev.* 30, 67–76. doi: 10.1093/epirev/mxn001

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Meyer, J. M., McEvoy, J. P., Davis, V. G., Goff, D. C., Nasrallah, H. A., Davis, S. M., et al. (2009). Inflammatory markers in schizophrenia: comparing antipsychotic effects in phase 1 of the clinical antipsychotic trials of intervention effectiveness study. *Biol. Psychiatry.* 66, 1013–1022. doi: 10.1016/j.biopsych.2009.06.005

Momiyama, Y., Hirano, R., Taniguchi, H., Nakamura, H., and Ohsuzu, F. (2001). Effects of interleukin-1 gene polymorphisms on the development of coronary artery disease associated with *Chlamydia pneumoniae* infection. *J. Am. Coll. Cardiol.* 38, 712–717. doi: 10.1016/s0735-1097(01)01438-3

Nielsen, J., Juel, J., Alzuhairi, K. S., Al Zuhairi, K. S. M., Friis, R., Graff, C., et al. (2015). Unrecognised myocardial infarction in patients with schizophrenia. *Acta Neuropsychiatr.* 27, 106–112. doi: 10.1017/neu.2014.41

Paredes-Gamero, E. J., Craveiro, R. B., Pesquero, J. B., França, J. P., Oshiro, M. E., and Ferreira, A. T. (2006). Activation of P2Y1 receptor triggers two calcium signaling pathways in bone marrow erythroblasts. *Eur. J. Pharmacol.* 534, 30–38. doi: 10.1016/j.ejphar.2006.01.010

Ruiz, C., Zitnik, M., and Leskovec, J. (2021). Identification of disease treatment mechanisms through the multiscale interactome. *Nat. Commun.* 12:1796. doi: 10.1038/s41467-021-21770-8

Saberi, K., Pasbakhsh, P., Omidi, A., Borhani-Haghighi, M., Nekoonam, S., Omidi, N., et al. (2019). Melatonin preconditioning of bone marrow-derived mesenchymal stem cells promotes their engraftment and improves renal regeneration in a rat model of chronic kidney disease. *J. Mol. Hist.* 50, 129–140. doi: 10.1007/s10735-019-09812-4

Sag, C. M., Santos, C. X., and Shah, A. M. (2014). Redox regulation of cardiac hypertrophy. *J. Mol. Cell Cardiol.* 73, 103–111. doi: 10.1016/j.yjmcc.2014.02.002

Sakaguchi, A., Nishiyama, C., and Kimura, W. (2020). Cardiac regeneration as an environmental adaptation. *Biochim. Biophys. Acta Mol. Cell Res.* 1867:118623. doi: 10.1016/j.bbamcr.2019.118623

Sheikh, F., Ross, R. S., and Chen, J. (2009). Cell-cell connection to cardiac disease. *Trends Cardiovasc. Med.* 19, 182–190. doi: 10.1016/j.tcm.2009.12.001

Sirota, P., Gavrieli, R., and Wolach, B. (2003). Overproduction of neutrophil radical oxygen species correlates with negative symptoms in schizophrenic patients: parallel studies on neutrophil chemotaxis, superoxide production and bactericidal activity. *Psychiatry Res.* 121, 123–132. doi: 10.1016/s0165-1781(03)00222-1

Sohn, M., Moga, D. C., and Talbert, J. (2015). Mental disorder comorbidity and in-hospital mortality among patients with acute myocardial infarction. *Geriatric. Ment. Health Care* 3, 7–11. doi: 10.1016/j.gmhc.2015.04.002

Wu, S. I., Chen, S. C., Liu, S. I., Sun, F. J., Juang, J. J., Lee, H. C., et al. (2015). Relative risk of acute myocardial infarction in people with schizophrenia and bipolar disorder: a population-based cohort study. *PLoS One* 10:e0134763. doi: 10.1371/journal.pone.0134763

Yang, C. J., Yang, J., Yang, J., and Fan, Z. X. (2016). Thioredoxin-1 (Trx1) engineered mesenchymal stem cell therapy is a promising feasible therapeutic approach for myocardial infarction. *Int. J. Cardiol.* 206, 169–170. doi: 10.1016/j.ijcard.2015.10.150

Yoshioka, K., Saitoh, O., and Nakata, H. (2001). Heteromeric association creates a P2Y-like adenosine receptor. *Proc. Natl. Acad. Sci. U.S.A.* 98, 7617–7622. doi: 10.1073/pnas.121587098

Zhang, X. Y., Chen, D. C., Xiu, M. H., Wang, F., Qi, L. Y., Sun, H. Q., et al. (2009). The novel oxidative stress marker thioredoxin is increased in first-episode schizophrenic patients. *Schizophr. Res.* 113, 151–157. doi: 10.1016/j.schres.2009.05.016

Zhou, H., Li, D., Zhu, P., Ma, Q., Toan, S., Wang, J., et al. (2018). Inhibitory effect of melatonin on necroptosis via repressing the Ripk3-PGAM5-CypD-mPTP pathway attenuates cardiac microvascular ischemia-reperfusion injury. *J. Pineal. Res.* 65:e12503. doi: 10.1111/jpi.12503

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 June 2021
doi: 10.3389/fgene.2021.607248

Computational Identification of miRNAs and Temperature-Responsive lncRNAs From Mango (*Mangifera indica* L.)

Nann Miky Moh Moh[1,2], *Peijing Zhang*[2†], *Yujie Chen*[2,3] *and Ming Chen*[2,3*]

[1] *Biotechnology Research Department, Ministry of Education, Kyaukse, Myanmar,* [2] *State Key Laboratory of Plant Physiology and Biochemistry, Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, China,* [3] *College of Life Sciences and Food, Inner Mongolia University for the Nationalities, Tongliao, China*

OPEN ACCESS

Edited by:
Tatiana V. Tatarinova, University of La Verne, United States

Reviewed by:
Yuriy L. Orlov, I.M. Sechenov First Moscow State Medical University, Russia
Evgenii Chekalin, Michigan State University, United States

***Correspondence:**
Ming Chen
mchen@zju.edu.cn

†Present address:
Peijing Zhang, Liangzhu Laboratory, Zhejiang University Medical Centre, Hangzhou, China

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

Received: *16 September 2020*
Accepted: *26 April 2021*
Published: *07 June 2021*

Citation:
Moh NMM, Zhang P, Chen Y and Chen M (2021) Computational Identification of miRNAs and Temperature-Responsive lncRNAs From Mango (Mangifera indica L.). Front. Genet. 12:607248. doi: 10.3389/fgene.2021.607248

Mango is a major tropical fruit in the world and is known as the king of fruits because of its flavor, aroma, taste, and nutritional values. Although various regulatory roles of microRNAs (miRNAs) and long non-coding RNAs (lncRNAs) have been investigated in many plants, there is yet an absence of such study in mango. This is the first study to provide information on non-coding RNAs (ncRNAs) of mango with the aims of identifying miRNAs and lncRNAs and discovering their potential functions by interaction prediction of the miRNAs, lncRNAs, and their target genes. In this analysis, about a hundred miRNAs and over 7,000 temperature-responsive lncRNAs were identified and the target genes of these ncRNAs were characterized. According to these results, the newly identified mango ncRNAs, like other plant ncRNAs, have a potential role in biological and metabolic pathways including plant growth and developmental process, pathogen defense mechanism, and stress-responsive process. Moreover, mango lncRNAs can target miRNAs to reduce the stability of lncRNAs and can function as molecular decoys or sponges of miRNAs. This paper would provide information about miRNAs and lncRNAs of mango and would help for further investigation of the specific functions of mango ncRNAs through wet lab experiments.

Keywords: mango, *Mangifera indica*, miRNA, lncRNA, stress response, target genes, interaction network

Abbreviations: ABA, abscisic acid; ACO1, 1-aminocyclopropane-1-carboxylate oxidase 1; ADH1, alcohol dehydrogenase 1; AMFE, adjusted minimal free energy; BLAST, basic local alignment search tool; bp, base pair; C, cytosine; CNGC, cyclic nucleotide-gated channel; CPC, coding potential calculator; CRlncRNAs, cold-responsive lncRNAs; DCL1, dicer-like 1 enzyme; EST, expressed sequence tag; FDR, false discovery rate; G, guanine; GO, gene ontology; GSS, genome survey sequence; HRlncRNAs, heat-responsive lncRNAs; Hsp70, heat shock protein 70; KAT2, 3-ketoacyl-CoA thiolase 2; kcal, kitocalorie; KEGG, Kyoto Encyclopedia of Genes and Genomes; lncRNA, long non-coding RNA; Lrr, leucine-rich repeat receptor; MFE, minimal free energies; MFEI, minimal free energy index; miRNA, microRNA; mRNA, messenger RNA; NCBI, National Center for Biotechnology Information; ncRNA, non-coding RNA; ndG, normalized binding free energy; Nr, non-redundant; nt, nucleotide; ORF, open reading frame; piRNA, piwi-interacting RNA; RC12B, related cDNA 12 B; RISC, RNA-induced silencing complex; RNA, ribonucleic acid; SCL14, scarecrow-like 14; siRNA, small interfering; snoRNA, small nucleolar RNA; SPL, Squamosa promoter binding protein-like; STP13, sugar transport protein 13; UPE, maximum energy of unpairing.

INTRODUCTION

Non-coding RNAs (ncRNAs) are RNA molecules that have no or little protein-coding potential and are not translated into proteins although they are transcribed from DNA. These ncRNAs can be classified according to their length. Small ncRNAs including microRNA (miRNA), small interfering RNA (siRNA), small nucleolar RNA (snoRNA), and piwi-interacting RNA (piRNA) are shorter than 200 nucleotides (nt) in length, and long non-coding RNA (lncRNAs) are longer than 200 nt (Blignaut, 2012).

The miRNAs are small (18–24 nt), endogenous, and regulatory RNA molecules derived from their long self-complementary precursor sequences which can fold into hairpin secondary structures (Ambros et al., 2003). In plants, these long primary precursor miRNAs are transcribed by RNA polymerase II or RNA polymerase III and then processed by dicer-like 1 enzyme (DCL1) into miRNA/miRNA* duplex which is the mature miRNA sequence and its opposite miRNA strand (miRNA*) (Kurihara and Watanabe, 2004; Panda et al., 2014). Finally, the mature miRNAs are incorporated into an RNA-induced silencing complex (RISC) (Bartel, 2004). The binding of miRNAs to their targeted mRNAs in a perfect or nearly perfect complementarity suggests a method for identifying their targets by BLAST analysis (Zhang et al., 2006a) or other related publicly available tool like psRNATarget[1] (Dai and Zhao, 2011). Many experimental researches have proved that miRNAs are involved in many important biological and metabolic processes. In plants, miRNAs play a fundamental role in almost all biological and metabolic processes including plant growth, development, signal transduction, and various stress responses by binding to their target genes (Rhoades et al., 2002).

Long non-coding RNAs are a family of regulatory RNAs having a minimal length of 200 nt. Most lncRNAs are transcribed by RNA polymerase II although some are transcribed by RNA polymerase III (Dieci et al., 2007; Geisler and Coller, 2013; Zhang and Chen, 2013). LncRNAs can interact with ncRNAs such as miRNAs (Jalali et al., 2013). LncRNAs not only can target miRNAs to reduce the stability of lncRNAs but also can function as molecular decoys or sponges of miRNAs (Salmena et al., 2011). Moreover, lncRNAs can compete with miRNAs to bind to their target mRNAs and are the precursors for the generation of miRNAs to silence target mRNAs (Yoon et al., 2014). Many lines of evidence showed that plant lncRNAs play an important role in fundamental biological processes including growth and development and abiotic stress responses (Xin et al., 2011). However, the molecular basis of how lncRNAs function and mediate gene regulation is still poorly understood (Megha et al., 2015).

The genus *Mangifera* belongs to the family Anacardiaceae and contains about 69 different species. *Mangifera indica* L. (mango) is the most common species among them (Mukherjee, 1972; Slippers et al., 2005). Mango is one of the main tropical fruits over the world and is believed to have originated from Asia (Hirano et al., 2010). The well-known countries for mango cultivation are China, India, Thailand, Pakistan, Mexico, Philippines, and Myanmar. The annual production of mango is approximately 42 million tons which is second after banana production (Galán Saúco, 2010). Mango is called as the king of fruits because of its special characteristic flavor, pleasant aroma, taste, and nutritional values. Both ripe and raw fruits can be used as food products such as pickles, juice, jam, powder, sauce, cereal flakes, and so on (Siddiq et al., 2012). Moreover, various parts of mango trees have been used for medical purposes a long time ago, mostly in Southeast Asian and African countries (Mukherjee, 1953). *In vitro* and *in vivo* studies have indicated the various pharmacological potentials of *M. indica* such as anticancer, anti-inflammatory, antidiabetic, antioxidant, antifungal, antibacterial, anthelmintic, gastroprotective, hepatoprotective, immunomodulatory, antiplasmodial, and antihyperlipidemic effects (Lauricella et al., 2017). As a tropical plant, mango is susceptible to cold temperature (Sudheeran et al., 2018), and its floral morphogenesis, photosynthesis, and stomatal limitation are induced by chilling temperature (Núñez-Elisea and Davenport, 1994; Allen et al., 2000). Heat treatment can also affect mango genes involved in stress response and pathogen defense mechanism, genes involved in chlorophyll degradation and photosynthesis, and genes involved in sugar and flavonoid metabolism (Luria et al., 2014).

Although mango is a popular plant with many important usages, its ncRNA data are still limited. Over 10,000 miRNA data of several plants can be accessed in the miRNA database, miRBase[2], but mango miRNAs and their functions have not yet been identified. The regulatory roles of lncRNAs and the molecular basis of lncRNA-mediated gene regulation are also still poorly understood in plants including mango. So, the aims of this research work are to identify and study about the miRNAs and lncRNAs of mango and to examine their potential functions by the interaction prediction of the miRNAs, lncRNAs, and their target genes.

MATERIALS AND METHODS

Data Collection

A total of 10,415 plant miRNAs (release 21) were downloaded from the miRBase database (see text footnote 2), and the redundancy sequences were removed. The resulting 6,042 non-redundant known miRNAs were used as the reference for the prediction of conserved miRNAs.

For the identification of mango miRNAs, 107,744 mango unigenes collected from the mango RNA-Seq database[3] were used (Tafolla-Arellano et al., 2017). These unigenes were derived from the peels of Keitt mango cultivar.

A total of 277,071 RNA transcripts from Zill (Wu et al., 2014), Shelly (Luria et al., 2014; Sivankalyani et al., 2016), and Keitt (Tafolla-Arellano et al., 2017) mango cultivars were used for the identification of lncRNAs.

[1]http://plantgrn.noble.org/psRNATarget/

[2]http://www.mirbase.org/cgi-bin/browse.pl

[3]http://bioinfo.bti.cornell.edu/cgi-bin/mango/index.cgi

Identification of miRNAs and Their Precursors

First, the homology search of mango unigenes against non-redundant plant miRNAs was performed by using BLASTn (BLAST+ 2.7.1) with an *e*-value cutoff of 10. The following criteria were used to choose the candidate miRNAs: the length of the candidate miRNA should be greater than or equal to 18 nt without gap and the number of mismatches between mango sequences and plant miRNAs should not be more than 2. The sequences of 100 nt upstream and 100 nt downstream from the BLAST hit were extracted for precursor sequences. If the length of the query sequence was less than 200 nt, the entire sequence was selected. BLASTx against NCBI non-redundant (nr) protein databases was used to remove the protein-coding sequences from the extracted precursor sequences with an *e*-value cutoff of 0.01. The secondary structures of the remaining precursor sequences were predicted by using the Zuker folding algorithm in MFOLD (version 2.3) software (Zuker, 2003) with default parameters. The workflow for the identification of miRNAs is briefly described in **Figure 1**. Based on the parameters suggested by Dr. Zhang in the identification and characterization of new plant miRNAs using EST analysis (Zhang et al., 2005), the potential pre-miRNAs were predicted as follows:

1. The minimum length of precursor should be at least 60 nt;
2. The pre-miRNA sequence should be folded into an appropriate stem-loop hairpin secondary structure;
3. It should contain the mature miRNA within one arm of the hairpin;

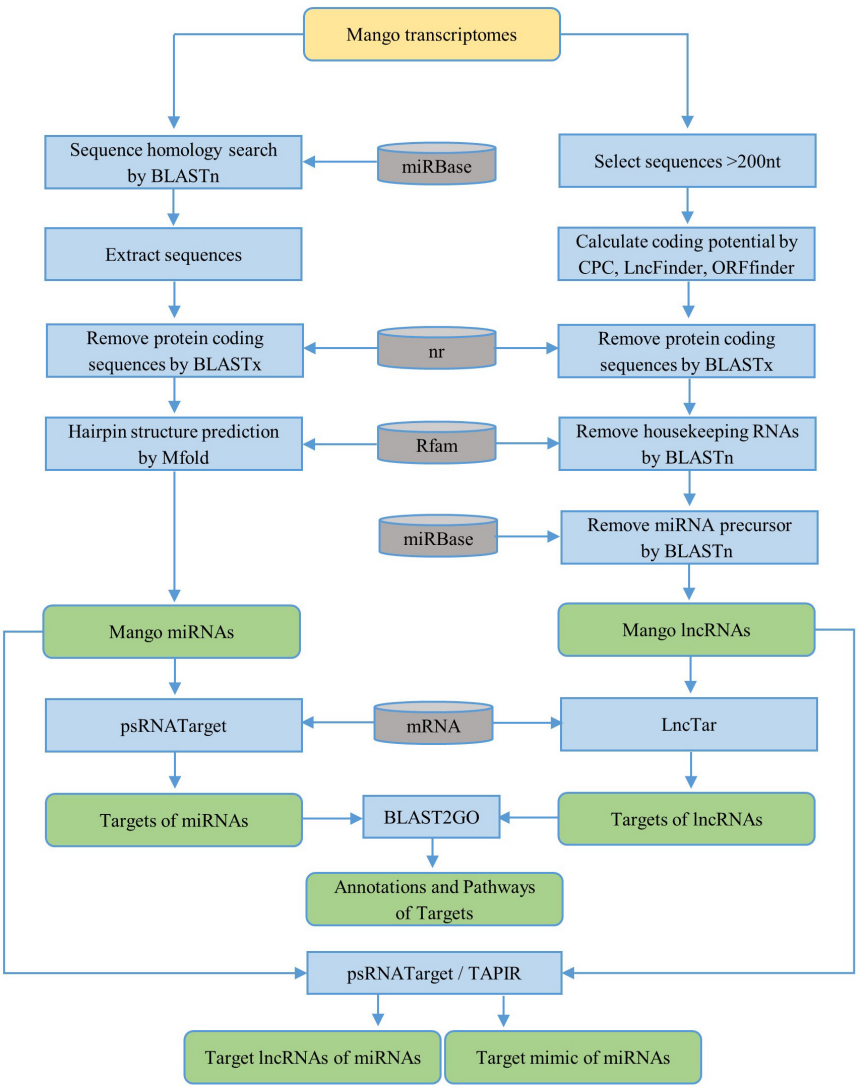


FIGURE 1 | Workflow for the identification of ncRNAs [microRNAs (miRNAs) and long non-coding RNAs (lncRNAs)] and their targets. The yellow rounded rectangle represents the data input and the green rounded rectangles represent the data output. The blue rectangles are the processing steps and the gray cans represent databases.

4. The predicted mature miRNAs and its opposite miRNA* sequence in the other arm of the hairpin should not have more than 6 nt mismatches;
5. Loop or break should not be contained between the miRNA/miRNA* duplex;
6. Maximum size of a bulge in the mature miRNA sequences should not be more than 3 nt;
7. The predicted secondary structures should have higher minimal free energies (MFE) and minimal free energy index (MFEI);
8. MFEI of pre-miRNA should be greater than 0.7;
9. A+U content should be within 30–70%.

The following equations were used to calculate the MFEI and adjusted minimal free energy (AMFE):

$$\mathrm{AMFE} = \left(\frac{\mathrm{MFE}}{\text{length of pre} - \text{miRNA}}\right) \times 100$$

$$\mathrm{MFEI} = \frac{\mathrm{AMFE}}{(\mathrm{G}+\mathrm{C})\%}$$

Prediction of Candidate lncRNAs

To predict the lncRNAs, the transcripts smaller than 200 nt were firstly removed. The coding potential of the remaining transcripts was then calculated by CPC (coding potential calculator) (Kong et al., 2007) and LncFinder (Han et al., 2019). Only sequences with a CPC score less than -1 and LncFinder score less than 0.5 were used for further prediction. The protein coding sequences were removed by BLASTx against the NCBI nr protein databases. Also, the ORF Finder tool[4] was used to predict the open reading frame (ORF) of the remaining sequences, and the minimal ORF cutoff less than 102 amino acids was applied for the prediction. Then, housekeeping genes were removed against the Rfam_14.0 database[5] with *e*-value 0.001. Finally, to remove the lncRNAs acting as precursors of known or novel miRNAs, lncRNAs were aligned with precursors of known non-redundant plant miRNAs from the miRBase database[6] using BLASTn with default parameters (**Figure 1**).

The remaining transcriptome sequences that were not captured as lncRNAs were used as queries against the NCBI nr protein database using BLASTx with a cutoff *e*-value of 1e-5. The sequences with BLAST hits were then analyzed to remove the housekeeping RNAs. The final sequences were identified as protein coding sequences in this study for target gene analysis.

Identification of Significantly Expressed Temperature-Responsive lncRNAs

Mango is a tropical plant and is susceptible to cold temperature (Sudheeran et al., 2018), and its floral morphogenesis, photosynthesis, and stomatal limitation are induced by chilling temperature (Núñez-Elisea and Davenport, 1994; Allen et al., 2000). A high temperature can also affect mango genes involved in stress response and pathogen defense mechanism, genes involved in chlorophyll degradation and photosynthesis, and genes involved in sugar and flavonoid metabolism (Luria et al., 2014). Therefore, temperature-responsive lncRNAs were identified. From the resulting lncRNA transcripts, the temperature-responsive lncRNAs were filtered by two parameters. The mango lncRNAs with an adjusted *p*-value of 0.05 and a log2 fold change of greater than 2 or less than -2 were identified as the significantly expressed lncRNAs.

Target Gene Prediction of miRNAs and lncRNAs

Mango mRNAs downloaded from the NCBI database and mango protein coding sequences previously identified were used for the target gene prediction of miRNAs. The putative target sites of miRNAs were identified by aligning miRNA sequences using the plant target prediction tool, psRNATarget (2017 release) server (see text footnote 1) (Dai and Zhao, 2011). To reduce the number of false predictions, the maximum expectation threshold was set to the value of 3.0. The cutoff length of nucleotides for complementarity scoring, hsp (high-scoring segment pair) size, was set as the length of the mature miRNAs. The maximum energy of unpairing (UPE) of the target site was set as 25 kcal. The flanking length around the target site for target accessibility analysis was 17 bp upstream and 13 bp downstream. The range of central mismatch leading to translation inhibition was adjusted as 9–11 nt. No gap and no more than four mismatches between miRNA and its target (G-U pair count as 0.5 mismatch) were allowed. The target genes of mango lncRNAs were predicted by using the LncTar tool (version 1.0) (Li et al., 2015) with the normalized binding free energy (ndG) cutoff value less than 0.1.

Prediction of lncRNAs as miRNA Target or Target Mimic

To predict the lncRNAs as the target genes of miRNAs, psRNATarget (2017 release) (Dai and Zhao, 2011) was used as previously mentioned in the interaction prediction of miRNAs and mRNAs. For target mimic prediction, the TAPIR server (version 1.2) (Bonnet et al., 2010) was used in this study. TAPIR is a web server for the prediction of plant miRNA targets including target mimics.

Functional Annotation and Pathway Analysis of Target Genes

The gene ontology (GO) analysis of the identified target transcripts was executed by combining both BLASTx data and InterProScan analysis data by means of the BLAST2GO_5.2.5 software (Conesa et al., 2005). The GO enrichment analysis was done by using Fisher's exact test with multiple testing correction of false discovery rate (FDR). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was also performed for a better understanding of the functions of the target genes.

Conservation Analysis of lncRNAs

The analysis of the conservation of mango lncRNAs was detected by using BLASTn against all lncRNA sequences from the plant

[4]https://www.ncbi.nlm.nih.gov/orffinder/

[5]http://rfam.xfam.org

[6]http://www.mirbase.org/

lncRNA database, CANTATAdb_2.0 (Szcześniak et al., 2016), with *e*-value cutoff 1e-20.

Interaction Network of miRNAs, lncRNAs, and Their Target Genes

Finally, the interaction network of miRNAs, lncRNAs, and their target genes was visualized by using Cytoscape_3.7.2 (Shannon et al., 2003).

RESULTS

Identification and Characterization of Mango miRNAs

Most of the plant miRNAs are evolutionarily conserved from species to species (Dezulian et al., 2005; Weber, 2005), and this indicates the powerful strategy for the identification of new miRNAs by using the already known miRNAs (Zhang et al., 2006b). Many conserved miRNAs have been identified from the expressed sequence tag (EST) (Zhang et al., 2005; Frazier and Zhang, 2011) and genome survey sequence (GSS) (Pan et al., 2007) by using this homology search approach. For mango, there are no GSS data and the available EST data for mango were only 1,709 and it was not sufficient for the identification of miRNA. Hence, unigenes (107,744) (see text footnote 3) were used for the identification of miRNAs in this study. Unigene is a unique transcript that is transcribed from a genome, and many miRNAs have been identified from the unigenes of many plant species such as *Artemisia annua* (Pérez-Quintero et al., 2012), coconut (Naganeeswaran et al., 2015), litchi fruit (Yao et al., 2015), and black pepper (Asha et al., 2016).

From the mango unigenes, we have identified 104 miRNAs by following the identification workflow explained in **Figure 1**. The length of the resulting mature miRNAs is in the range of 18–22 nt. Among them, nearly 40% (41 miRNAs) of mango mature miRNAs are in the length of 18 nt and 6 miRNAs have the length of 22 nt. Thirty-two miRNAs, 17 miRNAs, and 8 miRNAs are 19, 20, and 21 nt of length, respectively (**Figure 2A**). The potential 104 pre-miRNAs of mango were predicted based on the parameters by Dr. Zhang (Zhang et al., 2005), and the MFEI values were also calculated as the MFEI gave the best prediction of miRNAs (Zhang et al., 2006c). The precursor length of mango miRNAs (MmiRs) was varied significantly from 67 to 144 nt with an average length of 94 nt. We denoted the name of mango miRNA as MmiR with the numbers. The secondary structure of

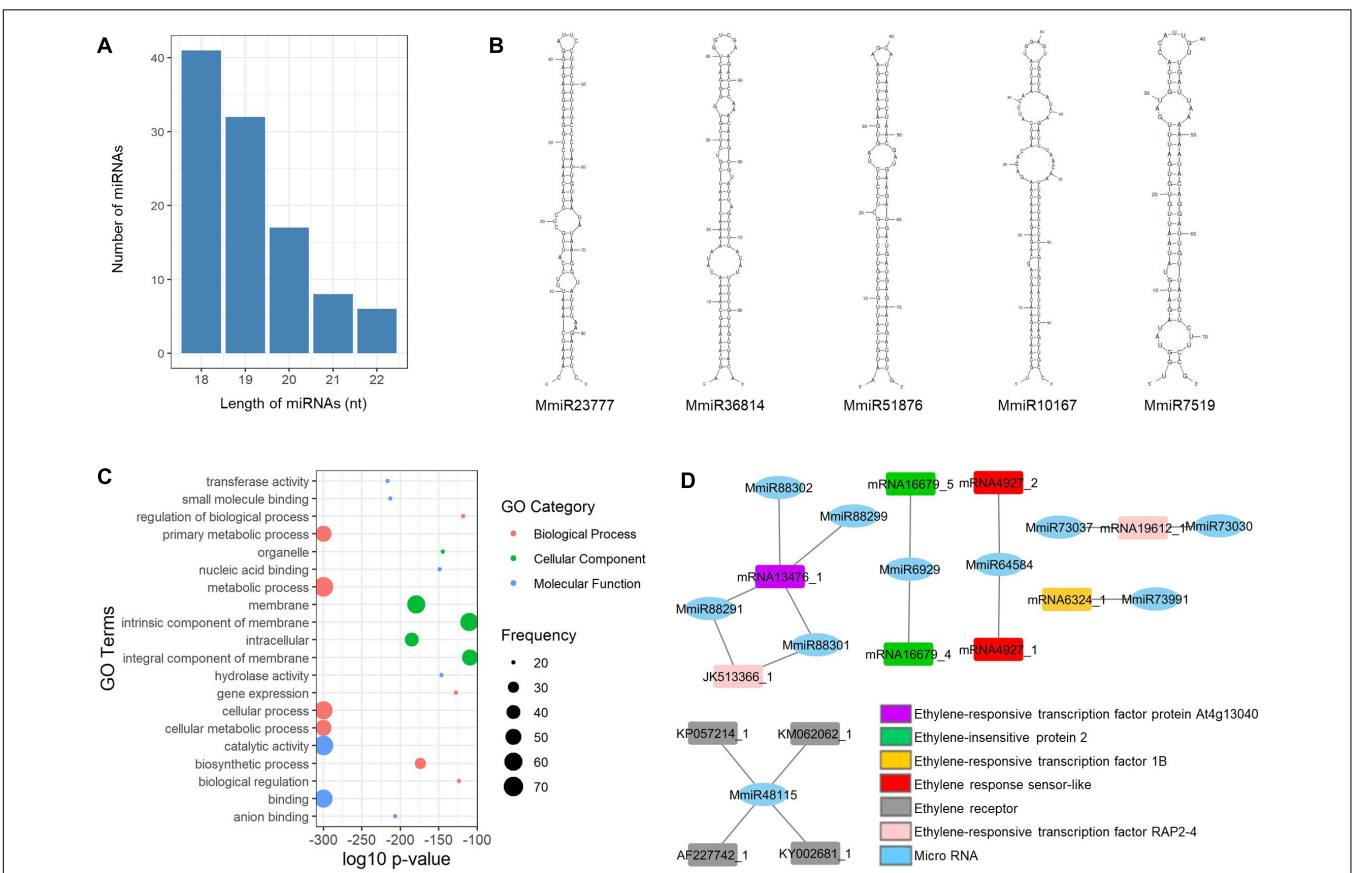


FIGURE 2 | Mango miRNAs. **(A)** The number and length distribution of miRNAs; **(B)** hairpin structures of five precursor miRNAs involved in the developmental process of mango; **(C)** GO enrichment analysis of highly enriched top 20 target genes of miRNAs; **(D)** interaction among the 10 mango miRNAs and six ethylene-responsive target genes.

precursor sequences was predicted by Zuker folding algorithm in MFOLD. **Figure 2B** shows the hairpin structures of five miRNAs (MmiR23777, MmiR36814, MmiR51876, MmiR10167, and MmiR7519) involved in the developmental process of mango according to the result of GO enrichment and KEGG pathway analysis. The average MFE of pre-miRNAs is 29.92. The MFEI values were also calculated and were in the range of 0.7–1.45 with the average MFEI of 0.84. The A+U content was in the range of 42–70% with an average of 62.9% (**Supplementary Table 1**).

Target Gene Analysis of miRNAs

According to the result of target gene prediction by the psRNATarget server (Dai and Zhao, 2011), all the newly identified mango miRNAs could bind to their targets and a total of 2,347 target genes were predicted for 104 mango miRNAs. The predicted target genes were annotated and assigned to GO terms by BLAST2GO. The top 20 highly enriched GO terms of miRNA target genes are visualized in **Figure 2C**. The metabolic process and cellular process are highly enriched GO terms of the biological process, and membrane and intracellular are highly enriched GO terms of the cellular component. In molecular function, catalytic activity and binding are highly enriched GO terms.

From KEGG pathway analysis, 310 targets were involved in the 103 different KEGG pathways. Purine metabolism was the pathway with the highest target genes: 136. The predicted miRNAs, their target genes, target descriptions, target GO terms, and target KEGG pathways are shown in **Supplementary Table 2**.

Identification and Characterization of Mango lncRNAs

For the identification of lncRNAs, a total of 277,071 RNA transcripts from Zill (Wu et al., 2014), Shelly (Luria et al., 2014; Sivankalyani et al., 2016), and Keitt (Tafolla-Arellano et al., 2017) mango cultivars were used. First, the sequences less than 200 nt were removed because lncRNAs were always longer than 200 nt. Then, the coding transcripts were removed by their protein-coding potential, homology with known proteins, and potential ORFs. Finally, the housekeeping RNAs and the precursor of miRNAs were removed. After a series of filtering steps, a total of 31,226 candidate lncRNAs were predicted.

The temperature-responsive lncRNAs were then defined by fold change value and FDR adjusted *p*-value to filter out the significantly expressed mango lncRNAs. The lncRNAs with fold change value less than -2 were defined as the downregulated lncRNAs and the lncRNAs greater than 2 as the upregulated lncRNAs. The FDR adjusted *p*-value was set to 0.05. As a result, 24 lncRNAs were significantly expressed to heat stress (55°C hot water brushing) and 7,586 lncRNAs to cold stress (5, 8, or 12°C) (**Figure 3A**). We denoted the name of the heat-responsive lncRNAs of mango as HRlnc with numbers and the cold-responsive lncRNAs as CRlnc with numbers. 24 HRlncRNA

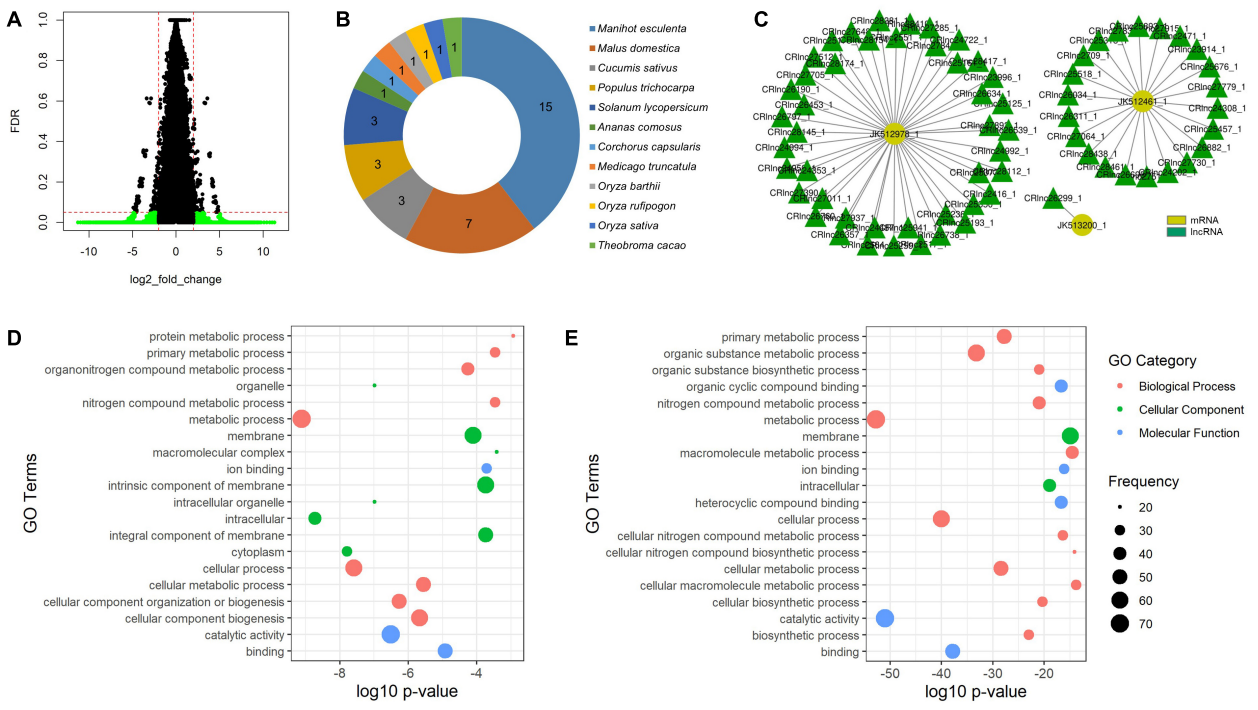


FIGURE 3 | Temperature-responsive lncRNAs of mango. **(A)** Volcano plot of cold-responsive lncRNAs of mango; green plots represent significantly expressed cold-responsive lncRNAs with false discovery rate (adjusted *p*-value) of less than 0.05 and log2 fold change of less than –2 or greater than 2; **(B)** conserved number of mango lncRNAs in different plant species; **(C)** interaction subnetwork between three low temperature-responsive proteins and cold-responsive lncRNAs of mango; **(D)** GO enrichment analysis of highly enriched genes targeted by heat-responsive lncRNAs; **(E)** GO enrichment analysis of highly enriched target genes of cold-responsive lncRNAs.

sequences are shown in the **Supplementary Data Sheet 1** and 7,586 CRlncRNA sequences are in the **Supplementary Data Sheet 2**. In heat-responsive lncRNAs, 18 lncRNAs were upregulated and 6 lncRNAs were downregulated. The length of heat-responsive lncRNAs ranged from 213 to 1,186 nt (**Supplementary Table 3**). Among the 7,619 cold-responsive lncRNAs, 4,335 were upregulated and 3,251 were downregulated. The length of cold-responsive lncRNAs was in the range of 201–2,746 nt (**Supplementary Table 4**).

Conservation Analysis of lncRNAs

The mango lncRNAs were searched by using BLASTn against the plant lncRNA database, CANTATAdb, with *e*-value cutoff 1e-20 to check their evolutionary conservation. As a result, no heat-responsive lncRNAs were conserved and 22 cold-responsive lncRNAs were conserved with 12 different plant species. Among these different plants, *Manihot esculenta* (cassava) is the most highly conserved plant and 15 cold-responsive mango lncRNAs were conserved with its lncRNAs (**Figure 3B**).

Target Gene Prediction of lncRNAs

To analyze the interaction of the newly identified lncRNAs of mango with protein-coding genes, the lncRNA target prediction tool LncTar was used. A total of 1,998 mango mRNAs downloaded from NCBI were used for the target prediction of lncRNAs. From the resulting data, 6,975 lncRNAs interacted with 1,985 target mRNAs. To analyze the functional overview of the identified lncRNAs, the targets of the identified lncRNAs were predicted by BLAST2GO. Among the 24 heat-responsive lncRNAs, 8 lncRNAs had 115 target genes (SCL14, STP13, Hsp70, At4g39970, ACO1, and so on) involved in plant development and stress response. In cold-responsive lncRNAs, 6,951 lncRNAs interacted with 1,985 target genes. The WRKY proteins are a large family of transcriptional regulators in higher plant and 64 cold-responsive lncRNAs interact with the WRKY gene family in this study (**Figure 3C**).

Moreover, functional prediction of the target genes of the identified lncRNAs was performed by GO enrichment analysis. For the target genes of heat-responsive lncRNAs, metabolic process, cellular process, cellular component biogenesis, and cellular metabolic process were highly enriched in the biological process. In the cellular component analysis, GO terms associated with membranes and intracellular were highly enriched. Catalytic activity and binding GO terms were highly enriched in molecular function analysis (**Figure 3D**). For the target genes of cold-responsive lncRNAs, metabolic processes and cellular processes were highly enriched for biological processes. In the cellular component analysis, GO terms related to membranes and intracellular were highly enriched. For molecular function analysis, most of the enriched GO terms were related to catalytic activity and binding (**Figure 3E**).

From the results of the KEGG pathway analysis, heat-responsive lncRNAs had target genes involved in 17 KEGG pathways (**Supplementary Table 5**). Among these different pathways, amino sugar and nucleotide sugar metabolism was the most significant pathway and eight target genes were involved in this pathway. For cold-responsive lncRNAs, 209 target genes had been mapped to 87 KEGG pathways (**Supplementary Table 6**). JK513026_1, alcohol dehydrogenase 1 (ADH1, EC:1.1.1.1), was the most enriched target gene and involved in 12 different pathways.

Prediction of lncRNAs as miRNA Targets

To analyze the direct interaction of miRNAs and lncRNAs of mango, the psRNATarget server (Dai and Zhao, 2011) was used to predict the target lncRNAs of miRNAs. The resulting data showed that three heat-responsive lncRNAs interacted with six miRNAs (**Supplementary Table 7**). For cold-responsive lncRNAs, 763 lncRNAs had 1,203 pairs of interactions with 89 miRNAs (**Supplementary Table 8**).

The miRNA target mimicry search was also performed by using TAPIR. No heat-responsive lncRNA acts as the target mimic of miRNAs. However, 20 cold-responsive lncRNAs were predicted as the target mimics of 20 miRNAs (**Supplementary Table 9**). CRlnc31221 was the target mimic of MmiR5408 which targeted eight cold-responsive lncRNAs and 47 target genes (**Figure 4B**). The schematic diagram of the interaction between MmiR5408 and its target mimic cold-responsive lncRNA, CRlnc31221, is shown in **Figure 4C**.

The interaction network of mango ncRNAs (miRNAs, lncRNAs, and mimic) and their target genes was visualized by using Cytoscape, which contained a total of 5,388 pairs of interaction among miRNAs, lncRNAs, and their targets (**Figure 4A**). These interactions were 4,155 pairs of 104 miRNAs and 2,347 mRNAs, 1,203 pairs of 89 miRNAs and 763 cold-responsive lncRNAs, six pairs of six miRNAs and four heat-responsive lncRNAs, and 24 pairs of 20 miRNAs and their 20 target mimics.

DISCUSSION

Identification, Characterization, and Target Gene Prediction of miRNAs

The length of the predicted 104 mature miRNAs was in the range of 18–22 nt, but the length of precursor miRNAs varied significantly from 67 to 144 nt with an average length of 94 nt. The predicted mango miRNAs belong to 86 different families. Among them, over 70% of the miRNA families have only one family member. The highest five family members were found in the miR2673 family followed by miR159, with four family members. The remaining miRNAs have a family member of two or three. Therefore, we can see that the mango miRNA distribution across various families is highly heterogeneous.

Previous studies have already proved that plant miRNAs bind to their targets in a perfect or nearly perfect complementarity (Bartel, 2004, Zhang et al., 2006a) and the psRNATarget server (Dai and Zhao, 2011) was used to search the target gene of mango miRNAs in this study. Both mRNAs collected from NCBI and identified in this study were used as the target candidates of miRNAs due to the absence of *M. indica* target candidates in the psRNATarget server. Some previous studies indicated that miR156 was a master regulator of the juvenile phase in plants and it targeted the Squamosa promoter binding protein-Like (SPL) gene family to regulate the transition from vegetative phase to floral phase in *Arabidopsis* (Gandikota et al., 2007;

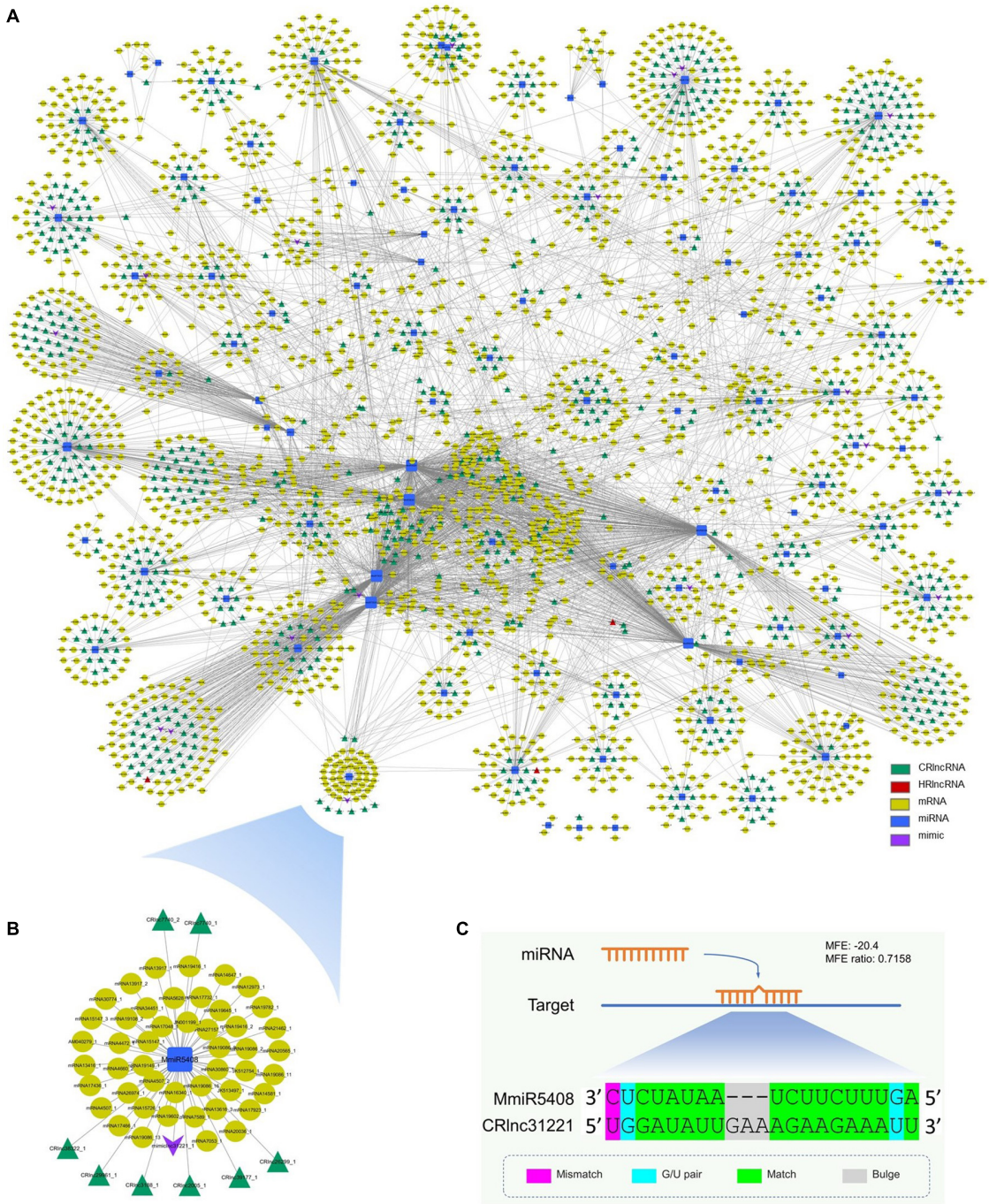


FIGURE 4 | Interaction network among mango ncRNAs and their targets; green triangles represent cold-responsive lncRNAs, red triangles represent heat-responsive lncRNAs, yellow ellipses are used for target genes, blue rectangles are for miRNAs, and purple V-shapes are for target mimics. **(A)** Network of interaction among newly identified miRNAs, newly identified lncRNAs (cold-responsive lncRNAs and heat-responsive lncRNAs), and newly identified target mimic of miRNAs and mRNAs; **(B)** subnetwork of interaction among MmiR5408, its target mimic CRlnc31221, target CRlncRNAs, and 47 target genes; **(C)** schematic diagram of the interaction between MmiR5408 and its target mimic cold-responsive lncRNA, CRlnc31221.

Wu et al., 2009; Yamaguchi et al., 2009), maize (Chuck et al., 2007) and rice (Jiao et al., 2010; Jeong et al., 2011). In mango, MmiR105772, a family of miR156, also bound to its target SPL6, and thus, the predicted targets of mango miRNAs were in the agreement with the previously published papers in other plants. The resulting data from psRNATarget also showed that only one miRNA (MmiR1653) had a single target gene which was a member of the miR482 family and

bound to monodehydroascorbate reductase four enzyme, an important gene related to the nutritional quality of mango fruit (Pandit et al., 2010). All the other miRNAs could target multiple genes and some miRNAs had over 200 target genes. For example, MmiR73030 had 230 target genes and these target genes are involved in 16 KEGG pathways such as biosynthesis of antibiotics, purine metabolism, sulfur metabolism, glycerophospholipid metabolism, T cell receptor signaling pathway, steroid degradation, and so on.

Sivankalyani et al. (2016) published that the mango stress-response pathways were activated by cyclic nucleotide-gated channel (CNGC) and leucine-rich repeat receptor (Lrr). In this study, we found that MmiR90392 targeted CNGC1, and MmiR68471 and MmiR68478 targeted Lrr2. Moreover, MmiR10167 and MmiR15558 are bound to the stress WRKY transcription factor 44 which plays a major role in plant defense to biotic and abiotic stresses. MmiR78769 and MmiR101928 are also bound to their target genes of phospholipase A and phospholipase D which were key factors in plant responses to biotic and abiotic stresses (Xue et al., 2007). The ethylene response could improve the tolerance of mango fruit to chilling stress (Lederman et al., 1997), and 10 mango miRNAs identified in this study had six ethylene-responsive target genes such as ethylene-insensitive protein and ethylene-responsive transcription factor (**Figure 2D**). So, these newly identified mango miRNAs have potential roles in the chilling stress-responsive process of mango.

Two mango miRNAs (MmiR23777 and MmiR36814) also targeted the auxin efflux carrier which had a potential role in mango plant organ development (Li et al., 2012). A total of 17 miRNAs interacted with auxin-related genes. MmiR51876 was a miRNA that targeted auxin-responsive protein. The pentose and glucuronate interconversion pathway, phenlypropanoid biosynthesis pathway, and alpha-linolenic acid metabolism pathway were KEGG pathways involved in the adventitious root formation of mango cotyledon segments (Li et al., 2017). In this study, nine miRNAs bound to eight target genes are involved in these three pathways for mango root formation. MmiR10167 bound to target genes is involved in phenlypropanoid biosynthesis pathway and MmiR7519 bound to target genes is involved in alpha-linolenic acid metabolism pathway. From these findings, it was observed that these five mango miRNAs (MmiR23777, MmiR36814, MmiR51876, MmiR10167, and MmiR7519) are involved in the developmental process of mango (**Figure 2B**).

Identification, Characterization, and Target Gene Prediction of lncRNAs

As the genome sequence of mango is not available till now, the *de novo* assembled transcriptome sequences were used for the identification of lncRNAs in this study. A total of 277,071 RNA transcripts from Zill (Wu et al., 2014), Shelly (Luria et al., 2014; Sivankalyani et al., 2016), and Keitt (Tafolla-Arellano et al., 2017) mango cultivars studied by former researchers were used, and a total of 31,226 candidate lncRNAs were predicted in this study. Heat and cold stresses can affect the important mechanisms of mango such as stress response, defense mechanism, sugar and flavonoid metabolism, photosynthesis, and floral morphogenesis (Núñez-Elisea and Davenport, 1994; Allen et al., 2000; Luria et al., 2014; Sudheeran et al., 2018). Therefore, the temperature-responsive lncRNAs were identified from the resulting 31,226 lncRNAs. Among them, 24 lncRNAs were significantly expressed to heat stress and 7,586 lncRNAs to cold stress. The most significantly expressed downregulated heat-responsive lncRNA was HRlnc25944 with a fold change value of 6.22. HRlnc11351 and HRlnc27371 were the mostly expressed upregulated lncRNAs with a fold change value greater than 7. For the cold-responsive lncRNAs, CRlnc10871 was the mostly expressed downregulated lncRNA (FC value -11.19), and CRlnc26299, CRlnc30496, and CRlnc36473 were the most significantly expressed lncRNAs with a fold change value greater than 11.

No heat-responsive lncRNAs were conserved, but 0.29% of cold-responsive lncRNAs were conserved with 12 different plant species (**Figure 3B**). Among them, the highest conserved lncRNAs were CRlnc32663 and CRlnc47883, each of which was conserved with four different lncRNAs of other plants. CRlnc32663 was conserved with four different lncRNAs of three different plant species such as *Manihot esculenta* (cassava), *Malus domestica* (apple), and *Populus trichocarpa* (the black cottonwood). CRlnc47883 was also conserved with four lncRNAs of *Oryza rufipogon* (brownbeard rice), *Oryza barthii* (wild rice), and *Solanum lycopersicum* (tomato).

For heat-responsive lncRNAs, 8 bound to 115 target genes were involved in plant development and stress response. HRlnc11351 was the most significantly expressed upregulated lncRNAs with a fold change value of 7.55 and bound to six heat shock proteins. In cold-responsive lncRNAs, CRlnc26299 was one of the most significantly expressed upregulated lncRNAs and bound to RC12B (JK513200_1), which is a low-temperature and salt-responsive protein found in *Arabidopsis thaliana* (Medina et al., 2001). The WRKY proteins are a large family of transcriptional regulators in higher plants and exhibited variable expression patterns in response to chilling stress in cucumber (Ling et al., 2011), mango (Sivankalyani et al., 2016), and rice (Ramamoorthy et al., 2008). In this study, 64 cold-responsive lncRNAs interact with the WRKY gene family. So, we can observe that the cold-responsive lncRNAs of mango have interaction with the target genes that are expressed at low-temperature stress.

Gene ontology enrichment analysis and KEGG pathway analysis were performed for a better understanding of the target genes of newly identified lncRNAs. From the GO enrichment analysis result, we could see that both types of heat-responsive lncRNAs and cold-responsive lncRNAs were highly enriched in metabolic processes and cellular processes in the biological process analysis. In the cellular component analysis, GOs related to membranes, intracellular, and cytoplasm were highly enriched for both types of lncRNAs. Meanwhile, most of the enriched GO terms in both types of lncRNAs were related to catalytic activity and binding for molecular function analysis. Therefore, we could see that the GO terms highly enriched in both heat-responsive and cold-responsive lncRNAs were not quite different.

Among the 17 KEGG pathways of the target genes of the heat-responsive lncRNAs, amino sugar and nucleotide sugar

metabolism was the most significant pathway, and eight target genes were involved in this pathway. As mentioned above, HRlnc11351 was the most significantly expressed upregulated lncRNAs and its target gene, JK513625_1, is 3-ketoacyl-CoA thiolase 2 (KAT2, EC:2.3.1.16), which could be mapped to nine different pathways such as benzoate degradation; fatty acid elongation; biosynthesis of unsaturated fatty acids; alpha-linolenic acid metabolism; fatty acid degradation; valine, leucine, and isoleucine degradation; biosynthesis of antibiotics: geraniol degradation; and ethylbenzene degradation according to the result of the KEGG pathway analysis. In *Arabidopsis*, KAT2 is an enzyme that catalyzes the β-oxidation of fatty acid and involves in abscisic acid (ABA) signal transduction (Jiang et al., 2011). The phytohormone ABA plays an important role in plant development and adaptation to diverse environmental stresses. Therefore, HRlnc11351 may be involved and played an important role in mango development and stress response by targeting KAT2. For cold-responsive lncRNAs, 209 target genes had been mapped to 86 KEGG pathways. JK513026_1, alcohol dehydrogenase 1 (ADH1, EC:1.1.1.1), was the most enriched target gene and involved in 12 different pathways including glycolysis/gluconeogenesis; metabolism of xenobiotics by cytochrome P450; glycine, serine, and threonine metabolism; methane metabolism; fatty acid degradation; and so on. In plants, ADH genes are involved in mediating stress responses and developments. In mango, ADH1 has an important role in fruit ripening (Singh et al., 2010), and thus, cold-responsive lncRNAs that target the ADH1 gene may play an important role in the mango fruit ripening process. According to the KEGG pathway analysis results, purine metabolism and biosynthesis of antibiotics were the highly enriched pathways among 86 pathways and more than 50 target genes were enriched in each pathway.

Interaction Between lncRNAs and miRNAs

The interaction between miRNAs and lncRNAs showed that most of the miRNAs had targeted more than one lncRNAs and only eight miRNAs had single target lncRNAs. The number of lncRNAs targeted by a single miRNA was in the range of 1–90. A total of 90 target lncRNAs were found for MmiR73030, which also targeted 230 mRNAs. This miRNA had the highest target numbers in both lncRNAs and mRNAs.

Long non-coding RNAs not only can be targeted by miRNAs to reduce the stability of lncRNAs but also can function as molecular decoys or sponges of miRNAs (Salmena et al., 2011; Wu et al., 2013). So, the miRNA target mimicry search was performed by using TAPIR, which is a web server for the prediction of plant miRNA targets including target mimics. Although no heat-responsive lncRNA acts as the target mimic of miRNAs, 20 cold-responsive lncRNAs were predicted as the target mimics of 20 miRNAs. CRlnc31221 was the target mimic of MmiR5408 which targeted 8 cold-responsive lncRNAs and 47 target genes. These target genes were involved in starch and sucrose metabolism, inositol phosphate metabolism, and phenylpropanoid biosynthesis pathways, which are important for plant growth and development and the plant's response toward biotic and abiotic stresses. During target mimicry, the interactions between miRNAs and their authentic targets were blocked by binding of decoy RNA to miRNAs *via* partially complementary sequences (Franco-Zorrilla et al., 2007). So, the target mimicry of CRlnc31221 had the potential regulation effect to the interaction between the target genes and MmiR5408 (**Figure 4B**).

CONCLUSION

In conclusion, this study identified 104 miRNAs and 7,610 temperature-responsive lncRNAs from mango transcriptome sequences, and the interactions of these ncRNAs with their target genes were also predicted. MmiR105772 is bound to SPL6 gene that regulates the transition from the vegetative phase to the floral phase of plants. MmiR1653 is bound to monodehydroascorbate reductase 4 enzyme that regulates the nutritional quality of mango fruit. MmiR78769 and MmiR101928 are also bound to their target genes of phospholipase A and phospholipase D which were key factors in plant responses to biotic and abiotic stresses. HRlnc11351 may be involved and has an important role in mango development and stress response by targeting KAT2, an enzyme that catalyzes β-oxidation of fatty acid and is involved in abscisic acid (ABA) signal transduction. Cold-responsive lncRNAs that target the ADH1 gene may play an important role in mango fruit ripening process because ADH1 has an important role in mango fruit ripening. CRlnc26299, one of the most significantly expressed upregulated cold-responsive lncRNAs, is bound to RC12B (JK513200_1), which is the low-temperature and salt-responsive protein. According to these results, the newly identified mango ncRNAs, like other plant ncRNAs, have a potential role in metabolic pathways including plant growth and developmental process, pathogen defense mechanism, and stress-responsive process. Therefore, the resulting data of this project may help for further prediction of the specific functions of mango ncRNAs through wet lab experiments.

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

NM and MC designed the research. NM and PZ performed the bioinformatics analysis. NM and YC analyzed the plant gene data. All authors approved the final manuscript.

FUNDING

This research was supported by the Talented Young Scientist Program organized by the China Ministry of Science and Technology. MC's Lab is grateful for the support from MOST

(2018YFC0310600 and 2016YFA0501704), NSFC (31771477, 31571366 and 32070677), and JCIC-MCP/CIC-MCP.

ACKNOWLEDGMENTS

The authors thank all lab members especially Yong Jing and Cong Feng for their suggestions during this research work. This manuscript has been released as a preprint at Research Square (https://doi.org/10.21203/rs.3.rs22698/v1) (Moh et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.607248/full#supplementary-material

Supplementary Table 1 | miRNAs data.

Supplementary Table 2 | miRNAs targets data.

Supplementary Table 3 | Heat responsive lncRNAs data.

Supplementary Table 4 | Cold responsive lncRNAs data.

Supplementary Table 5 | KEGG pathways of the target of heat responsive lncRNAs.

Supplementary Table 6 | KEGG pathways of the target of cold responsive lncRNAs.

Supplementary Table 7 | Heat responsive lncRNAs as target of miRNAs.

Supplementary Table 8 | Cold responsive lncRNAs as target of miRNAs.

Supplementary Table 9 | Cold responsive lncRNAs as target mimic of miRNAs.

Supplementary Data Sheet 1 | Heat responsive lncRNAs sequences.

Supplementary Data Sheet 2 | Cold responsive lncRNAs sequences.

REFERENCES

Allen, D. J., Ratner, K., Giller, Y. E., Gussakovsky, E. E., Shahak, Y., and Ort, D. R. (2000). An overnight chill induces a delayed inhibition of photosynthesis at midday in mango (*Mangifera indica* L.). *J. Exp. Bot.* 51, 1893–1902.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., et al. (2003). A uniform system for microRNA annotation. *RNA* 9, 277–279. doi: 10.1261/rna.2183803

Asha, S., Sreekumar, S., and Soniya, E. (2016). Unravelling the complexity of microRNA-mediated gene regulation in black pepper (*Piper nigrum* L.) using high-throughput small RNA profiling. *Plant Cell Rep.* 35, 53–63. doi: 10.1007/s00299-015-1866-x

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Blignaut, M. (2012). Review of Non-coding RNAs and the epigenetic regulation of gene expression: a book edited by Kevin Morris. *Epigenetics* 7, 664–666. doi: 10.4161/epi.20170

Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* 26, 1566–1568. doi: 10.1093/bioinformatics/btq233

Chuck, G., Cigan, A. M., Saeteurn, K., and Hake, S. (2007). The heterochronic maize mutant Corngrass1 results from overexpression of a tandem microRNA. *Nat. Genet.* 39, 544–549. doi: 10.1038/ng2001

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

Dai, X., and Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.* 39(suppl. 2), W155–W159.

Dezulian, T., Palatnik, J. F., Huson, D., and Weigel, D. (2005). Conservation and divergence of microRNA families in plants. *Genome Biol.* 6:13.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., and Pagano, A. (2007). The expanding RNA polymerase III transcriptome. *Trends Genet.* 23, 614–622. doi: 10.1016/j.tig.2007.09.001

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., et al. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* 39, 1033–1037. doi: 10.1038/ng2079

Frazier, T. P., and Zhang, B. (2011). *Identification of Plant MicroRNAs Using Expressed Sequence tag Analysis Plant Reverse Genetics*. New York, NY: Springer, 13–25.

Galán Saúco, V. (2010). "Worldwide mango production and market: current situation and future prospects," in *Paper Presented at the IXth International Mango Symposium*, (Leuven: International Society for Horticultural Science), 992.

Gandikota, M., Birkenbihl, R. P., Höhmann, S., Cardon, G. H., Saedler, H., and Huijser, P. (2007). The miRNA156/157 recognition element in the 3′UTR of the *Arabidopsis* SBP box gene SPL3 prevents early flowering by translational inhibition in seedlings. *Plant J.* 49, 683–693. doi: 10.1111/j.1365-313x.2006.02983.x

Geisler, S., and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.* 14, 699–712. doi: 10.1038/nrm3679

Han, S., Liang, Y., Ma, Q., Xu, Y., Zhang, Y., Du, W., et al. (2019). LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief. Bioinform.* 20, 2009–2027. doi: 10.1093/bib/bby065

Hirano, R., Htun Oo, T., and Watanabe, K. (2010). Myanmar mango landraces reveal genetic uniqueness over common cultivars from Florida, India, and Southeast Asia. *Genome* 53, 321–330. doi: 10.1139/g10-005

Jalali, S., Bhartiya, D., Lalwani, M. K., Sivasubbu, S., and Scaria, V. (2013). Systematic transcriptome wide analysis of lncRNA-miRNA interactions. *PloS One* 8:e53823. doi: 10.1371/journal.pone.0053823

Jeong, D. H., Park, S., Zhai, J., Gurazada, S. G. R., De Paoli, E., Meyers, B. C., et al. (2011). Massive analysis of rice small RNAs: mechanistic implications of regulated microRNAs and variants for differential target RNA cleavage. *Plant Cell* 23, 4185–4207. doi: 10.1105/tpc.111.089045

Jiang, T., Zhang, X. F., Wang, X. F., and Zhang, D. P. (2011). Arabidopsis 3-ketoacyl-CoA thiolase-2 (KAT2), an enzyme of fatty acid β-oxidation, is involved in ABA signal transduction. *Plant Cell Physiol.* 52, 528–538. doi: 10.1093/pcp/pcr008

Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., et al. (2010). Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nat. Genet.* 42:541. doi: 10.1038/ng.591

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35(supp. 2), W345–W349.

Kurihara, Y., and Watanabe, Y. (2004). *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc. Natl. Acad. Sci. U.S.A.* 101, 12753–12758. doi: 10.1073/pnas.0403115101

Lauricella, M., Emanuele, S., Calvaruso, G., Giuliano, M., and D'Anneo, A. (2017). Multifaceted health benefits of *Mangifera indica* L.(Mango): the inestimable value of orchards recently planted in Sicilian rural areas. *Nutrients* 9:525. doi: 10.3390/nu9050525

Lederman, I. E., Zauberman, G., Weksler, A., Rot, I., and Fuchs, Y. (1997). Ethylene-forming capacity during cold storage and chilling injury development in 'Keitt'mango fruit. *Postharvest Biol. Technol.* 10, 107–112. doi: 10.1016/s0925-5214(96)00060-9

Li, J., Ma, W., Zeng, P., Wang, J., Geng, B., Yang, J., et al. (2015). LncTar: a tool for predicting the RNA targets of long noncoding RNAs. *Brief. Bioinform.* 16, 806–812. doi: 10.1093/bib/bbu048

Li, Y. H., Zhang, H. N., Wu, Q. S., and Muday, G. K. (2017). Transcriptional sequencing and analysis of major genes involved in the adventitious root formation of mango cotyledon segments. *Planta* 245, 1193–1213. doi: 10.1007/s00425-017-2677-9

Li, Y. H., Zou, M. H., Feng, B. H., Huang, X., Zhang, Z., and Sun, G. M. (2012). Molecular cloning and characterization of the genes encoding an auxin efflux carrier and the auxin influx carriers associated with the adventitious root formation in mango (*Mangifera indica* L.) cotyledon segments. *Plant Physiol. Biochem.* 55, 33–42. doi: 10.1016/j.plaphy.2012.03.012

Ling, J., Jiang, W., Zhang, Y., Yu, H., Mao, Z., Gu, X., et al. (2011). Genome-wide analysis of WRKY gene family in *Cucumis sativus*. *BMC Genomics* 12:471.

Luria, N., Sela, N., Yaari, M., Feygenberg, O., Kobiler, I., Lers, A., et al. (2014). De-novo assembly of mango fruit peel transcriptome reveals mechanisms of mango response to hot water treatment. *BMC Genomics* 15:957.

Medina, J. n, Catalá, R., and Salinas, J. (2001). Developmental and stress regulation of RCI2A and RCI2B, two cold-inducible genes of *Arabidopsis* encoding highly conserved hydrophobic proteins. *Plant Physiol.* 125, 1655–1666. doi: 10.1104/pp.125.4.1655

Megha, S., Basu, U., Rahman, M. H., and Kav, N. N. (2015). *The Role of Long Non-Coding RNAs in Abiotic Stress Tolerance in Plants Elucidation of Abiotic Stress Signaling in Plants*. New York, NY: Springer, 93–106.

Moh, N. M. M., Zhang, P., Chen, Y., and Chen, M. (2020). Computational identification of miRNAs and temperature responsive lncRNAs from Mango (*Mangifera indica*, L.). *Preprint* doi: 10.21203/rs.3.rs-22698/v1

Mukherjee, S. (1953). The mango—its botany, cultivation, uses and future improvement, especially as observed in India. *Econ. Bot.* 7, 130–162. doi: 10.1007/bf02863059

Mukherjee, S. (1972). Origin of mango (*Mangifera indica*). *Econ. Bot.* 26, 260–264. doi: 10.1007/bf02861039

Naganeeswaran, S., Fayas, T., Rachana, K., and Rajesh, M. (2015). Computational prediction and characterization of miRNA from coconut leaf transcriptome. *J. Appl. Hortic.* 17, 12–17. doi: 10.37855/jah.2015.v17i01.03

Núñez-Elisea, R., and Davenport, T. L. (1994). Flowering of mango trees in containers as influenced by seasonal temperature and water stress. *Sci. Hortic.* 58, 57–66. doi: 10.1016/0304-4238(94)90127-9

Pan, X., Zhang, B., Francisco, M. S., and Cobb, G. P. (2007). Characterizing viral microRNAs and its application on identifying new microRNAs in viruses. *J. Cell. Physiol.* 211, 10–18. doi: 10.1002/jcp.20920

Panda, D., Dehury, B., Sahu, J., Barooah, M., Sen, P., and Modi, M. K. (2014). Computational identification and characterization of conserved miRNAs and their target genes in garlic (*Allium sativum* L.) expressed sequence tags. *Gene* 537, 333–342. doi: 10.1016/j.gene.2014.01.010

Pandit, S. S., Kulkarni, R. S., Giri, A. P., Köllner, T. G., Degenhardt, J., Gershenzon, J., et al. (2010). Expression profiling of various genes during the fruit development and ripening of mango. *Plant Physiol. Biochem.* 48, 426–433. doi: 10.1016/j.plaphy.2010.02.012

Pérez-Quintero, ÁL., Sablok, G., Tatarinova, T. V., Conesa, A., Kuo, J., and López, C. (2012). Mining of miRNAs and potential targets from gene oriented clusters of transcripts sequences of the anti-malarial plant, *Artemisia annua*. *Biotechnol. Lett.* 34, 737–745. doi: 10.1007/s10529-011-0808-0

Ramamoorthy, R., Jiang, S. Y., Kumar, N., Venkatesh, P. N., and Ramachandran, S. (2008). A comprehensive transcriptional profiling of the WRKY gene family in rice under various abiotic and phytohormone treatments. *Plant Cell Physiol.* 49, 865–879. doi: 10.1093/pcp/pcn061

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513–520.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

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

Siddiq, M., Akhtar, S., and Siddiq, R. (2012). "Mango pocessing, products and nutrition," in *Tropical and Subtripical Fruits: Posthrvest and Physiology, Processing and Packaging*, ed. M. Siddiq (Ames, IA: John Wiley & Sons), 277–297. doi: 10.1002/9781118324097.ch15

Singh, R. K., Sane, V. A., Misra, A., Ali, S. A., and Nath, P. (2010). Differential expression of the mango alcohol dehydrogenase gene family during ripening. *Phytochemistry* 71, 1485–1494. doi: 10.1016/j.phytochem.2010.05.024

Sivankalyani, V., Sela, N., Feygenberg, O., Zemach, H., Maurer, D., and Alkan, N. (2016). Transcriptome dynamics in mango fruit peel reveals mechanisms of chilling stress. *Front. Plant Sci.* 7:1579. doi: 10.3389/fpls.2016.01579

Slippers, B., Johnson, G. I., Crous, P. W., Coutinho, T. A., Wingfield, B. D., and Wingfield, M. J. (2005). Phylogenetic and morphological re-evaluation of the *Botryosphaeria* species causing diseases of *Mangifera indica*. *Mycologia* 97, 99–110. doi: 10.3852/mycologia.97.1.99

Sudheeran, P. K., Feygenberg, O., Maurer, D., and Alkan, N. (2018). Improved cold tolerance of mango fruit with enhanced anthocyanin and flavonoid contents. *Molecules* 23:1832. doi: 10.3390/molecules23071832

Szcześniak, M. W., Rosikiewicz, W., and Makałowska, I. (2016). CANTATAdb: a collection of plant long non-coding RNAs. *Plant Cell Physiol.* 57, e8–e8.

Tafolla-Arellano, J. C., Zheng, Y., Sun, H., Jiao, C., Ruiz-May, E., Hernández-Oñate, M. A., et al. (2017). Transcriptome analysis of mango (*Mangifera indica* L.) fruit epidermal peel to identify putative cuticle-associated genes. *Sci. Rep.* 7:46163.

Weber, M. J. (2005). New human and mouse microRNA genes found by homology search. *FEBS J.* 272, 59–73. doi: 10.1111/j.1432-1033.2004.04389.x

Wu, G., Park, M. Y., Conway, S. R., Wang, J. W., Weigel, D., and Poethig, R. S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in *Arabidopsis*. *Cell* 138, 750–759. doi: 10.1016/j.cell.2009.06.031

Wu, H. J., Wang, Z. M., Wang, M., and Wang, X. J. (2013). Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants. *Plant Physiol.* 161, 1875–1884. doi: 10.1104/pp.113.215962

Wu, H. X., Jia, H. M., Ma, X. W., Wang, S. B., Yao, Q. S., Xu, W. T., et al. (2014). Transcriptome and proteomic analysis of mango (*Mangifera indica* Linn) fruits. *J. Proteomics* 105, 19–30. doi: 10.1016/j.jprot.2014.03.030

Xin, M., Wang, Y., Yao, Y., Song, N., Hu, Z., Qin, D., et al. (2011). Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biol.* 11:61. doi: 10.1186/1471-2229-11-61

Xue, H., Chen, X., and Li, G. (2007). Involvement of phospholipid signaling in plant growth and hormone effects. *Curr. Opin. Plant Biol.* 10, 483–489. doi: 10.1016/j.pbi.2007.07.003

Yamaguchi, A., Wu, M. F., Yang, L., Wu, G., Poethig, R. S., and Wagner, D. (2009). The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of LEAFY, FRUITFULL, and APETALA1. *Dev. Cell* 17, 268–278. doi: 10.1016/j.devcel.2009.06.007

Yao, F., Zhu, H., Yi, C., Qu, H., and Jiang, Y. (2015). MicroRNAs and targets in senescent litchi fruit during ambient storage and post-cold storage shelf life. *BMC Plant Biol.* 15:181. doi: 10.1186/s12870-015-0509-2

Yoon, J. H., Abdelmohsen, K., and Gorospe, M. (2014). "Functional interactions among microRNAs and long noncoding RNAs," in *In Paper Presented at the Seminars in Cell & Developmental Biology*, Vol. 34, (Cambridge, MA: Academic Press), 9–14. doi: 10.1016/j.semcdb.2014.05.015

Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006a). Conservation and divergence of plant microRNA genes. *Plant J.* 46, 243–259. doi: 10.1111/j.1365-313x.2006.02697.x

Zhang, B., Pan, X., Cobb, G. P., and Anderson, T. A. (2006b). Plant microRNA: a small regulatory molecule with big impact. *Dev. Biol.* 289, 3–16.

Zhang, B., Pan, X., Cox, S., Cobb, G., and Anderson, T. (2006c). Evidence that miRNAs are different from other RNAs. *Cell. Mol. Life Sci. CMLS* 63, 246–254. doi: 10.1007/s00018-005-5467-7

Zhang, B. H., Pan, X. P., Wang, Q. L., George, P. C., and Anderson, T. A. (2005). Identification and characterization of new plant microRNAs using EST analysis. *Cell Res.* 15, 336–360. doi: 10.1038/sj.cr.7290302

Zhang, Y. C., and Chen, Y. Q. (2013). Long noncoding RNAs: new regulators in plant development. *Biochem. Biophys. Res. Commun.* 436, 111–114.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415. doi: 10.1093/nar/gkg595

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.

The reviewer, YO declared a past co-authorship with one of the authors, MC, to the handling editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 14 June 2021
doi: 10.3389/fgene.2021.674534

Genome of the Single Human Chromosome 18 as a "Gold Standard" for Its Transcriptome

Ekaterina Ilgisonis, Nikita Vavilov, Elena Ponomarenko, Andrey Lisitsa, Ekaterina Poverennaya, Victor Zgoda, Sergey Radko and Alexander Archakov*

Institute of Biomedical Chemistry, Moscow, Russia

The cutoff level applied in sequencing analysis varies according to the sequencing technology, sample type, and study purpose, which can largely affect the coverage and reliability of the data obtained. In this study, we aimed to determine the optimal combination of parameters for reliable RNA transcriptome data analysis. Toward this end, we compared the results obtained from different transcriptome analysis platforms (quantitative polymerase chain reaction, Illumina RNASeq, and Oxford Nanopore Technologies MinION) for the transcriptome encoded by human chromosome 18 (Chr 18) using the same sample types (HepG2 cells and liver tissue). A total of 275 protein-coding genes encoded by Chr 18 was taken as the gene set for evaluation. The combination of Illumina RNASeq and MinION nanopore technologies enabled the detection of at least one transcript for each protein-coding gene encoded by Chr 18. This combination also reduced the probability of false-positive detection of low-copy transcripts due to the simultaneous confirmation of the presence of a transcript by the two fundamentally different technologies: short reads essential for reliable detection (Illumina RNASeq) and long-read sequencing data (MinION). The combination of these technologies achieved complete coverage of all 275 protein-coding genes on Chr 18, identifying transcripts with non-zero expression levels. This approach can improve distinguishing the biological and technical reasons for the absence of mRNA detection for a given gene in transcriptomics.

Keywords: proteomics, transcriptomics, threshold, human genome, qPCR, Illumina RNASeq, Oxford Nanopore Technologies MinION

OPEN ACCESS

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

Reviewed by:
Spyros Oikonomopoulos, McGill University and Génome Québec Innovation Centre, Canada
Diana Mechtcheriakova, Medical University of Vienna, Austria
Nikolai Ravin, Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences (RAS), Russia

****Correspondence:***
Ekaterina Ilgisonis
ilgisonis.ev@gmail.com

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

Received: *01 March 2021*
Accepted: *17 May 2021*
Published: *14 June 2021*

Citation:
Ilgisonis E, Vavilov N, Ponomarenko E, Lisitsa A, Poverennaya E, Zgoda V, Radko S and Archakov A (2021) Genome of the Single Human Chromosome 18 as a "Gold Standard" for Its Transcriptome. Front. Genet. 12:674534. doi: 10.3389/fgene.2021.674534

INTRODUCTION

One of the key steps in transcriptome profiling is to determine the criteria for uncovering gene expression; that is, to establish the appropriate threshold for identifying whether or not a gene is expressed. Despite the widespread use of sequencing methods, it is commonly recognized that the choice of threshold (i.e., the cutoff level after which the signal is considered reliable) depends on the specific task being solved, sample type, and technology used (Sha et al., 2015). In particular, different sequencing technologies use different units to measure expression levels, such as reads per kilobase per million (RPKM), transcripts per kilobase per million, fragments per kilobase million, copies per cell, or number of cycles (Bullard et al., 2010).

Regardless of the chosen measurement unit, there is a tendency for an increase in the cutoff level to cause a decrease in the number of registered transcripts, thereby increasing the reliability

of detection (Łabaj and Kreil, 2016; Zhao et al., 2020). This tendency has also been confirmed in targeted polymerase chain reaction (PCR)-based transcriptome mining, in which increasing the number of cycles in droplet digital PCR transcriptome profiling confirmed the presence of transcripts that scored below the cutoff level in the sample (Radko et al., 2019).

However, there is a need for a "gold standard" transcriptome data analysis, which would enable obtaining complete transcriptome coverage of the genome of interest, such as that encoded by a single chromosome. In this study, we sought to establish such a gold standard using human chromosome 18 (Chr 18) as an example. We performed comparative analyses of sequencing from previously published transcriptome datasets (Zgoda et al., 2013; Ponomarenko et al., 2014; Poverennaya et al., 2016; Radko et al., 2019)obtained with three different methods applied to the same sample of biological materials: quantitative PCR (qPCR), Illumina RNASeq (Illumina), and the recently developed nanopore sequencing platform MinION developed by Oxford Nanopore Technologies (ONT) (Jain et al., 2016). ONT can produce long reads of more than10^4 nucleotides, which is an advantage compared with the Illumina platform that produces reads for sequences up to 300 nucleotides in length (Slatko et al., 2018). The disadvantage of ONT is that long reads contain errors at a rate of approximately one lost or misread site per 100 sequenced nucleotides (Amarasinghe et al., 2020). At present, ONT is the only sequencing technology that offers real-time analysis (for rapid insights) in fully scalable formats from the pocket to population scale, which can enable analyses of native DNA or RNA, and can sequence fragments of any length to achieve short to ultra-long read lengths. Transcript sets encoded by 275 protein-coding genes on Chr 18 measured using these three

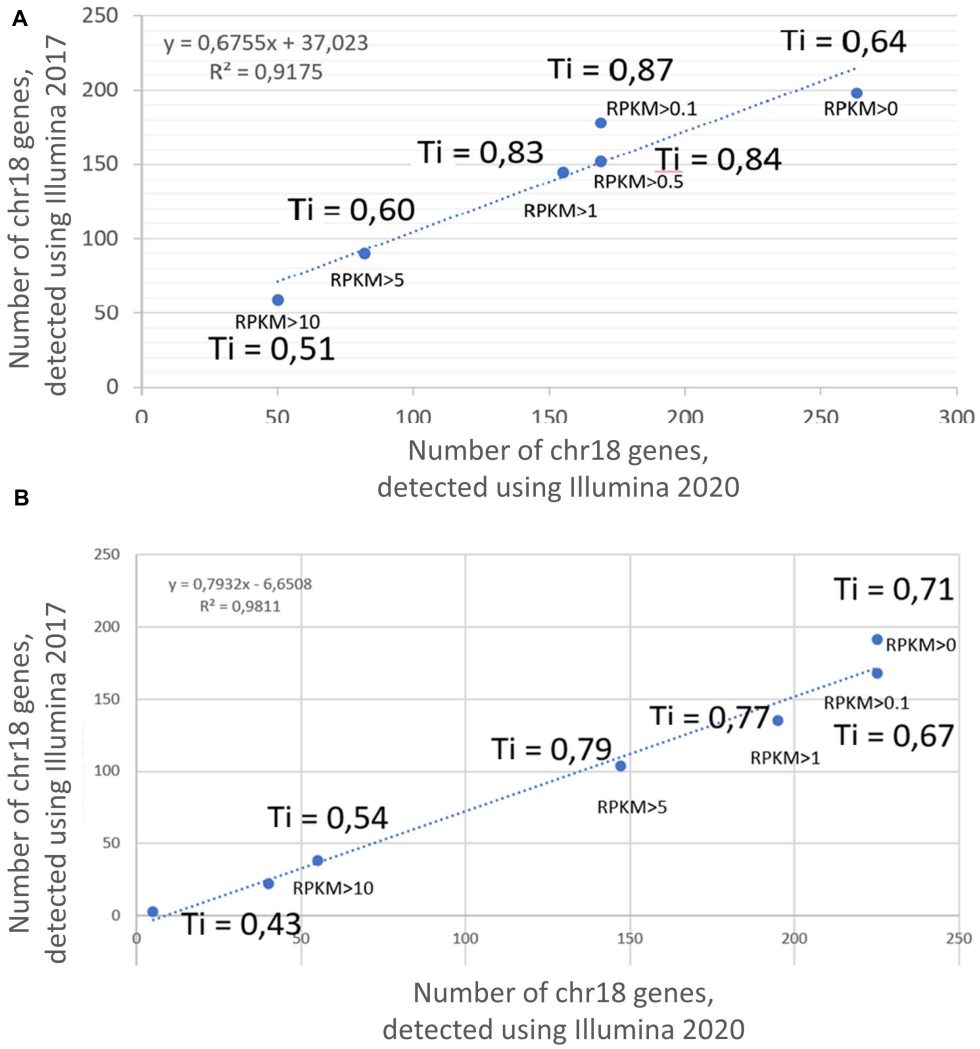


FIGURE 1 | Correlation between the results of transcriptome profiling of the **(A)** HepG2 cell line and **(B)** liver, using the Illumina platform in 2017 and 2020. X axis corresponds to the number of genes, detected using Illumina 2020; Y axis corresponds to the number of genes, detected using Illumina 2017. Ti, Tanimoto index.

independent approaches (qPCR, Illumina, and ONT) in the HepG2 cell line and human liver tissue samples were used for this comparative analysis.

The aim of this study was to establish the optimal technology or combination of technologies for transcriptome analysis based on obtaining the maximum number of detected products at the mRNA level along with complete transcriptome coverage depending on the selected cutoff level for each platform. It is presumed that the lowest possible cutoff level leads to maximum coverage because of the reduction in unreliable results. The confirmation of low-copy transcripts with the three different technologies could therefore be used to judge the reliability of the results obtained. These results can be applied to establishing gold standard approaches for transcriptome analyses of other human chromosomes in the future.

MATERIALS AND METHODS

Data

The results of transcriptome profiling using three technologies (qPCR, RNASeq, and ONT) of Chr 18 genes in the liver tissue and HepG2 cell line obtained by Russian Consortium were analyzed. The details of the samples, sample preparation, and experimental procedures are described in Krasnov et al. (2020). It is necessary to specify, that our study deals only with RNA transcriptome data. Datasets were previously published in the Russian Proteomic Consortium annual reports (Ponomarenko et al., 2014; Poverennaya et al., 2016; Archakov et al., 2019).

Tanimoto Index

Bajusz et al. (2015) demonstrated that the Tanimoto index (Rogers and Tanimoto, 1960) is one of the best measures for assessing similarity, and is now widely used in chemoinformatics and bioinformatics. In particular, they ranked the performances and correlations of eight similarity metrics, which were statistically analyzed using the sum of ranking differences and analysis of variance. They found that the Cosine, Dice, Tanimoto, and Soergel similarity metrics had equivalent high performance, whereas the similarity measures derived from Euclidean and Manhattan distances were far from optimal. Based on this finding, we used the Tanimoto index to estimate the similarity among the results of transcriptomic profiling using the three different technologies.

Specifically, the coefficient of semantic similarity T (**a**, **b**) between two objects **a** and **b** is calculated using the Tanimoto normalization equation (Rogers and Tanimoto, 1960):

$$T(\mathrm{a}, \mathrm{b}) = \frac{|P_{\mathrm{ab}}|}{|P_a| + |P_b| - |P_{\mathrm{ab}}|} \quad (1)$$

where *Pa* indicates the variety of transcripts *a*, *Pb* indicates the variety of transcripts *b*, and *Pab* indicates the variety of transcripts shared in *a* and *b*.

If the Tanimoto index is within 1.0–0.7, it is considered that the two sets are identical, Tanimoto index values from 0.75 to 0.55 indicate that the similarity is much weaker, and values of 0.55 and below indicate that the arrays differ considerably.

Cutoff Level

There is currently no standard guideline for defining the low expression or noise threshold in transcriptomics; therefore, the researchers suggest the approach to determining a threshold

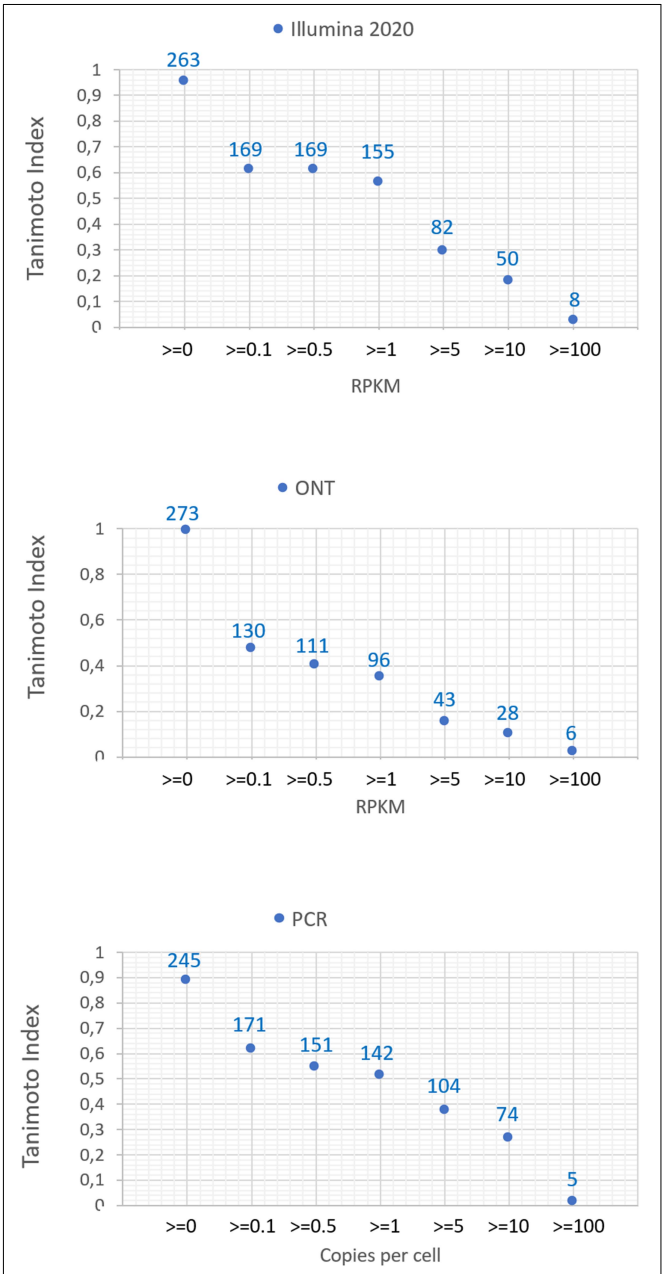


FIGURE 2 | Dependence of the number of detected transcripts of chromosome 18 for various platforms (Illumina 2020 data, ONT, and qPCR) on the cutoff level and the concordance of the results obtained with the known genome of chromosome 18.

for expression above noise: to compare the number of genes expressed at different cutoffs across all samples (Koch et al., 2018). In this work, we used cutoff levels that have been generally recommended in the related literature and compared the number of transcripts obtained depending on the cutoff level. In particular, we applied the following cutoff levels for comparison: 0 (Dall'Agnol et al., 2014), 0.1 (Abdullah et al., 2016), 1 (Xu et al., 2016; Łabaj and Kreil, 2016), 5 (Yang and Chen, 2019), and 10(Wright et al., 2013).

This approach takes into account a variety of factors, including the sequencing depth, batch effects, and technical variability. The resulting cutoff value will not only impact the number of genes to be trimmed from the original dataset but may also affect the interpretation of individual gene expression graphs.

Reliability of the Results

In our work, we proceed from considerations that the more technologies a transcript has been detected, the more reliable its detection is. If a transcript is detected by only one technology, we do not know if this is due to the peculiarities of a particular sequencing technology or a false-positive result. At least two reasons can lead to false positive results. First, the presence of DNA in the RNA preparation. The second is the erroneous mapping of readings to genes due to the read length or the high error rate. Within the framework of this study, we cannot accurately determine the reason for the occurrence of unreliable results, since the main purpose of this study is to compare the results obtained by various technological platforms. Moreover, the lower the abundance of transcript, the less reliable the result is usually considered to be.

RESULTS AND DISCUSSION

Transcriptomic profiling using the Illumina platform (RNASeq) was reported in two different studies by Poverennaya et al. (2017) and by Vavilov et al. (2020). **Figure 1** shows the results obtained in 2017 and 2020 at different RPKM levels, demonstrating 90% correspondence; therefore, only the results obtained in 2020 were used for the comparative analysis among the three technologies in this study.

The Tanimoto index showed a tendency to increase with an increase in the cutoff level, which was a consistent trend for both the HepG2 cell line (**Figure 1A**) and in the liver tissue (**Figure 1B**). The greatest similarity between the transcripts obtained in 2017 and 2020 was found at cutoff levels of >0.1, >0.5, and >1 for the HepG2 cell line and >0.1, >5, and >5 for liver tissue. In addition, the qualitative composition of the transcripts detected by the Illumina platform in 2017 and 2020 at different cutoff levels did not differ significantly, especially observed at RPKM cutoff levels of 0, 0.1, and 1. However, the composition of the arrays at an RPKM cutoff level of >5 differed significantly between years both in the HepG2 cell line and in the liver tissue (Tanimoto index of 0.51 and 0.43, respectively). This discrepancy between the arrays is most likely due to the lifespan of the transcripts and that highly abundant transcripts disintegrate faster, which would lead to differences in transcript detection when samples are analyzed 3 years apart.

The number of common transcripts detected by the different technologies varied depending on the cutoff level. **Figure 2** shows that the largest number of detected transcripts corresponded to a cutoff level of >0. With an increase in the cutoff level

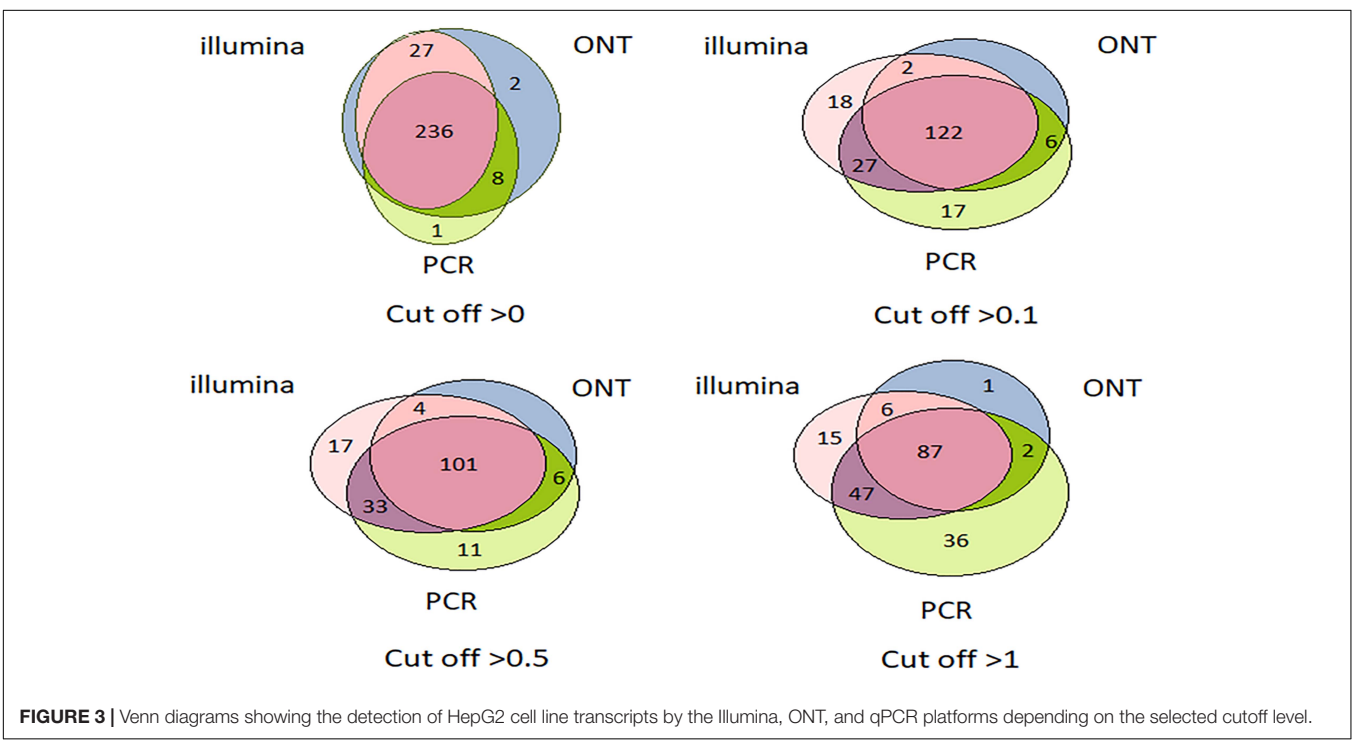


FIGURE 3 | Venn diagrams showing the detection of HepG2 cell line transcripts by the Illumina, ONT, and qPCR platforms depending on the selected cutoff level.

to 0.1, the number of detected transcripts dropped sharply. This may be attributed to noise pollution of the signal in the range from 0 to 0.1.

Regardless of the specific technology used, employing the cutoff level of 0.1 led to a decrease in the number of detected transcripts by 40–60%, and a cutoff level of 0.1 and above led to a decrease in the detected transcripts by 40–50%. The Tanimoto index decreased to 0.6, and then further decreased to 0 at higher cutoffs, indicating that transcripts for most of genes of Chr18 remained unrecorded. This may be due to the contamination of DNA in the RNA preparation or the erroneous mapping of readings to genes. To reveal the most reliable results, the intersection of sets of transcripts obtained by the three different technologies (Illumina, ONT, and qPCR) in the HepG2 cell line and in the liver tissue were compared.

Venn diagrams representing the number of intersecting (common) transcripts according to different cutoff levels for different technologies in the HepG2 cell line and liver tissue are shown in **Figure 3** and **Figure 4**, respectively. In HepG2 cells, a total of 236 transcripts were common to all three technologies, whereas 138 transcripts in the liver tissue were commonly identified; however, the total number of registered transcripts was 273 and 267, respectively. With an increase in the cutoff level to 0.1 and higher, the number of common transcripts obtained with the three platforms decreased sharply, whereas the number of transcripts detected by each platform increased. This increase in the number of intersecting transcripts with a decrease in the cutoff level reflects an increase in the sensitivity of each technology, making it possible to exclude the significant role of unreliable results in the expression of the Chr 18 genome (RPKM $>$ 0), despite the theoretical existence of such a possibility.

The intersection of the results obtained by the three technologies was maximal at the minimum cutoff level ($>$0)

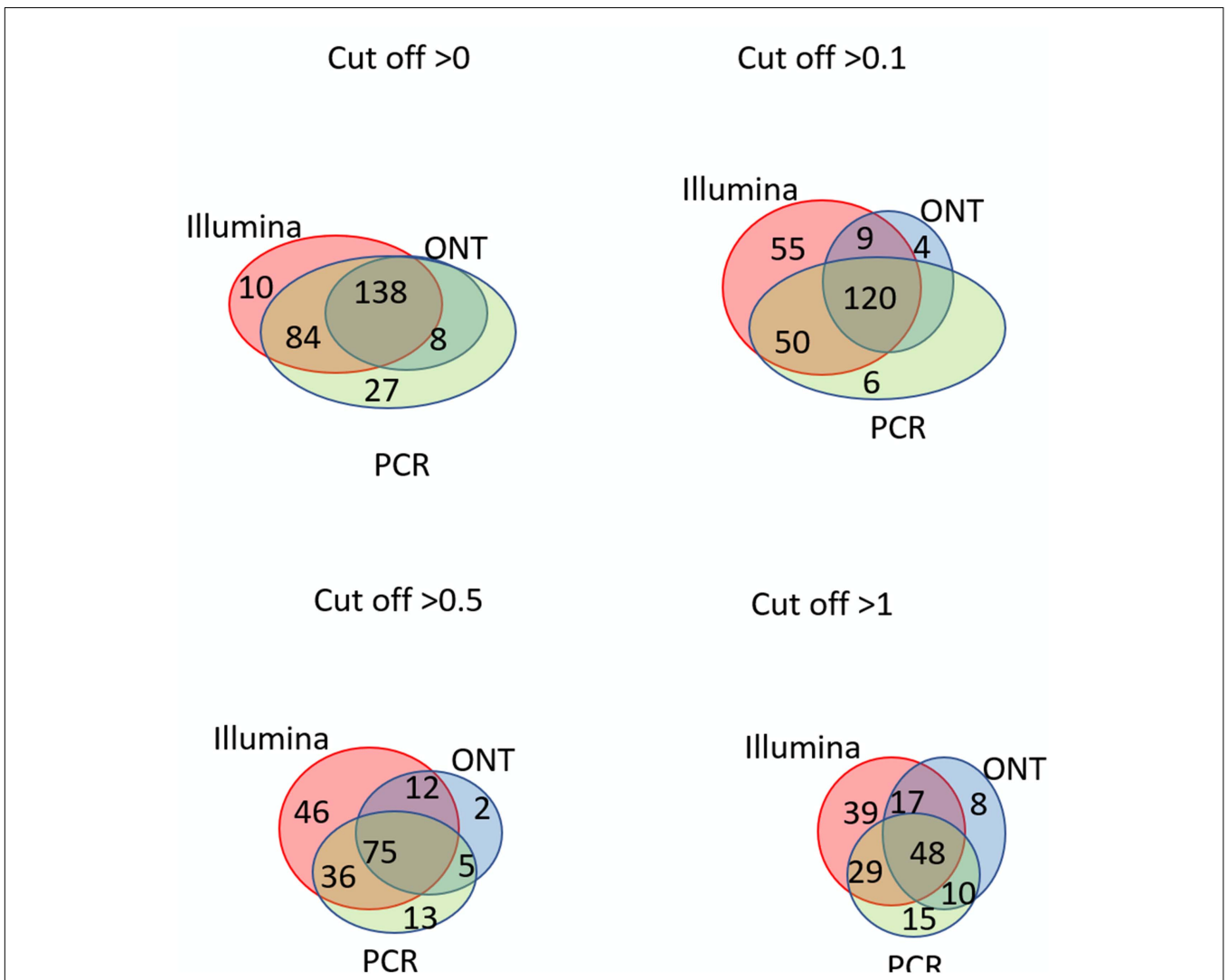


FIGURE 4 | Venn diagrams showing the detection of chromosome 18 transcripts in the liver by the Illumina, ONT, and qPCR platforms depending on the selected cutoff level.

for both the liver and HepG2 cells (**Figures 3**, **4**). Importantly, this shows that applying the same minimal cutoff with different technologies will reveal the same reliable transcripts.

Interestingly, at different cutoff levels, the different technologies showed different patterns of increase in specific transcripts that were detected with only one technology. The maximum increase in the number of transcripts detected by a single technology in the HepG2 cell line was 36, which was obtained using qPCR at a cutoff level ≥ 1, and was 55 using Illumina in the liver tissue. Therefore, different transcripts are detected by different platforms according to variation in sensitivities, highlighting the importance of using several technologies to obtain a reliable transcriptome.

Figures 3, **4** further show that an increase in the cutoff level leads to a decrease in the total portion of transcripts detected by the three technologies. In the HepG2 cell line, at a cutoff level >0, over 236 transcripts were obtained by the three platforms, which represents more than 50% of the Chr 18 genome, and at a cutoff level >1, the number of common transcripts sharply dropped to 48, representing only 20% of the chromosome genome. The same trend was found for the liver tissue.

Transcripts that were not detected by any technology at any cutoff level corresponded to two proteins: Q6ZTR6 and Q9HC47. According to the UniProt database (accession date—02.2021) (Apweiler et al., 2004), these proteins also could not be confirmed (**Figure 5**). Q6ZTR6 is annotated as a "predicted" protein, and Q9HC47 corresponds to cutaneous T-cell lymphoma-associated antigen 1 protein, which is annotated at a PE2 level (protein evidence confirmed at the transcript level). These findings suggested that these missing transcripts did not actually correspond to missing protein detection on these platforms. Ten transcripts were obtained using only ONT technology, which could be considered false positives (**Figure 5**). To assess this possibility, we screened the complete genomes of the liver and HepG2 cell lines obtained from an RNASeq database (accession date—02.2021) (Edgar et al., 2002), demonstrating that these unique transcripts found using ONT technology have no homologous sequence to genes on any other chromosomes besides Chr 18. This suggested that these undetected transcripts are likely the result of extremely underrepresented gene expression on Chr 18 (**Supplementary Material**). Of course, detection of these transcripts could be a results of DNA contamination or wrong mapping of poor quality nanopore reads, but we cannot estimate it in the course of this research.

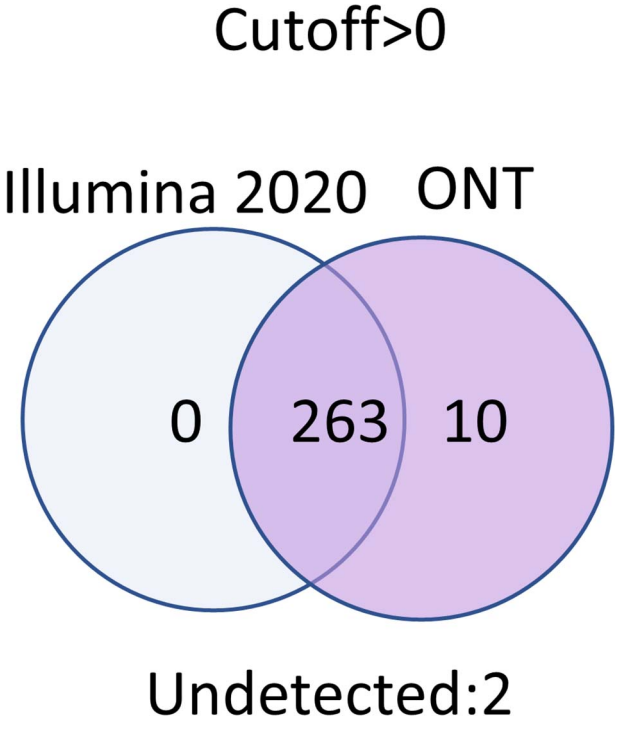


FIGURE 5 | Results obtained using the Illumina and ONT technologies at a cutoff level >0.

Thus, the use of two technologies, Illumina and ONT, enabled the identification of transcripts corresponding to all experimentally observed proteins derived from genes located on human Chr 18, with the exception of two transcripts that were also not confirmed at the protein level in the Nextprot database (accession date—02.2021) (Zahn-Zabal et al., 2020).

CONCLUSION

The greatest coverage of the human genome encoded by Chr 18 was achieved at a cutoff level of >0. Among the three technologies compared (qPCR, Illumina, and ONT), Illumina sequencing and nanopore technology (ONT) complement each other well in terms of non-overlapping common transcripts and detection the complete set of protein-coding genes encoded by the chromosome. In particular, the combined use of Illumina RNASeq and ONT revealed 98–100% of transcripts of the Chr 18 genome at a cutoff level of 0. We also found an expected result that the lowest possible cutoff level leads to maximum coverage due to the lack of unreliable results. However, confirmation of the existence of low-copy transcripts when using all three technologies could further ensure the reliability of the results obtained. This was evidenced by the comparison of the Tanimoto index, which decreased with an increasing cutoff level (**Figure 2**). At a cutoff level of 0.1 and higher, the Tanimoto index was reduced to 0.6 or less, which indicates that under these conditions, the transcriptome obtained would differ significantly from the full Chr 18 exome.

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

EI: manuscript draft. NV: analysis and interpretation of data. ElP: project administration. AL: critical revision. EkP: analysis and

interpretation of data. VZ: critical revision. SR: acquisition of data. AA: study conception and design. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Russian Science Foundation (RSF Grant 20-15-00410; http://www.rscf.ru/). The authors are grateful to the "Human Proteome" Core Facility, Institute of Biomedical Chemistry (IBMC) for performing data processing.

ACKNOWLEDGMENTS

We would like to thank editage (www.editage.com) for english language editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.674534/full#supplementary-material

REFERENCES

Abdullah, H. M., Akbari, P., Paulose, B., Schnell, D., Qi, W., Park, Y., et al. (2016). Transcriptome profiling of *Camelina sativa* to identify genes involved in triacylglycerol biosynthesis and accumulation in the developing seeds. *Biotechnol. Biofuels* 9:136. doi: 10.1186/s13068-016-0555-5

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* 21:30. doi: 10.1186/s13059-020-1935-5

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 32, D115–D119. doi: 10.1093/nar/gkh131

Archakov, A. I., Aseev, A. L., Bykov, V. A., Grigoriev, A. I., Govorun, V. M., Ilgisonis, E. V., et al. (2019). Challenges of the human proteome project: 10-year experience of the Russian Consortium. *J. Proteome Res.* 18, 4206–4214. doi: 10.1021/acs.jproteome.9b00358

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why Is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* 7:20. doi: 10.1186/s13321-015-0069-3

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in MRNA-seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Dall'Agnol, H. P. M. B., Baraúna, R. A., de, P. H. C. G., Sá, R. T.J. Ramos, Nóbrega, F., Nunes, C. I. P., et al. (2014). Omics profiles used to evaluate the gene expression of *Exiguobacterium antarcticum* B7 during cold adaptation. *BMC Genomics* 15:986. doi: 10.1186/1471-2164-15-986

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Jain, M., Olsen, H. E., Paten, B., and Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* 17, 1–11. doi: 10.1186/s13059-016-1103-0

Koch, C. M., Chiu, S. F., Akbarpour, M., Bharat, A., Ridge, K. M., Bartom, E. T., et al. (2018). A beginner's guide to analysis of RNA sequencing data. *Am. J. Respir. Cell Mol. Biol.* 59, 145–157. doi: 10.1165/rcmb.2017-0430TR

Krasnov, G. S., Radko, S. P., Ptitsyn, K. G., Shapovalova, V. V., Timoshenko, O. S., Khmeleva, S. A., et al. (2020). Human Chr18: 'Stakhanovite' genes, missing and UPE1 proteins in liver tissue and HepG2 cells. *BioRxiv* [Preprint]. doi: 10.1101/2020.11.04.358739

Łabaj, P. P., and Kreil, D. P. (2016). Sensitivity, specificity, and reproducibility of RNA-seq differential expression calls. *Biol. Direct* 11:66. doi: 10.1186/s13062-016-0169-7

Ponomarenko, E. A., Kopylov, A. T., Lisitsa, A. V., Radko, S. P., Kiseleva, Y. Y., Kurbatov, L. K., et al. (2014). Chromosome 18 transcriptoproteome of liver tissue and HepG2 cells and targeted proteome mapping in depleted plasma: update 2013. *J. Proteome Res.* 13, 183–190. doi: 10.1021/pr400883x

Poverennaya, E. V., Ilgisonis, E. V., Ponomarenko, E. A., Kopylov, A. T., Zgoda, V. G., Radko, S. P., et al. (2017). Why are the correlations between MRNA and protein levels so low among the 275 predicted protein-coding genes on human chromosome 18. *J. Proteome Res.* 16, 4311–4318. doi: 10.1021/acs.jproteome.7b00348

Poverennaya, E. V., Kopylov, A. T., Ponomarenko, E. A., Ilgisonis, E. V., Zgoda, V. G., Tikhonova, O. V., et al. (2016). State of the art of chromosome 18-Centric HPP in 2016: transcriptome and proteome profiling of liver tissue and HepG2 cells. *J. Proteome Res.* 15, 4030–4038. doi: 10.1021/acs.jproteome.6b00380

Radko, S. P., Poverennaya, E. V., Kurbatov, L. K., Ponomarenko, E. A., Lisitsa, A. V., and Archakov, A. I. (2019). The 'Missing' proteome: undetected proteins, not-translated transcripts, and untranscribed genes. *J. Proteome Res.* 18, 4273–4276. doi: 10.1021/acs.jproteome.9b00383

Rogers, D. J., and Tanimoto, T. T. (1960). A Computer program for classifying plants. *Science* 132, 1115–1118.

Sha, Y., Phan, J. H., and Wang, M. D. (2015). "Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2015-Novem:6461–64* (New York, NY: Institute of Electrical and Electronics Engineers Inc), doi: 10.1109/EMBC.2015.7319872

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of next generation sequencing technologies (and bioinformatics) in cancer. *Curr. Protoc. Mol. Biol.* 122:e59. doi: 10.1002/cpmb.59

Vavilov, N. E., Zgoda, V. G., Tikhonova, O. V., Farafonova, T. E., Shushkova, N. A., Novikova, S. E., et al. (2020). Proteomic analysis of Chr 18 proteins using 2D fractionation. *J. Proteome Res.* 19, 4901–4906. doi: 10.1021/acs.jproteome.0c00856

Wright, H. L., Thomas, H. B., Moots, R. J., and Edwards, S. W. (2013). RNA-seq reveals activation of both common and cytokine-specific pathways following neutrophil priming. *PLoS One* 8:e58598. doi: 10.1371/journal.pone.0058598

Xu, J., Gong, B., Wu, L., Thakkar, S., Hong, H., and Tong, W. (2016). Comprehensive assessments of RNA-seq by the SEQC consortium: FDA-led efforts advance precision medicine. *Pharmaceutics* 8:8. doi: 10.3390/pharmaceutics8010008

Yang, J. R., and Chen, X. (2019). Dosage sensitivity of X-linked genes in human embryonic single cells. *BMC Genomics* 20:42. doi: 10.1186/s12864-019-5432-8

Zahn-Zabal, M., Michel, P. A., Gateau, A., Nikitin, F., Schaeffer, M., Audot, E., et al. (2020). The neXtProt knowledgebase in 2020: data, tools and usability improvements. *Nucleic Acids Res.* 48, D328–D334. doi: 10.1093/nar/gkz995

Zgoda, V. G., Kopylov, A. T., Tikhonova, O. V., Moisa, A. A., Pyndyk, N. V., Farafonova, T. E., et al. (2013). Chromosome 18 transcriptome profiling and targeted proteome mapping in depleted plasma, liver tissue and HepG2 cells. *J. Proteome Res.* 12, 123–134. doi: 10.1021/pr300821n

Zhao, S., Ye, Z., and Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA* 26, 903–909. doi: 10.1261/RNA.074922.120

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.

frontiers in Genetics

DATA REPORT
published: 17 June 2021
doi: 10.3389/fgene.2021.662928

Nanopore and Illumina Genome Sequencing of *Fusarium oxysporum* f. sp. *lini* Strains of Different Virulence

Ekaterina M. Dvorianinova[1,2†], Elena N. Pushkova[1†], Roman O. Novakovskiy[1†], Liubov V. Povkhova[1,2], Nadezhda L. Bolsheva[1], Ludmila P. Kudryavtseva[3], Tatiana A. Rozhmina[3], Nataliya V. Melnikova[1] and Alexey A. Dmitriev[1]*

[1] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, [2] Moscow Institute of Physics and Technology, Moscow, Russia, [3] Federal Research Center for Bast Fiber Crops, Torzhok, Russia

OPEN ACCESS

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

Reviewed by:
Izabela Makałowska,
Adam Mickiewicz University, Poland
Vladimir Zhukov,
All-Russian Research Institute of Agricultural Microbiology of the Russian Academy of Agricultural Sciences, Russia

***Correspondence:**
Alexey A. Dmitriev
alex_245@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: 01 February 2021
Accepted: 06 May 2021
Published: 17 June 2021

Citation:
Dvorianinova EM, Pushkova EN, Novakovskiy RO, Povkhova LV, Bolsheva NL, Kudryavtseva LP, Rozhmina TA, Melnikova NV and Dmitriev AA (2021) Nanopore and Illumina Genome Sequencing of Fusarium oxysporum f. sp. lini Strains of Different Virulence. Front. Genet. 12:662928. doi: 10.3389/fgene.2021.662928

Keywords: *Fusarium oxysporum*, flax pathogen, different virulence, genome sequencing, nanopore sequencing, *de novo* genome assembly

INTRODUCTION

Fusarium oxysporum f. sp. *lini* is the one of notoriously known pathogens of flax (*Linum usitatissimum* L.), causing wilt. In the case of young seedlings, the disease may lead to complete yield losses (Kommedahl et al., 1970). At the same time, flax is an important crop used for manufacturing oil of high nutritional value, food supplements, industrial products, and fiber (Jhala and Hall, 2010; Czemplik et al., 2011; Kezimana et al., 2018; Parikh and Pierce, 2019). However, flax production is dependent on the resistance of flax varieties to phytopathogens, whereas *F. oxysporum* demonstrates considerable genetic diversity (Edel et al., 2001; Michielse and Rep, 2009).

At present, phytopathogenic strains of *F. oxysporum* are classified into numerous *formae speciales* according to their ability to colonize different hosts (Armstrong and Armstrong, 1981). Unfortunately, this system cannot provide researchers with enough information on a type of the pathogen and the severity of the infection it causes (Edel-Hermann and Lecomte, 2019). Like other representatives of the species, *forma specialis lini* is morphologically and genetically heterogeneous. Strains of the pathogen differ in such characteristics as the type of sporulation, conidia formation, pigment production, and the rate of growth on different media. Moreover, the degree of pathogenicity varies within the *forma* and depends on the infected crop variety (Kommedahl et al., 1970).

For molecular classification of *F. oxysporum*, numerous markers were used (Baayen et al., 2000; Lievens et al., 2008; Baysal et al., 2010; Sharma et al., 2014; van Dam et al., 2018; Srinivas et al., 2019; Sasseron et al., 2020), and genes associated with virulence were considered as targets for molecular discrimination of strains of the fungus (Lievens et al., 2008; van Dam et al., 2018). It was shown that *secreted in xylem* (*SIX*) genes are associated with pathogenicity of *F. oxysporum*, and the majority of them are distributed within the sequence of one chromosome (Rep et al., 2004; Houterman et al., 2007; Kashiwa et al., 2017; Carvalhais et al., 2019). Importantly, the combination of these genes differs between and within *formae* (Lievens et al., 2009). *Secreted in xylem* are also responsible for the resistance of cultivars to certain pathogen races due to gene-for-gene interaction between *SIX* genes of a pathogen and R genes (resistance genes) of a plant. Breaking the resistance of plant lineages to wild-type *F. oxysporum* could be accomplished by deletion of a certain *SIX* gene, which is recognized by the immune system of a plant (Houterman et al., 2008). However, in the case of *F. oxysporum* f. sp. *lini*, the connection between the degree of pathogenicity of a strain and the set of its *SIX* genes is not investigated deeply, and gaining insights into the mechanisms of pathogenicity offers an opportunity for effective disease control.

Besides, *F. oxysporum* representatives differ not only in the content of genes and their sequences but also in the number of chromosomes due to chromosomal rearrangements and the mobility of lineage-specific chromosomes (Davière et al., 2001; Ma et al., 2010; Schmidt et al., 2013; Vlaardingerbroek et al., 2016; Wang et al., 2020). Having learned the statistics of deposited *F. oxysporum* assemblies, one may conclude that genome sizes vary from about 50 to 70 Mb and differ between many *formae* (data of the NCBI Genome database, https://www.ncbi.nlm.nih.gov/genome/browse/#!/eukaryotes/707/).

Summing up, *F. oxysporum* f. sp. *lini* appears to be heterogeneous. However, there is a lack of next-generation sequencing data concerning the genome structure of the flax pathogen and the molecular basis of pathogenicity in relation to diverse virulence of the strains. In this work, we have chosen six strains of *F. oxysporum* f. sp. *lini* of low, medium, and high virulence (two strains per each degree of pathogenicity), performed genome sequencing on Oxford Nanopore Technologies (ONT) and Illumina platforms, and obtained *de novo* assemblies of the sequenced strains.

MATERIALS AND METHODS

Fungal Material

F. oxysporum f. sp. *lini* samples were provided by the Institute for Flax (Torzhok, Russia). The strains were of the following pathogenicity degrees: the low (strains #456, #482), the medium (#476, #525), and the high one (#483, #39). Mycelium was grown in test tubes on potato dextrose agar medium for 3 weeks (Alpha Biosciences, USA).

DNA Extraction and Purification

Following the previously developed protocol, pure high-molecular-weight DNA of the fungal samples was obtained (Krasnov et al., 2020). In brief, the DNA was extracted using the CTAB method followed by its isolation with Blood and Cell Culture DNA Mini Kit (Qiagen, USA). The quality and concentration of the DNA were evaluated on a NanoDrop 2000C spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit 2.0 fluorometer (Life Technologies, USA). The assessment of DNA length and the control of RNA absence were performed by electrophoresis in a 0.8% agarose gel (Lonza, Switzerland).

DNA Library Preparation and Sequencing on the Oxford Nanopore Technologies Platform

Before library preparation, DNA fragments up to 10 kb were removed from the samples with a Short Read Eliminator Kit (Circulomics, USA), and the remaining DNA was purified with AMPure XP beads (Beckman Coulter, USA) in a ratio of 1:0.7 (sample:beads). SQK-LSK109 Ligation Sequencing Kit (ONT, UK) for 1D genomic DNA sequencing was used to prepare the library. During this procedure, minor modifications were introduced to the recommended protocol that included sample barcoding with the EXP-NBD103 (Native Barcoding Expansion) kit (ONT). Namely, at the steps of DNA recovery at 20°C and ligation the time of incubation was increased to 20 and 60 min respectively. Sequencing was performed on a MinION (ONT) instrument with a FLO-MIN-106D (R9.4.1) flow cell (ONT).

DNA Library Preparation and Sequencing on the Illumina Platform

Upon DNA shearing on a S220 ultrasonic homogenizer (Covaris, USA), the library was prepared from 1 μg of fragmented DNA with the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, UK) according to the manufacturer's protocol with size selection of adaptor-ligated DNA of about 600–800 bp. A 2100 Bioanalyzer instrument (Agilent Technologies, USA) and a Qubit 2.0 fluorometer (Life Technologies) were used to evaluate the quality and concentration of the DNA library respectively. The DNA library was sequenced on a MiSeq instrument (Illumina, USA) with a read length of 300+300 bp.

Preliminary Data Analysis

On genome sequencing of five strains (#456, #476, #482, #483, #525), we obtained 2.7 Gb of ONT data (374–756 Mb per strain, N50 = 32–44 kb) and 26.7 million Illumina paired-end reads (2.9–7.9 million 300+300 bp reads per strain). Since *F. oxysporum* f. sp. *lini* genome size is about 60–70 Mb (Kanapin et al., 2020; Krasnov et al., 2020), the obtained data corresponded to 10x genome coverage with ONT reads and 50x coverage with Illumina reads on average. The Nanopore and Illumina sequencing data for these five strains and isolate #39, which was sequenced by us earlier, were deposited in NCBI under the BioProject accession number PRJNA721899. For further use in performing *de novo* genome assemblies of the five sequenced strains, ONT fast5 files were basecalled with Guppy 3.6.1 (https://community.nanoporetech.com/protocols/Guppy-protocol/v/GPB_2003_v1_revU_14Dec2018) using the dna_r9.4.1_450bps_hac.cfg config file. Adapter removal and demultiplexing were carried out with Porechop 0.2.4 (https://github.com/rrwick/Porechop). Reads with an average quality below 6 ($Q < 6$) were discarded with Trimmomatic 0.39 (Bolger et al., 2014). To perform the initial genome assemblies, we chose the Canu tool (version 2.1) developed for long-read datasets (Koren et al., 2017), as it provided us with the best results when assembling the genome of isolate #39 in our previous study (Krasnov et al., 2020). The approximate size of the genomes was set as 60 Mb, and the other Canu parameters were kept default. The resulting quality of the assemblies was judged by QUAST features (version 5.0.2) (Mikheenko et al., 2018) and BUSCO completeness (version 4.1.2, hypocreales_odb10 dataset) (Seppey et al., 2019). BUSCO and QUAST statistics for the Canu genome assemblies of the five strains were the following: completeness laid in a range of 61.9–94.5%, N50 was 0.2–2.0 Mb, L50 varied between 9 and 82 (**Table 1**, Canu). These assemblies were not good enough for further analysis that can be explained by the low coverage of genomes with ONT reads (about 6–13x). So, we decided to perform hybrid assemblies of the genomes of the five strains with MaSuRCA (version 3.4.2, CA algorithm) (Zimin et al., 2013) using both ONT and Illumina data, as the obtained coverage with Illumina reads was high (30–80x) and MaSuRCA showed decent results for isolate #39 earlier (Krasnov et al., 2020). The N50 parameter of the hybrid assemblies of

TABLE 1 | QUAST and BUSCO statistics for Canu and MaSuRCA genome assemblies of six *F. oxysporum* f. sp. *lini* strains.

Strain number		**456**	**482**	**476**	**525**	**483**	**39**
Pathogenicity		**low**		**medium**		**high**	
ONT data volume, Gb*		0.38	0.76	0.59	0.37	0.56	4.87
Canu assemblies							
QUAST	Number of contigs	338	53	191	323	226	36
statistics	Total length, Mb	51.4	51.1	64.0	49.2	63.4	69.5
	N50, kb	204	2,036	634	193	436	4,152
	L50	80	9	35	82	41	6
	GC, %	48.05	47.70	47.95	48.28	47.95	47.99
BUSCO	Complete, %	68.0	94.5	85.6	61.9	86.4	97.8
statistics	Single, %	67.4	93.9	84.8	61.3	85.6	96.8
	Duplicated, %	0.6	0.6	0.8	0.6	0.8	1.0
MaSuRCA assemblies							
QUAST	Number of contigs	113	43	67	123	97	74
statistics	Total length, Mb	61.7	53.4	65.9	63.1	63.8	69.3
	N50, kb	1,063	2,363	2,485	961	1,594	2,336
	L50	17	7	9	21	11	5
	GC, %	47.86	47.30	47.87	48.11	47.98	48.07
BUSCO	Complete, %	96.5	99.7	99.1	98.2	98.4	99.7
statistics	Single, %	95.0	98.0	97.5	97.2	97.2	97.9
	Duplicated, %	1.5	1.7	1.6	1.0	1.2	1.8

**The volume of ONT data after basecalling with Guppy, adapter trimming with Porechop, and quality filtration with Trimmomatic is presented.*

strains #456, #476, #482, #483, and #525 laid in a range of 1.0–2.5 Mb, L50 was 7–21, and the BUSCO completeness varied from 96.5 to 99.7% (**Table 1**, MaSuRCA). These statistics indicate fairly high contiguity and completeness of the obtained MaSuRCA assemblies. They were deposited in NCBI under the BioProject accession number PRJNA721899. Besides, we tested if ONT reads were crucial for assembly quality and assembled the genome of strain #525 from only Illumina reads using MaSuRCA, as the smallest amount of ONT data was obtained for this strain (0.37 Mb). It appeared that even 5–10x coverage of a genome with long ONT reads significantly improved assembly contiguity—N50 = 961 kb and L50 = 21 for the hybrid assembly against N50 = 92 kb and L50 = 145 for the assembly from Illumina reads only. Each of the received assemblies of the five strains contained a circular contig representing a complete mitochondrial genome which can be used in phylogenetic studies of *Fusarium* species. The obtained assemblies are a useful resource for comparative genomic studies of the flax pathogen strains.

The comparison of the *F. oxysporum* f. sp. *lini* strains studied in the present work was also performed based on Illumina reads mapped to reference genomes. We trimmed Illumina reads of the six strains with Trimmomatic 0.39 (trailing:28) and discarded the short ones (minlen:50); the remaining data were mapped to reference assemblies of *F. oxysporum* using Bowtie 2 (version 2.3.5.1) (Langmead et al., 2019), which also reported the overall alignment rate, and the alignments were sorted with samtools (version 1.10) (Li et al., 2009). Samtools was also used to determine the coverage of contigs by Illumina reads in each assembly, and the overall coverage was calculated as the sum of covered bases in each contig divided by the number of all bases in an assembly. The only two complete *F. oxysporum* genomes deposited in GenBank were downloaded to be used for mapping—the assembly of strain Fo5176 (GCA_013112355.1, https://www.ncbi.nlm.nih.gov/genome/707?genome_assembly_id=1472496) by Fokkens et al. (2020) and the assembly of strain Fo47 (GCA_013085055.1, https://www.ncbi.nlm.nih.gov/genome/707?genome_assembly_id=910449) by Wang et al. (2020). Besides, the genome of *F. oxysporum* f. sp. *lini* isolate #39 was reassembled with Canu, using the earlier obtained data (Krasnov et al., 2020) and according to the same strategy as described above (ONT read processing with Guppy, Porechop, Trimmomatic, assembling with Canu), and with MaSuRCA (basecalling with Guppy, assembling from ONT and Illumina reads with MaSuRCA) (**Table 1**). Polishing of the Canu assembly was conducted using ONT reads according to the previously optimized scheme—two iterations with the Racon polisher and one with Medaka (Dmitriev et al., 2020). The accuracy of the Canu assembly was also improved by polishing with POLCA (from MaSuRCA) using Illumina reads (Zimin and Salzberg, 2020). The MaSuRCA assembly does not need additional polishing (Zimin et al., 2017). The resulting Canu assembly consisted of 41 contigs with N50 of 4.15 Mb, had a length of 69.5 Mb, and completeness of 99.8%. As expected, the polished Canu assembly outperformed the MaSuRCA one in terms of QUAST and BUSCO statistics, so the Canu-assembled and polished

TABLE 2 | Percentages of aligned Illumina reads against the assembly of *F. oxysporum* f. sp. *lini* isolate #39 and the complete genomes of *F. oxysporum* strains Fo47 and Fo5176 and percentages of the genomes covered by Illumina reads for six *F. oxysporum* f. sp. *lini* strains.

Strain number		**456**	**482**	**476**	**525**	**483**	**39**
Pathogenicity		**low**		**medium**		**high**	
Illumina data volume, Gb		4.76	3.22	3.32	3.00	1.72	1.78
Genome to be aligned against	**Genome size, Mb**	**Percentage of aligned reads, %**					
F. oxysporum f. sp. *lini* #39	69.5	82.85	78.48	83.45	93.17	83.99	95.03
F. oxysporum Fo5176 (GCA_013112355.1)	68.0	74.30	76.72	74.44	78.07	75.23	77.09
F. oxysporum Fo47 (GCA_013085055.1)	50.4	63.75	75.11	63.79	63.11	65.35	62.87
Genome to be covered	**Genome size, Mb**	**Percentage of a genome covered, %**					
F. oxysporum f. sp. *lini* #39	69.5	89.29	71.40	95.97	95.51	88.34	99.34
F. oxysporum Fo5176 (GCA_013112355.1)	68.0	82.74	71.02	85.75	83.26	80.94	95.84
F. oxysporum Fo47 (GCA_013085055.1)	50.4	88.23	85.61	89.90	84.69	87.82	83.54

genome of *F. oxysporum* f. sp. *lini* isolate #39 was used for further mapping.

Illumina reads of the six differently virulent strains of *F. oxysporum* f. sp. *lini* were separately mapped to the two complete genomes of strains Fo5176 and Fo47 and also the assembly of isolate #39 obtained anew (**Table 2**). The overall alignment of the mapped reads corresponding to the six strains against the genomes of isolate #39 and strains Fo5176 and Fo47 as well as the percentages of the genome covered were higher for isolate #39 than for Fo5176 and Fo47. Fo5176 infects *Arabidopsis thaliana* (Fokkens et al., 2020) and has a comparable to isolate #39 genome size (68.0 and 69.5 Mb respectively), while Fo47 is classified as an endophyte and has a significantly smaller genome size−50.4 Mb (Wang et al., 2020). The percentages of aligned reads and covered genome fractions varied between *F. oxysporum* f. sp. *lini* strains; however, we did not reveal the similarity between genomes of the strains with the same pathogenicity degree. For example, high-virulent isolate #39 had less resemblance to another high-virulent strain #483 than to strains #476 and #525 with medium virulence. At the same time, low-virulent strain #482 had the most significant difference from isolate #39. Thus, mapping our Illumina sequencing data of differently virulent strains enabled us to evaluate broad-scale differences in genomes. Besides, using the present Illumina data, one can also reveal single nucleotide polymorphisms (SNPs) or small insertions/deletions in the loci of interest and perform a comparison of the *F. oxysporum* f. sp. *lini* strains with *F. oxysporum* strains of other *formae speciales*. At the same time, the genomes of the six studied strains, even of those ones with similar virulence, differed in the amount of mapped Illumina data, and, therefore, mapping Illumina reads to a reference genome can result in the loss of significant part of information. Besides, chromosomal rearrangements may be implicated in gene regulation, and, for evaluation of the role of such variations in fungal pathogenicity, *de novo* genome assemblies obtained from long reads are indispensable (Demené et al., 2021). Thus, both Illumina and ONT data obtained by us for *F. oxysporum* f. sp. *lini* strains of different virulence are essential for a comprehensive comparative analysis of the genomes and identification of genetic differences associated with pathogenicity.

Another example of the use of the obtained dataset is the identification of genes of interest and their further analysis. For instance, we performed the search for genes involved in a process of plant colonization. The blastn analysis showed that the obtained assemblies of all the strains but low-virulent #482 contained regions with high homology (*e*-value $< 10^{-10}$, identity of 99% on average) to partial coding sequences of virulence genes found in *forma specialis lini*: *SIX1* (NCBI: KM893920.1), *SIX7* (KM893928.1), *SIX10* (KP964982.1), *SIX12* (KP964992.1), *SIX13* (KP964998.1) (Laurence et al., 2015; Taylor et al., 2016). The identified *SIX* sequences were found arranged in clusters. For example, in the genome of isolate #39, partial coding sequences of *SIX1*, *SIX7*, *SIX10*, *SIX13* mapped completely against the locus of 400 kb of tig00000022, and *SIX7* mapped twice against this locus. Besides, homologs of *SIX7*, *SIX12* sequences were present in tig00000001, and those of *SIX7*, *SIX10*, *SIX12*, *SIX13* were found in a 100 kb locus of tig00000029. Similar clustering of *SIX* sequences was observed in the assemblies of strains #456, #476, #483, and #525. Thus, the obtained data are valuable for identifying genes of a particular family, including those responsible for pathogenicity and playing an important role in the interaction between *F. oxysporum* and plants.

One more illustration of a way to use our dataset is the study of the role of DNA modifications in the regulation of *F. oxysporum* genome. DNA methylation is implicated in gene expression regulation, repression of transposable elements, and chromatin remodeling. To date, it is known that cytosines in fungi can be methylated within the context of CpG sites, CN (N for any nucleotide) pairs, and long clusters of cytosines;

the type of the most methylated motif varies between fungal species, while the overall level of methylation in fungal genomes is relatively low (He et al., 2020). Several approaches of the ONT data use for the evaluation of methylation level throughout the genome of an object were developed (Xu and Seki, 2020; Tourancheau et al., 2021). We assessed DNA methylation levels in the genome of isolate #39, for which high coverage with ONT reads was obtained (about 70x), using the nanopolish tool (https://nanopolish.readthedocs.io/en/latest/quickstart_call_methylation.html), which is trained for methylation evaluation within the CpG context. The basecalled ONT reads were mapped against the assembly of isolate #39 with minimap2, alignments were derived using samtools, and methylation levels of CpG sites were estimated with nanopolish. Methylation across the whole genome was low (**Supplementary Data 1**) that is in concordance with other studies on DNA methylation in fungi (Bewick et al., 2019). Nevertheless, in the assembly of *F. oxysporum* f. sp. *lini* isolate #39, more than 500 CpG sites (with coverage >20 ONT reads) had methylation levels ≥0.5. Then, due to poor annotation of *F. oxysporum* genomes, we performed a blast-search for the regions containing these CpG sites in fungi taxa. Most of the methylated CpG sites were not assigned to specific functional elements of the genome; however, there were also CpG sites that were located in promoter regions, transposons, and gene bodies. Therefore, our dataset is useful for the evaluation of the role of DNA methylation in genome regulation of *F. oxysporum* f. sp. *lini*.

Finally, the obtained genome assemblies of *F. oxysporum* f. sp. *lini* strains can be used in gene expression studies, for example, as a reference for transcriptome assembly with further expression analysis or in selection of conservative regions suitable for primer design for the assessment of gene expression by quantitative PCR in different *F. oxysporum* strains or even different *Fusarium* species.

CONCLUSIONS

This work mainly focused on genome sequencing of strains of the flax pathogen *F. oxysporum* f. sp. *lini*, possessing diverse pathogenicity degrees, on two platforms—ONT and Illumina. The collected data allowed us to assemble the genomes of five strains and reassemble the genome of isolate #39 (the used data were obtained in our previous work), which lay the basis for further investigation of *F. oxysporum* virulence mechanisms and contribute to understanding the general structure of the pathogen population. Due to *F. oxysporum* f. sp. *lini* includes a vast number of genotypes, it is of high significance to study the origins of pathogenicity at molecular level. Our dataset can be of great use for researchers working on breeding resistant flax varieties and developing methods to prevent the disease and economic losses.

DATA AVAILABILITY STATEMENT

The obtained data can be found in NCBI under the BioProject accession number PRJNA721899.

AUTHOR CONTRIBUTIONS

TR, NM, and AD conceived and designed the work. ED, EP, RN, LP, NB, and LK performed the experiments. ED, RN, TR, NM, and AD analyzed the data. ED, NM, and AD wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work (sequencing and analysis of genomes of *F. oxysporum* f. sp. *lini* strains) was funded by RFBR according to the research project 19-34-90055. The maintenance of flax pathogen collection is carried out under the financial support of the Ministry of Science and Higher Education of the Russian Federation, state assignment number 075-00853-19-00.

ACKNOWLEDGMENTS

We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, EIMB RAS for providing genome sequencing techniques and computing power. This work was performed using the equipment of EIMB RAS Genome center (http://www.eimb.ru/ru1/ckp/ccu_genome_ce.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.662928/full#supplementary-material

Supplementary Data 1 | Methylation levels of CpG sites in the genome assembly of *F. oxysporum* f. sp. *lini* isolate #39. Data derived from Nanopore sequencing data using nanopolish. Sites with at least 20x coverage are presented.

REFERENCES

Armstrong, G. M., and Armstrong, J. K. (1981). "Formae speciales and races of *Fusarium oxysporum* causing wilt diseases," in *Fusarium: Disease, Biology, and Taxonomy*, eds P. E. Nelson, T. A. Toussoun, and R. J. Cook (Pennsylvania, PA: Pennsylvania State University Press), 391–399.

Baayen, R. P., O'Donnell, K., Bonants, P. J., Cigelnik, E., Kroon, L. P., Roebroeck, E. J., et al. (2000). Gene genealogies and AFLP analyses in the *Fusarium oxysporum* complex identify monophyletic and nonmonophyletic formae speciales causing wilt and rot disease. *Phytopathology* 90, 891–900. doi: 10.1094/PHYTO.2000.90.8.891

Baysal, O., Siragusa, M., Gumrukcu, E., Zengin, S., Carimi, F., Sajeva, M., et al. (2010). Molecular characterization of *Fusarium oxysporum* f. *melongenae* by ISSR and RAPD markers on eggplant. *Biochem. Genet.* 48, 524–537. doi: 10.1007/s10528-010-9336-1

Bewick, A. J., Hofmeister, B. T., Powers, R. A., Mondo, S. J., Grigoriev, I. V., James, T. Y., et al. (2019). Diversity of cytosine methylation across the fungal tree of life. *Nat. Ecol. Evol.* 3, 479–490. doi: 10.1038/s41559-019-0810-9

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

Carvalhais, L. C., Henderson, J., Rincon-Florez, V. A., O'Dwyer, C., Czislowski, E., Aitken, E. A. B., et al. (2019). Molecular diagnostics of banana Fusarium wilt targeting secreted-in-xylem genes. *Front. Plant Sci.* 10:547. doi: 10.3389/fpls.2019.00547

Czemplik, M., Boba, A., Kostyn, K., Kulma, A., Mituła, A., Sztajnert, M., et al. (2011). "Flax engineering for biomedical application," in *Biomedical Engineering, Trends, Research and Technologies*, ed S. Olsztynska (London: IntechOpen), 407–434. doi: 10.5772/13570

Davière, J.-M., Langin, T., and Daboussi, M.-J. (2001). Potential role of transposable elements in the rapid reorganization of the *Fusarium oxysporum* genome. *Fungal Genet. Biol.* 34, 177–192. doi: 10.1006/fgbi.2001.1296

Demené, A., Laurent, B., Cros-Arteil, S., Boury, C., and Dutech, C. (2021). Chromosomal rearrangements but no change of genes and transposable elements repertoires in an invasive forest-pathogenic fungus. *bioRxiv* 2021.03.09.434572. doi: 10.1101/2021.03.09.434572

Dmitriev, A. A., Pushkova, E. N., Novakovskiy, R. O., Beniaminov, A. D., Rozhmina, T. A., Zhuchenko, A. A., et al. (2020). Genome sequencing of fiber flax cultivar atlant using oxford nanopore and illumina platforms. *Front. Genet.* 11:590282. doi: 10.3389/fgene.2020.590282

Edel, V., Christian, S., Gautheron, N., Recorbet, G., and Alabouvette, C. (2001). Genetic diversity of *Fusarium oxysporum* populations isolated from different soils in France. *FEMS Microbiol. Ecol.* 36, 61–71. doi: 10.1111/j.1574-6941.2001.tb00826.x

Edel-Hermann, V., and Lecomte, C. (2019). Current status of *Fusarium oxysporum* formae speciales and races. *Phytopathology* 109, 512–530. doi: 10.1094/PHYTO-08-18-0320-RVW

Fokkens, L., Guo, L., Dora, S., Wang, B., Ye, K., Sanchez-Rodriguez, C., et al. (2020). A chromosome-scale genome assembly for the *Fusarium oxysporum* strain Fo5176 to establish a model Arabidopsis-fungal pathosystem. *G3 (Bethesda)* 10, 3549–3555. doi: 10.1534/g3.120.401375

He, C., Zhang, Z., Li, B., and Tian, S. (2020). The pattern and function of DNA methylation in fungal plant pathogens. *Microorganisms* 8:227. doi: 10.3390/microorganisms8020227

Houterman, P. M., Cornelissen, B. J., and Rep, M. (2008). Suppression of plant resistance gene-based immunity by a fungal effector. *PLoS Pathog.* 4:e1000061. doi: 10.1371/journal.ppat.1000061

Houterman, P. M., Speijer, D., Dekker, H. L., CG, D.E. K., Cornelissen, B. J., and Rep, M. (2007). The mixed xylem sap proteome of *Fusarium oxysporum*-infected tomato plants. *Mol. Plant Pathol.* 8, 215–221. doi: 10.1111/j.1364-3703.2007.00384.x

Jhala, A. J., and Hall, L. M. (2010). Flax (*Linum usitatissimum* L.): current uses and future applications. *Aust. J. Basic Appl. Sci.* 4, 4304–4312.

Kanapin, A., Samsonova, A., Rozhmina, T., Bankin, M., Logachev, A., and Samsonova, M. (2020). The genome sequence of five highly pathogenic isolates of *Fusarium oxysporum* f. sp. *lini*. *Mol. Plant Microbe Interact.* 33, 1112–1115. doi: 10.1094/MPMI-05-20-0130-SC

Kashiwa, T., Kozaki, T., Ishii, K., Turgeon, B. G., Teraoka, T., Komatsu, K., et al. (2017). Sequencing of individual chromosomes of plant pathogenic *Fusarium oxysporum*. *Fungal Genet. Biol.* 98, 46–51. doi: 10.1016/j.fgb.2016.12.001

Kezimana, P., Dmitriev, A. A., Kudryavtseva, A. V., Romanova, E. V., and Melnikova, N. V. (2018). Secoisolariciresinol diglucoside of flaxseed and its metabolites: biosynthesis and potential for nutraceuticals. *Front. Genet.* 9:641. doi: 10.3389/fgene.2018.00641

Kommedahl, T., Christensen, J. J., and Frederiksen, R. A. (1970). *A Half Century of Research in Minnesota on Flax Wilt Caused by Fusarium oxysporum*. Minneapolis, MN: University of Minnesota.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27, 722–736. doi: 10.1101/gr.215087.116

Krasnov, G. S., Pushkova, E. N., Novakovskiy, R. O., Kudryavtseva, L. P., Rozhmina, T. A., Dvorianinova, E. M., et al. (2020). High-quality genome assembly of *Fusarium oxysporum* f. sp. *lini*. *Front. Genet.* 11:959. doi: 10.3389/fgene.2020.00959

Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* 35, 421–432. doi: 10.1093/bioinformatics/bty648

Laurence, M. H., Summerell, B. A., and Liew, E. C. Y. (2015). *Fusarium oxysporum* f. sp. *canariensis*: evidence for horizontal gene transfer of putative pathogenicity genes. *Plant Pathol.* 64, 1068–1075. doi: 10.1111/ppa.12350

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

Lievens, B., Houterman, P. M., and Rep, M. (2009). Effector gene screening allows unambiguous identification of *Fusarium oxysporum* f. sp. *lycopersici* races and discrimination from other formae speciales. *FEMS Microbiol. Lett.* 300, 201–215. doi: 10.1111/j.1574-6968.2009.01783.x

Lievens, B., Rep, M., and Thomma, B. P. (2008). Recent developments in the molecular discrimination of formae speciales of *Fusarium oxysporum*. *Pest Manag. Sci.* 64, 781–788. doi: 10.1002/ps.1564

Ma, L.-J., Van Der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., et al. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature* 464, 367–373. doi: 10.1038/nature08850

Michielse, C. B., and Rep, M. (2009). Pathogen profile update: *Fusarium oxysporum*. *Mol. Plant Pathol.* 10, 311–324. doi: 10.1111/j.1364-3703.2009.00538.x

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34, i142–i150. doi: 10.1093/bioinformatics/bty266

Parikh, M., and Pierce, G. N. (2019). Dietary flaxseed: what we know and don't know about its effects on cardiovascular disease. *Can. J. Physiol. Pharmacol.* 97, 75–81. doi: 10.1139/cjpp-2018-0547

Rep, M., van der Does, H. C., Meijer, M., van Wijk, R., Houterman, P. M., Dekker, H. L., et al. (2004). A small, cysteine-rich protein secreted by *Fusarium oxysporum* during colonization of xylem vessels is required for I-3-mediated resistance in tomato. *Mol. Microbiol.* 53, 1373–1383. doi: 10.1111/j.1365-2958.2004.04177.x

Sasseron, G. R., Benchimol-Reis, L. L., Perseguini, J., Paulino, J. F. C., Bajay, M. M., Carbonell, S. A. M., et al. (2020). *Fusarium oxysporum* f. sp. *phaseoli* genetic variability assessed by new developed microsatellites. *Genet. Mol. Biol.* 43:e20190267. doi: 10.1590/1678-4685-gmb-2019-0267

Schmidt, S. M., Houterman, P. M., Schreiver, I., Ma, L., Amyotte, S., Chellappan, B., et al. (2013). MITEs in the promoters of effector genes allow prediction of novel virulence genes in *Fusarium oxysporum*. *BMC Genomics* 14:119. doi: 10.1186/1471-2164-14-119

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). "BUSCO: assessing genome assembly and annotation completeness," in *Gene Prediction: Methods and Protocols*, ed M. Kollmar (New York, NY: Springer New York), 227–245. doi: 10.1007/978-1-4939-9173-0_14

Sharma, M., Nagavardhini, A., Thudi, M., Ghosh, R., Pande, S., and Varshney, R. K. (2014). Development of DArT markers and assessment of diversity in *Fusarium oxysporum* f. sp. *ciceris*, wilt pathogen of chickpea (*Cicer arietinum* L.). *BMC Genomics* 15:454. doi: 10.1186/1471-2164-15-454

Srinivas, C., Nirmala Devi, D., Narasimha Murthy, K., Mohan, C. D., Lakshmeesha, T. R., Singh, B., et al. (2019). *Fusarium oxysporum* f. sp. *lycopersici* causal agent of vascular wilt disease of tomato: Biology to diversity - a review. *Saudi J. Biol. Sci.* 26, 1315–1324. doi: 10.1016/j.sjbs.2019.06.002

Taylor, A., Vágány, V., Jackson, A. C., Harrison, R. J., Rainoni, A., and Clarkson, J. P. (2016). Identification of pathogenicity-related genes in *Fusarium oxysporum* f. sp. *cepae*. *Mol. Plant Pathol.* 17, 1032–1047. doi: 10.1111/mpp.12346

Tourancheau, A., Mead, E. A., Zhang, X. S., and Fang, G. (2021). Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat Methods* 18, 491–498. doi: 10.1038/s41592-021-01109-3

van Dam, P., de Sain, M., Ter Horst, A., van der Gragt, M., and Rep, M. (2018). Use of comparative genomics-based markers for discrimination of host specificity in *Fusarium oxysporum*. *Appl. Environ. Microbiol.* 84:e01868-17. doi: 10.1128/AEM.01868-17

Vlaardingerbroek, I., Beerens, B., Rose, L., Fokkens, L., Cornelissen, B. J., and Rep, M. (2016). Exchange of core chromosomes and horizontal transfer of

lineage-specific chromosomes in *Fusarium oxysporum*. *Environ. Microbiol.* 18, 3702–3713. doi: 10.1111/1462-2920.13281

Wang, B., Yu, H., Jia, Y., Dong, Q., Steinberg, C., Alabouvette, C., et al. (2020). Chromosome-scale genome assembly of *Fusarium oxysporum* strain Fo47, a fungal endophyte and biocontrol agent. *Mol. Plant Microbe Interact.* 33, 1108–1111. doi: 10.1094/MPMI-05-20-0116-A

Xu, L., and Seki, M. (2020). Recent advances in the detection of base modifications using the nanopore sequencer. *J. Hum. Genet.* 65, 25–33. doi: 10.1038/s10038-019-0679-0

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29, 2669–2677. doi: 10.1093/bioinformatics/btt476

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marcais, G., et al. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Res.* 27, 787–792. doi: 10.1101/gr.213405.116

Zimin, A. V., and Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput. Biol.* 16:e1007981. doi: 10.1371/journal.pcbi.1007981

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 18 June 2021
doi: 10.3389/fgene.2021.612492

Genotyping and Whole-Genome Resequencing of Welsh Sheep Breeds Reveal Candidate Genes and Variants for Adaptation to Local Environment and Socioeconomic Traits

*James Sweet-Jones[1], Vasileios Panagiotis Lenis[2,3], Andrey A. Yurchenko[4], Nikolay S. Yudin[4], Martin Swain[2] and Denis M. Larkin[1,4]**

[1] Royal Veterinary College, University of London, London, United Kingdom, [2] Institute of Biological, Environmental and Rural Sciences, University of Aberystwyth, Aberystwyth, United Kingdom, [3] School of Health and Life Sciences, Teesside University, Middlesbrough, United Kingdom, [4] The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), Novosibirsk, Russia

OPEN ACCESS

Edited by:
Anastasia Anashkina, Engelhardt Institute of Molecular Biology (RAS), Russia

Reviewed by:
Qianjun Zhao, Chinese Academy of Agricultural Sciences (CAAS), China
Siroj Bakoev, Centre for Strategic Planning and Management of Biomedical Health, Russia
Jianlin Han, International Livestock Research Institute (ILRI), Kenya

***Correspondence:**
Denis M. Larkin
dmlarkin@gmail.com

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

Received: *30 September 2020*
Accepted: *10 May 2021*
Published: *18 June 2021*

Citation:
Sweet-Jones J, Lenis VP, Yurchenko AA, Yudin NS, Swain M and Larkin DM (2021) Genotyping and Whole-Genome Resequencing of Welsh Sheep Breeds Reveal Candidate Genes and Variants for Adaptation to Local Environment and Socioeconomic Traits. Front. Genet. 12:612492. doi: 10.3389/fgene.2021.612492

Background: Advances in genetic tools applied to livestock breeding has prompted research into the previously neglected breeds adapted to harsh local environments. One such group is the Welsh mountain sheep breeds, which can be farmed at altitudes of 300 m above sea level but are considered to have a low productive value because of their poor wool quality and small carcass size. This is contrary to the lowland breeds which are more suited to wool and meat production qualities, but do not fare well on upland pasture. Herein, medium-density genotyping data from 317 individuals representing 15 Welsh sheep breeds were used alongside the whole-genome resequencing data of 14 breeds from the same set to scan for the signatures of selection and candidate genetic variants using haplotype- and SNP-based approaches.

Results: Haplotype-based selection scan performed on the genotyping data pointed to a strong selection in the regions of *GBA3*, *PPARGC1A*, *APOB,* and *PPP1R16B* genes in the upland breeds, and *RNF24*, *PANK2*, and *MUC15* in the lowland breeds. SNP-based selection scan performed on the resequencing data pointed to the missense mutations under putative selection relating to a local adaptation in the upland breeds with functions such as angiogenesis (*VASH1*), anti-oxidation (*RWDD1*), cell stress (*HSPA5*), membrane transport (*ABCA13* and *SLC22A7*), and insulin signaling (*PTPN1* and *GIGFY1*). By contrast, genes containing candidate missense mutations in the lowland breeds are related to cell cycle (*CDK5RAP2*), cell adhesion (*CDHR3*), and coat color (*MC1R*).

Conclusion: We found new variants in genes with potentially functional consequences to the adaptation of local sheep to their environments in Wales. Knowledge of these variations is important for improving the adaptative qualities of UK and world sheep breeds through a marker-assisted selection.

Keywords: sheep, signatures of selection, Wales, whole-genome resequencing, adaptation

INTRODUCTION

Since the domestication in the semi-arid Fertile Crescent of Iran and Turkey, sheep (*Ovis aries*) have undergone migration and selection to form established breeds that are well-suited to various local environments (Zeder, 2008). The process of natural or artificial positive/negative selection results in genomic regions of a decreased diversity, which are known as the signatures of selection (Kijas et al., 2012). Detecting signatures of selection is important for understanding the genetic mechanisms of the adaptation of breeds to their local environments. Previous investigations have unearthed genes relating to hypoxia in sheep adapted to high altitudes on the Qinghai-Tibetan Plateau of China (*THRB*) (Yang et al., 2016), fat deposition in sheep from arid deserts of northern Africa (*PCDH9*) (Kim et al., 2016), and metabolism in Russian sheep adapted to low temperatures (*POMC*) (Yurchenko et al., 2019). Whilst these examples represent the extreme circumstances, there are examples of selection in sheep adapted to more temperate climates. For instance, three upland sheep breeds from northern England were shown to demonstrate higher than expected frequencies of the known missense mutations in genes associated with reproductive success (*PRLP*) and presence of horns (*RXFP2*) (Bowles et al., 2014). Knowledge of such selection is essential for the continued improvement of breeds better suited to their environments and with a better socioeconomic trait potential in current selection programs, which is particularly pressing in terms of climate change (Bowles, 2015). Here, we investigated another example of previously neglected breeds adapted to local environments in Wales, United Kingdom, where sheep have historically been farmed at altitudes of >1,000 feet (∼300 m) above sea level on a rough pasture with shallow-rooted plants (Davies, 1935).

Sheep were introduced to Wales by Neolithic settlers, bringing a primitive breed similar to the contemporary Soay (6,000 years ago) with two likely further introductions by Roman white-faced, fine-wool breeds (2,000 years ago) and Norse black-faced breeds (1,000 years ago) (Ryder, 1964). The South Wales Welsh Mountain (SWWM) breed of today, a descendent of the Roman and Soay breeds, has been documented since the 16th century and is renowned for its hardiness, lambing ability, and sweet meat taste. Despite this potential economic gain, these mountain sheep suffer a trade-off to a kemp wool and small carcass size, leading them to be poorly exploited outside of Wales (Williams-Davies, 1981). To overcome this, the SWWM breed was cross-bred in the eighteenth century to form a number of more productive local breeds, which are now farmed on the lowlands (Ryder, 1964; Williams-Davies, 1981).

Previous work to reveal the population history of Welsh sheep showed that Welsh breeds clustered closely with each other based on haplotype sharing, forming two groups within this cluster, which aligned with the upland or lowland farming style. Exceptions to this were the Black Welsh Mountain (BWM) and Kerry Hill with Beulah, which remained distinct from all other breeds (Beynon et al., 2015). This distinction of BWM is supportive of the alternate ancestry from Norse breeds and deliberate selection based on coat color (Williams-Davies, 1981). Moreover, the divergence of Kerry Hill and Beulah may be the result of a genetic bottleneck or a founder effect that these breeds underwent. Additional support of this can be seen in the low effective population sizes and high inbreeding coefficients of these breeds (Beynon et al., 2015). Low effective population sizes endanger these breeds to risks of increased homozygosity and lack of genetic diversity due to inbreeding. This is a potential risk for UK upland sheep breeds who have remained geographically and genetically isolated due to their adaptability to thrive on pastures that other breeds cannot (Bowles et al., 2014; Heaton et al., 2014).

With this in mind, it is important, from a perspective of cultural significance, breed conservation, and breed improvement, to study the adaptation of Welsh breeds, which potentially offers insight into genomic adaption to upland farming, as well as lambing and meat quality. Likewise, lowland productive breeds offer a good comparison due to their stronger capabilities for traits related to socioeconomic gain. Through the use of a commercially available medium-density SNP genotyping array with a HapFLK and Decorrelated Composite of Multiple Statistics (DCMS) software and whole-genome resequencing data with a DCMS pipeline, this paper aims to identify the signatures of selection in the genomes of Welsh sheep breeds and candidate genes containing functional missense mutations in these regions.

MATERIALS AND METHODS

Data Source and Variant Calling

Genotyping data from 353 individuals across 18 Welsh breeds on the Illumina OvineSNP50 SNP array from Beynon et al. (2015) were used in this study. Illumina pair-ended read (150 bp) resequencing of 11 Welsh sheep samples from the same set, representing one sample per breed (**Supplementary Table 1**), was performed at the University of Aberystwyth to ∼13× raw coverage using Illumina HiSeq according to the manufacturer's protocol. Three remaining Welsh sheep resequenced genomes for Welsh Hardy-Speckled Faced, Dolgellau Welsh Mountain, and Talybont Welsh Mountain were downloaded from the National Centre for Biotechnology Information Sequence Read Archive, PRJNA160933 (Heaton et al., 2014). A dataset of resequenced Russian sheep samples (n = 40) adapted to a contrasting environment was used as an outgroup in this study (Sweet-Jones et al., 2021).

Reads were mapped by using the Burrows-Wheeler Aligner BWA-MEM (BWA V.0.7.10 (Li, 2013) to the reference sheep genome Oar_V.3.1. Reads were sorted using the Samtools V.0.1.18 (Li et al., 2009) and duplicates were marked with the Picard V.2.18[1]. Libraries were also merged using Picard. Base Quality Score Recalibration (BQSR) was performed, which account for the systematic errors in sequencing, using the Genome Analysis Toolkit (GATK V.3.8; McKenna et al., 2010). Samples then underwent variant calling with the GATK HaplotypeCaller function. Finally, all samples (n = 54) underwent joint calling to merge all reported variants into a single vcf file. Following this, hard filtering for quality scores assigned by BQSR was performed by GATK, using the filter

[1]http://broadinstitute.github.io/picard/

expression "[QD < 2.0| | FS > 60.00| | MQ < 40.00| | MQRankSum < −12.5| | ReadPosRankSum < −8.0]." All variants from resequenced data were converted into a Plink V.1.90 (Purcell et al., 2007) format to be run through the DCMS pipeline (Yurchenko et al., 2019).

HapFLK Statistics

Welsh breed genotyping data were separated into three groups of related breeds based on the clustering analysis performed by Beynon et al. (2015), resulting in one group of upland breeds and two groups of lowland breeds. Of the two lowland groups, one consisted of five lowland breeds and the other consisted of only the Kerry Hill and Beulah breeds (KHB) (**Table 1**). Plink-formatted (Purcell et al., 2007) files for the upland and lowland breed groups had genotypes from the sex chromosomes removed and were filtered to remove the rare alleles [–maf 0.05], low called SNPs [–geno 0.01], or poorly genotyped individuals [–mind 0.05]. FastPhase V.1.4 (Scheet and Stephens, 2006) was used to estimate the number of haplotype clusters (k) for each group (Lowland k = 48, KHB k = 25, and Upland k = 53). HapFLK software (Bonhomme et al., 2010) was used to obtain selection statistics for each group. This test uses the hierarchal population structure to identify the haplotype-based selective sweeps, which focuses on the inherited combinations of alleles. It has the advantage of an increased statistical power, reliably detecting the hard and soft selective sweeps, and is a realistic simulation of selection through the haplotypes. HapFLK p-values were calculated using the Python script scaling_chi2_hapflk.py (Bonhomme et al., 2010; Fariello et al., 2013). Adjusted p-values, or q-values, were calculated through the R qqman q-value function (Turner, 2014). Selected intervals were determined by boundaries of $q < 0.05$ with SNPs within an interval of $q < 0.01$ considered to be under a strong selection.

De-correlated Composite of Multiple Signals (DCMS)

Five established measures of selection and genetic diversity were both used on the genotyping (15 breeds) and whole genome resequencing (upland breeds $n = 7$; lowland breeds = 7) datasets as a DCMS (**Table 1**; Ma et al., 2015). Statistics used were: (i/ii) *H2*/*H1* and *H12*, which can distinguish the hard and soft selective sweeps by measuring the intensity of selection (Garud et al., 2015); (iii) Tajima's D comparing pairwise sequence differences and the number of segregating sites, detecting positive, negative, or balancing selection (Tajima, 1989); (iv) nucleotide diversity (*Pi*), average number of nucleotide differences between two sequences (Nei and Li, 1979); (v) F_{ST} fixation index, comparing single SNP frequencies across a population (Weir and Cockerham, 1984). By weighting the result of each statistic and generating a combined score, regions that overlap in the analysis outcomes gain a stronger evidence to be a region under selection.

TABLE 1 | Breed representation in the genotyping and resequencing datasets.

Breed	Abbreviation	Genotyped samples	Horns	Base color	Fleece	HapFLK group	Resequencing group
Badger Faced Welsh Mountain	BFWM	21	Yes	Black and white	Firm	Upland	NA*
Balwen	–	14	No	Black and white	Firm	Upland	Upland
Beulah	–	22	No	Black and white	Fine	KHB	Lowland
Black Welsh Mountain	BWM	24	Yes	Black	Firm	Upland	Upland
Brecknock Hill Cheviot	BHC	24	No	White	Fine	Upland	Upland
Clun Forest	–	17	No	Black	Fine	Lowland	Lowland
Dolgellau Welsh Mountain	DWM	–	Yes	White	Firm	NA*	Upland
Hardy Speckled Faced	HSF	24	Yes	White	Fine	Lowland	Lowland
Hill Radnor	–	21	No	Gray-brown	Fine	Lowland	Lowland
Kerry Hill	–	18	No	White	Fine	KHB	Lowland
Llandovery White Faced	LWF	24	No	White	Fine	Upland	Upland
Llanwenog	–	21	No	Black	Fine	Lowland	Lowland
Lleyn	–	22	No	White	Fine	Lowland	Lowland
South Wales Welsh Mountain	SWWM	17	Yes	White	Firm	Upland	Upland
Talybont Welsh Mountain	TWM	24	No	White	Firm	Upland	Upland
Welsh Mountain Hill Flock	WMHF	24	Yes	White	Firm	Upland	NA
Average/Total	–	21/317	No	–	–	102/40/173	7/7

**NA denotes the exclusion of breed from either the HapFLK or resequencing study due to the unavailability of samples.*

H2/H1 and H12

Autosomal SNPs were filtered for –maf 0.0000001 –geno 0.1 –mind 0.1 by Plink. SNPs were then phased by chromosome using ShapeIt2 (Delaneau et al., 2011) with 400 states and an effective population size of 100. Phased chromosomes were split into appropriate groupings per chromosome using Plink and H2/H1, H12 were calculated using the H1_H12.py Python script (Garud et al., 2015). H2/H1 and H12 values were calculated in windows of 25 SNPs using a step size of one SNP following our previous study (Yurchenko et al., 2019).

Tajima's *D*

Tajima's *D* for mutation index was calculated over the same intervals of 25 SNPs, whose lengths were calculated from the output of the H2/H1 and H12 statistics. Using the vcftools V.0.1.13 (Danecek et al., 2011), Tajima's D was calculated [–TajimaD 900000000] for each chromosome per group/breed file per window. Output files were concatenated per group.

Fixation Index

F_{ST} was calculated with Plink comparing each group to all others. All negative values were converted to zero, and data were smoothed with the R *runmed* function in windows of 31 SNPs.

Nucleotide Diversity

Plink-format file was split by chromosome per breed, and nucleotide diversity was calculated with the vcftools [–site-pi] option. Data were then smoothed using the R *runmed* function in windows of 31 SNPs.

Combining Statistics With DCMS

Output files from individual statistics were sorted and joined by SNP id. Genome-wide *p*-values through ranking results of each statistic were calculated in the R MINOTAUR *stat-to-p*-value function specifying the one-tailed tests (H2/H1, H12, F_{ST}–right tailed, Pi, and Tajima's D–left tailed). Covariance matrix was constructed based on sampling 300,000 randomly sampled SNPs using the R *CovNAMcd* function where the alpha = 0.75. DCMS statistics were calculated using the *DCMS* function and fitted to a normal distribution to examine normality implemented by the R MASS *rlm* function. These fitted DCMS values were converted to *p*-values using *pnorm*, and adjusted for a false discovery rate with the *qvalue* function. *Q*-values were parsed to determine selection with region boundaries set at a $q < 0.2$ and a threshold of $q < 0.01$.

Candidate Gene Search

For the regions defined by our pipelines as being under selection, genes were identified using a list of 26,958 genes of the Oar_V.3.1 genome downloaded from Ensembl BioMart v.98. Within each selected region, genes were then ranked based on their distances to the most significant SNP of that region with the closest gene being the top-ranking. The top 10 highest ranking genes from each region underwent literature review for their previous associations to adaptation to local environments or socioeconomic traits in animals. Genes with established links to these traits were identified as candidate genes in this study.

In the resequencing study, all SNPs were annotated with NGS-SNP (Grant et al., 2011) to identify missense mutations. Only genes with missense mutations in the regions under selection were considered for literature review. Additionally, missense SNPs with a strong support from the F_{ST} statistic ($F_{ST} \geq 0.3$) were analyzed with PolyPhen2 to predict their effects on protein structure and function (0 = benign and 1 = deleterious; Adzhubei et al., 2013).

Functional enrichment analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) v. 6.8 (Huang et al., 2009) using the same gene list downloaded from Ensembl Biomart v. 98. Enrichments were detected using the DAVID functional clustering tool, which verifies enrichments of similar GO terms across many different databases, to confer a stronger evidence of enrichment. Scores > 1.3-fold enrichment were considered.

Copy Number Variant (CNV) Analysis of Resequenced Genomes

CNV detection was performed to identify the regions in the genomes of Welsh sheep that had been duplicated or deleted with respect to the Texel reference genome. We identified CNVs in the resequenced samples with the cn.Mops R package (Klambauer et al., 2012) in a window length of 700 SNPs, using sequences that had undergone BQSR. This resulted in each individual being given a raw copy number (CN) per window. CN1–CN3 were considered to be normal and discounted from the results. Raw CNVs were merged into the CNV Regions (CNVRs) with the BedOps *bedmap* function using at least 50% reciprocal overlap in at least three individuals within the same group as criteria for inclusion. Duplicate CNVRs were removed, and the neighboring CNVRs were merged. This allowed us to be confident in our results by excluding the regions where CNVRs appear to overlap the signatures of selection, as these cast doubt over their reliability.

Data Visualization

HapFLK clusterplots and haplotype trees were visualized from the prepared R scripts available (Bonhomme et al., 2010; Fariello et al., 2013). All Manhattan plots were rendered in R by the *qqman* package setting the suggestive line (q = 0.05) and genome-wide line (q = 0.01) to indicate selection. Haplostrips for visualization of haplotype sharing (Marnetto and Huerta-Sánchez, 2017) in the regions under selection detected by DCMS were run using phased data.

RESULTS

Regions Under Selection Detected From Genotyping Data

Using the HapFLK software, signatures of selection were found in each grouping using 44,711 SNPs (lowland group contained one region, KHB-two, and upland-31; **Supplementary Tables 2–4**). Cluster plots and Haplotype Trees for the most significantly selected regions in the KBH group are shown in **Figure 1** and

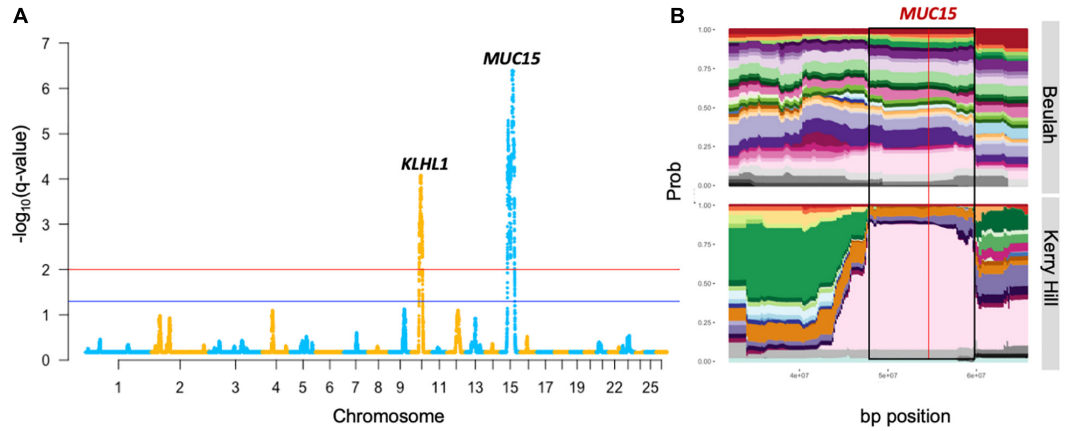


FIGURE 1 | Output of genomic regions under strong selection in the KHB group from the HapFLK study. This includes the Manhattan plot **(A)** with a gene proximal to the most significant SNP of the highly selected peaks. Cluster plot **(B)** shows the tracks of haplotype diversity for the peak highlighted on the Manhattan plot by black borders where the red line represents the most significant SNP.

for other groups, in **Supplementary Figure 1**. Lengths of the regions found under selection ranged from 1.81 to 64.78 Mb with a median length of 7.33 Mb. Moreover, DCMS also found the regions under selection in all 15 breeds, with at least one region of a strong selection ($q < 0.01$) in each breed using 47,366 SNPs. Lengths of the regions under selection ranged from 1 to 3.9 Mb where the median length of regions was 0.16 Mb. The number of regions detected in each breed ranged from 8 in Kerry Hill to 89 in Clun Forest (**Supplementary Tables 5–19**). In these regions under selection, 1,089 unique genes were found, with 179 occurring in multiple breeds including: *GBA3* (Hill Radnor, LWF, Lleyn, SWWM, TWM, and WMHF), *ENSOARG00000021104* (BFWM, Clun Forest, Hill Radnor, and Lleyn), *PCDH9* (BHC, Hill Radnor, and Llanwenog), and *PPARGCA1* (LWF, SWWM, and WMHF). Manhattan plots for all the breeds investigated are shown in **Figure 2** and **Supplementary Figure 2**.

There was a substantial overlap between the results for a region found on OAR6: 31.99–46.00 Mb ($q = 2.0 \times 10^{-5}$) in the upland breeds in the HapFLK study to a narrower region, OAR6: 40.20–43.38 Mb, in LWF ($q = 6.9 \times 10^{-15}$), WMHF ($q = 4.3 \times 10^{-11}$), and SWWM ($q = 8.73 \times 10^{-5}$). In all cases, the top ranked genes were *GBA3* linked to liver metabolism (Dekker et al., 2011) and *PPARGC1A* with known roles in mitochondrial biogenesis, fat deposition, and milk fatty acid composition in cattle (Fernandez-Marcos and Auwerx, 2011; Armstrong et al., 2018; Yuan et al., 2019; Li et al., 2016). On the HapFLK cluster plot for this region, low diversity was seen in the Balwen, BWM, LWF, and SWWM breeds, further supporting the idea of selection for this region. When plotted on a haplotype tree, it can be seen that Balwen had the longest branch compared to the other upland breeds, whilst the strongest signature of selection is seen in SWWM (**Supplementary Figure 1**).

Several other candidate regions detected by HapFLK overlapped with the regions found in the Welsh breeds by DCMS. OAR3: 23.23–33.86 Mb ($q = 5.0 \times 10^{-6}$) shared its top-ranked genes, *TDRD15* and *APOB,* with a region under selection in the upland Balwen breed ($q = 0.0003$). These genes are associated with cholesterol mobilization in Large White pigs (Bovo et al., 2019). OAR7: 47.55–54.20 Mb ($q = 0.0008$) overlapped with another region OAR7: 51.30–52.76 Mb in the upland BWM ($q = 5.0 \times 10^{-9}$), but these did not share the top-ranked genes. Strong overlapping candidate genes are presented in **Table 2**.

The most strongly selected region in the HapFLK study was seen on OAR15 32.15–72.10 Mb in the KHB lowland group, with the top-ranking gene being *MUC15* ($q = 4.1 \times 10^{-7}$), associated with a low fecal egg count in Spanish Churra sheep during gastrointestinal parasite infections (Periasamy et al., 2014; Benavides et al., 2015). This was supported by a low haplotype diversity seen in the Kerry Hill breed at this locus, but this signature was not seen in Beulah or in the DCMS results in either breeds.

In the other lowland group, the only region found under selection, OAR13: 47.16–54.37 Mb ($q = 0.01$), overlapped with another region found in the upland group, OAR13: 46.62–72.94 Mb ($q = 2.0 \times 10^{-6}$), but the top candidate genes were different. In the lowland breeds, the top-ranked genes were *RNF24* and *PANK2*, which have been found under selection in world sheep breeds in association to a loss of vision following domestication (Naval-Sanchez et al., 2018; Wang et al., 2019). Cluster plots for this region showed a decreased haplotype diversity amongst the selected region on OAR13 in the lowland breeds, especially Clun Forest and HSF. Furthermore, significant selection at this locus in HSF ($q = 0.0006$) was confirmed using the DCMS pipeline. In the case of the upland breeds, the top-ranked genes were *DHX35* and *PPP1R16B* which are related to innate viral immunity (Rahman et al., 2017) and endothelial cell proliferation (Pszczola et al., 2018), respectively.

The genes found within the regions under selection from HapFLK, functional enrichments seen for the KHB group included the DENN domain and connexin gap junctions, enriched 1.6- and 1.4-fold, respectively. The most highly enriched terms in the upland breeds were bactericidal permeability protein, major intrinsic protein, and semaphorins, which were all enriched over threefold. The most enriched cluster seen in

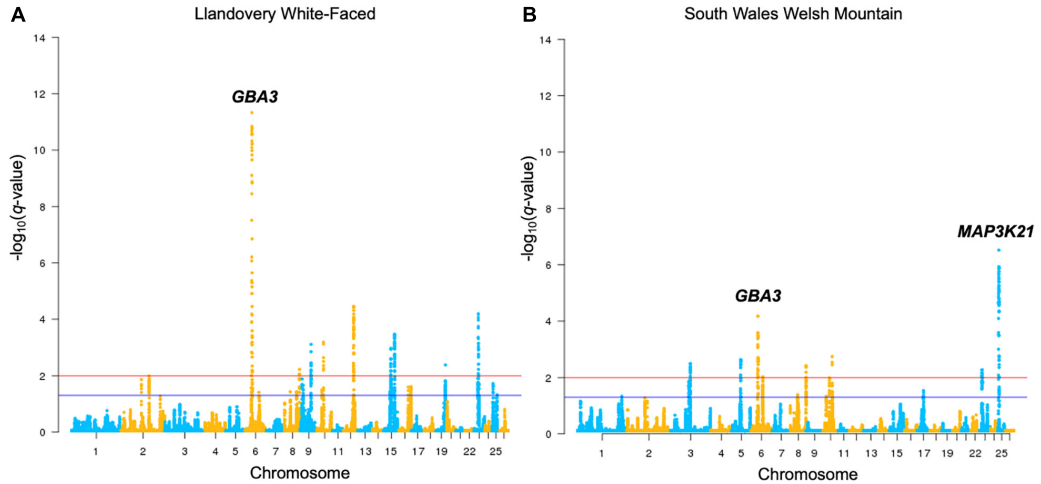


FIGURE 2 | Manhattan plots of Llandovery White-faced **(A)** and South Wales Welsh Mountain **(B)** genotyped breeds' DCMS output. All other breeds are shown in **Supplementary Figure 2**. Selection thresholds for a suggestive ($q < 0.05$) and strong ($q < 0.01$) selection shown by blue and red lines, respectively. Significant selection peaks have been annotated with the name of the top-ranked gene of that region.

TABLE 2 | Candidate genomic regions and genes from the HapFLK analysis and corresponding DCMS region.

OAR	Region start	Region end	Group	*q*-value	Candidate genes (rank)	Function	Overlapping DCMS result
1	258,588,662	275,499,089	Upland	0.0002	*SIM2(1), CLDN14(2)*		
3	23,229,012	33,863,536	Upland	4.0×10^{-6}	*TDRD15(1), APOB(2)*	Cholesterol efflux	Balwen
6	31,991,037	45,997,407	Upland	1.0×10^{-5}	*GBA3(1), PPARGC1A(2), DHX15(3), SOD3(4), KCNIP4(6)*	Metabolism	LWF, WMHF, SWWM
10	35,062,152	51,872,705	KHB	8.0×10^{-5}	*KLHL1(1), PCDH9(2)*	Neurodevelopment, cell adhesion	
13	46,620,758	72,941,741	Upland	2.0×10^{-6}	*DHX35(1), PPP1R16B(3), ADIG(6)*	Innate immunity, angiogenesis	BFWM
13	47,164,400	54,366,328	Lowland	0.01	*RNF24(1), PANK2(2), MAVS(3)*	Vision	HSF
15	32,152,675	65,710,885	Upland	4.0×10^{-7}	*MUC15(1), ANO3(3)*	Mucous production	

the genotyping DCMS analysis was the Type II keratin filaments from the Hill Radnor breed, which was enriched fivefold. This was followed by the Ribonuclease A in Clun forest, enriched 2.8-fold and leucine rich-repeats in Lleyn, enriched 2.7-fold (**Supplementary Table 20**).

Signatures of Selection Detected From Resequencing Data

Fifty-four (14 Welsh and 40 Russian sheep) resequenced genomes were aligned to the Oar_V.3.1 genome with a mean filtered coverage of 11.9× (**Supplementary Table 1**). A total of 41,643,098 SNPs were called, which were pruned to 38,276,494 SNPs after filtering. CNVRs covered 0.27% of the lowland and 0.24% of the upland genomes, overlapping 852 and 669 genes, respectively (**Supplementary Tables 21,22**). Three CNVRs from the lowland breeds and 52 CNVRs from the upland breeds overlapped the regions of selection. Some of these CNVRs had a high frequency in the population, including the regions under a strong selection on OAR24 in the lowland breeds, spanning the *CLCN7* gene and on OAR17, whereas in the upland breeds, these spanned the *IGLV4-69, ZNF280B*, and *PRAME* genes.

After excluding the regions overlapping CNVRs, DCMS found 2,996 regions under selection in the Welsh breeds (lowland = 514, upland = 2,482; **Supplementary Tables 23,24** and **Figure 3A**). These regions overlapped 104 and 430 genes in the lowland and upland breeds, respectively. The most significantly selected region in the upland group was OAR22: 15.3676–15.3679 Mb, which overlapped the *NOC3L* gene ($q = 1.8 \times 10^{-8}$), and for the lowland group, it was a single synonymous SNP located in *MYH11* ($q = 0.0006$).

Identification of Candidate Genes and Missense Mutations

In regions under selection, there were 12 missense mutations found in the lowland breeds (**Supplementary Table 25**) and 85 missense mutations found in the upland breeds (**Supplementary Table 26**). Of these, only missense mutations and their enclosing genes, which were top-ranking SNPs in their selected intervals are discussed below, leading to a total of 4 in the lowland breeds and 14 in the upland breeds. Clarification of the type of selection relied on a strong support from the *H2/H1* and *H12* statistics, which were considered to be haplotype-based selection, or F_{ST},

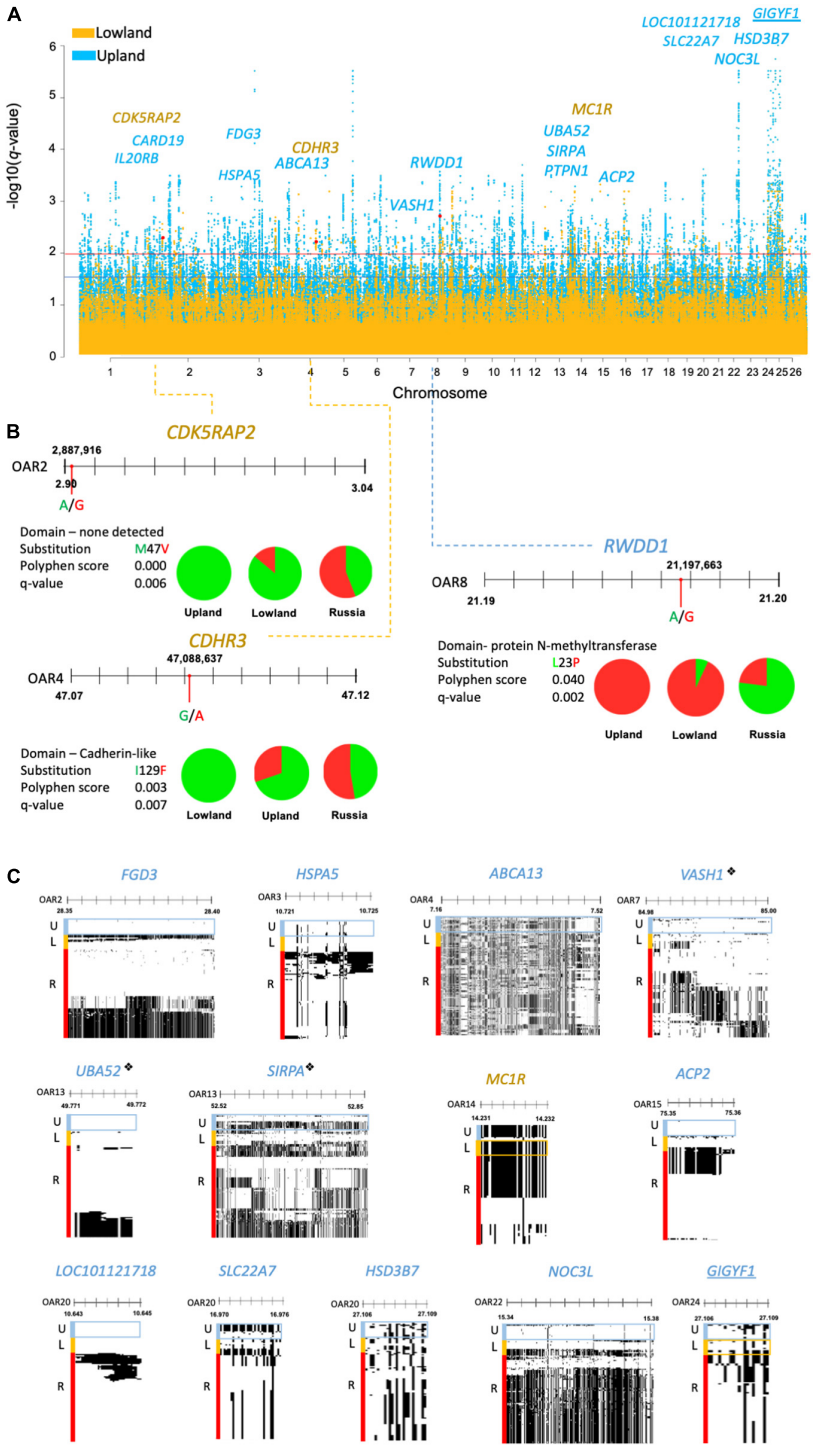


FIGURE 3 | (A) Manhattan plot of DCMS *q*-values of the lowland (yellow) and upland (blue) breeds showing missense mutations found under selection, highlighted in red with the corresponding gene names. Selection thresholds for a suggestive ($q < 0.05$) and strong ($q < 0.01$) selection shown by blue and red lines, respectively. Underlined gene names show selection in both the lowland and upland breeds. **(B)** Allele frequencies for missense mutations identified by a strong F_{ST} score represented by pie charts (green = reference allele, red = derived allele). This shows the location of missense mutations along their gene with nucleotide substitution highlighted by a red circle. Blue and yellow dotted lines point to the corresponding peak positions of the Manhattan plot. Amino acid substitution is shown, alongside the Polyphen score and *q*-value. **(C)** Haplostrip plots spanning genes containing the missense mutations but were selected on the basis of the H2/H1 and H12 haplotype statistics. Similar haplotypes are clustered together per population to demonstrate selection within these regions across the whole gene. These show the presence of reference (white) or derived (black) alleles making up different haplotypes. Populations of interest are highlighted in boxes corresponding to their colors on the Manhattan plot. "❖" is used to denote genes under selection by DCMS and HapFLK.

which was considered fixation of a variant in a population and was seen as a selection acting on individual SNPs.

Three missense mutations were found in the regions under selection ($q < 0.01$), with a strong support from the F_{ST} statistic ($F_{ST} \geq 0.3$). In the lowland breeds, these included the reference alleles of OAR2:2,887,916 in the *CDK5RAP2* gene ($q = 0.006$; $F_{ST} = 0.3$) and OAR4:47,087,846 in *CDHR3* ($q = 0.007$; $F_{ST} = 0.3$). For the upland breeds, only one missense mutation was found, relating to the derived allele of OAR8:21,197,663 in the *RWDD1* gene ($q = 0.002$; $F_{ST} = 0.5$). The Polyphen score of this selected mutation, L23P, was low and did not support the large change in the protein function (**Figure 3B**).

Missense SNPs in the regions under selection supported by the haplotype statistics (*H2/H1*, *H12*) were only found in the *MC1R* gene in the lowland breeds, and in the upland breeds: *FGD3*, *HSPA5*, *ABCA13*, *PTPN1*, *ACP2*, *LOC101121718*, *NOC3L*, *VASH1*, *SIRPA*, and *UBA52* genes. Alternatively, some missense SNPs received strong support from Tajima's *D* and *Pi*, including one in the *GIGYF1* gene found in the selected regions in both the upland and lowland breeds, as well as the *SLC22A7* and *HSD3B7* genes in the upland breeds (**Figure 3C**).

In the upland breeds, three of the genes found with missense SNPs were also identified in the regions under selection defined in the upland breeds of the genotyping analysis. These included: *VASH1* ($q = 0.003$) on OAR7, which also overlapped with BWM from the genotyping DCMS data; *UBA52* ($q = 0.01$) on OAR13, which also overlapped with BHC from the genotyping data, and *SIRPA* ($q = 0.007$). No missense mutations found in the lowland breeds overlapped with the regions under selection detected from the genotyping dataset.

Candidate Genes in Welsh Lowland Sheep Breeds

MC1R, found in a region under selection in the lowland breeds (**Table 3**), is known to cause an upregulation of tyrosinase in hair melanocytes, which led to an increase of black eumelanin pigment; however, in sheep, typically, white pheomelanin is produced due to selection for mutations in *MC1R* (Weatherhead and Logan, 1981; Yang et al., 2013). Another gene, *CDK5RAP2*, with a strong support from allele frequency statistics, is related to human neurodevelopment by recruiting tubulin subunits, which are important for cortical gyration (Issa et al., 2013). Mutations in this gene have also been linked to Hertwig's anemia mutant mouse models displaying blood cytopenia, aneuploidy due to impaired cell-cycle spindle checkpoints, and increased neuronal cell death (Lizarraga et al., 2010). Finally, *CDHR3*, encoding a cell adhesion protein is linked to childhood asthma and rhinovirus-C susceptibility, was also located in a region under selection in Asian sheep by HapFLK (Fariello et al., 2014; Bochkov et al., 2015).

Candidate Genes in Welsh Upland Sheep Breeds

Within the upland breeds, many genes found in the selected regions had functions related to cell stress and metabolism (**Table 3**). The most highly supported by DCMS, *RWDD1*, encodes a transcription factor related to sulfide metabolism in metozoans (Kang et al., 2008; Li et al., 2018). *HSPA5* also responds to cell stress mechanisms by promoting protein refolding in the endoplasmic reticulum in cancers or virally-infected cells (Booth et al., 2015). This gene has also been found under selection in

TABLE 3 | Candidate genes and missense mutations in the lowland and upland sheep breeds.

OAR	Region start	Region end	SNP	Breed	Gene (rank)	Reference allele	Alternative allele	Mutation	PolyPhen	*q*-value	Function
2	2,887,916	2,887,916	2,887,916	Lowland	*CDK5RAP2*(1)	A*	G	M47V	0	0.006	Cell cycle
2	28,318,917	28,318,992		Upland	*FGD3*(1)					0.003	GTP binding
3	10,725,499	10,725,501		Upland	*HSPA5*(1)					0.007	Cell stress
4	7,476,204	7,476,211	7,476,211	Upland	*ABCA13*(1)	T*	C	K2619Q	0.9	0.006	Membrane transport
4	47,088,764	47,089,276		Lowland	*CDHR3*(1)					0.004	Cell adhesion
7	84,989,884	84,989,669		Upland	*VASH1*(1)					0.004	Angiogenesis
8	21,197,663	21,198,137	21,197,663	Upland	*RWDD1*(1)	A	G*	L23P	0.04	0.002	Anti-oxidation
13	49,771,782	49,771,855		Upland	*UBA52*(1)					0.1	Ubiquitinoylation
13	52,848,831	52,848,943		Upland	*SIRPA*(1)					0.006	Cell recognition
13	52,848,831	52,848,943		Upland	*PTPN1*(1)					0.1	Insulin pathway
14	14,231,667	14,232,187		Lowland	*MC1R*(1)					0.004	Coat color
15	75,351,718	75,352,196		Upland	*ACP2*(1)					0.006	Development
20	10,644,119	10,644,602		Upland	*LOC10110726*(1)					0.009	Fat deposition
20	16,972,103	16,972,103		Upland	*SLC22A7*(1)					0.009	Cyclic nucleotide transport
22	15,368,530	15,368,926		Upland	*NOC3L*(1)					2.0×10^{-8}	Cell cycle
24	27,106,763	27,106,886		Upland	*HSD3B7*(1)					0.002	Bile synthesis
24	35,794,428	35,794,428		Lowland/Upland	*GIGFY*(1)					0.007	Insulin pathway

**Denotes selected allele.*

Chinese Yellow-Feathered chickens with regard to meat quality (Huang et al., 2020) and muscling in world pig breeds (Li et al., 2011).

Of the three genes found in the selected regions in both the genotyping and resequencing datasets, two related to cell stress mechanisms. The most significant region contains *VASH1*, encoding the vasohibin 1 signaling molecule, with known roles of negatively regulating angiogenesis (Chen et al., 2020), as well as promoting expression of antioxidation enzymes (Miyashita et al., 2012). Age-related downregulation of *VASH1* leads to a lower endothelial cell stress tolerance, posing as a risk factor for human vascular diseases in later life (Takeda et al., 2016). Secondly, *UBA52* encodes a protein with an ubiquitinate activity, with its downregulation causing cell cycle arrest and reduced protein synthesis essential for pre-implantation embryogenesis success in mouse models (Kobayashi et al., 2016).

Several genes related to metabolism and growth were found under selection in the upland breeds. The most significant of these, *HSD3B7*, is part of the *bile biosynthesis* pathway (Cheng et al., 2003). *PTPN1* is a risk-factor gene linked to diabetes and obesity (Olivier et al., 2004), which has a direct involvement in the *insulin* and *leptin signaling pathways*, and that mice lacking this gene were resistant to weight gain and intolerant to glucose (Elchebly et al., 1999). *GIGYF1* also has roles in enhancing the insulin receptor pathway, but additionally has been linked to translational repression (Giovannone et al., 2003; Tollenaere et al., 2019). Similar effects have been seen with *APC2*, which is linked to muscle mass in mice (Kärst et al., 2011). The Rho-GEF-containing gene *FGD3*, expressed in the growth plate of long bones, was previously found under selection in French Trotter and Gidran horses, as well as in Jutland and Japanese black cattle in association to birth weight (Takasuga et al., 2015; Grilz-Seger et al., 2019; Stronen et al., 2019).

Two membrane transport proteins, ABCA13 and SLC22A7, were found to be in the regions under selection in the upland breeds. *ABCA13* encodes a member of the ATP-binding cassette membrane transporter family, responsible for the active transport of biological substrates across cell membranes (Prades et al., 2002). Secondly, *SLC22A7*, a transmembrane solute carrier, with roles in cAMP and cGMP transport in mammalian tissues and, therefore, is important for intracellular signaling which may mobilize intracellular Ca^{2+}, activate protein kinases, or activate transcription factors (Kobayashi et al., 2005; Yan et al., 2016). The final gene shared between the genotyping and resequencing data was *SIRPA*, which is expressed by macrophages and polarizes M1 phagocytic macrophages to M2 antiphagocytic macrophages, which is a key survival strategy for tumors (Barclay and van den Berg, 2014).

Gene Ontology Enrichments Show Adaptations to Environment in the Resequenced Lowland and Upland Breeds

Nine functional category enrichments were found from genes within CNVRs in the lowland breeds and 14 enrichments in the upland breeds (**Supplementary Tables 27,28**). Some of these enrichments were shared between the lowland and upland breeds. These were semaphorins; ion transport, pleckstrin homology domain; SAND domains; and EGF-like domains. Exclusive enrichments in the lowland breeds' CNVRs included cell surface receptors, neuromuscular process, and DNA binding whereas exclusive enrichments in the upland breeds included Src homology-3 domain Rho signal transduction, Notch signaling, and chondrocyte differentiation (1.4-fold). Genes within the regions under selection in the upland breeds showed significantly enriched clusters including interleukin-1 and ATP-binding (**Supplementary Table 29**). No functional category enrichments were found in genes in the regions under selection in the lowland breeds.

DISCUSSION

Our study has demonstrated that regions under putative selection in Welsh sheep genomes contain candidate genes for adaptation to their local environment and production of socioeconomic traits. We used a large set of animals genotyped on a relatively small number of SNPs, applying the haplotype and point-based selection scan algorithms. This was combined with a relatively small number of resequenced individuals subjected to point-based selection scan. As a result, we detected genomic regions under selection in individual and groups of breeds, including candidate missense variants within these regions. Regions under selection detected by the three approaches followed the expected patterns seen previously that the haplotype-based approach would detect larger but fewer regions than the point-based approach, which detected smaller, but more numerous regions under selection (Yurchenko et al., 2019).

Exposure to altitudes has a range of deleterious effects caused by hypoxia, exposure to ultraviolet radiation, and generation of oxygen radicals. These, in turn, have been linked to a negative energy balance, dysregulated proteostasis, cellular stress mechanisms, and DNA damage (Askew, 2002; Pasiakos et al., 2017). Therefore, the presence of genes related to cell-stress and anti-oxidation in the regions under selection gives reassurance that the results from this study are relevant to the adaptations of Welsh sheep to altitudes. Genes related to hypoxia, however, were not identified in the regions under selection in the upland breeds, so it can be assumed that it is not a stress factor for these breeds because the altitudes they are farmed at are only moderate. Furthermore, body conditioning genes, such as *FDG3* and *ACP2,* in the upland breeds also suggest physical mechanisms of adaptation, such as increased fat deposition and muscle mass, however, these could also be linked to socioeconomic performance (Giovannone et al., 2003; Bento et al., 2004; Gu et al., 2011; Kärst et al., 2011; Grilz-Seger et al., 2019).

We observed selection at the loci of other top-ranking candidates from the haplotype analysis with known roles in energy consumption, liver metabolism, milking, fat deposition, and angiogenesis (Yang et al., 2016; Armstrong et al., 2018; Pszczola et al., 2018). These findings, showing the top-ranking genes sharing functions of that in the resequencing study, provide many candidate genes relating to survivability and socioeconomic traits in Welsh upland sheep.

Further evidence of selection in these breeds can be seen from functional term enrichments in genes found in the selected regions and in CNVRs.

Lowland breeds showed less signatures of selection, however, they mainly had selection in regions containing genes known to be associated with domestication, which are commonly reported signatures of selection in world sheep breeds (Kijas et al., 2012; Wei et al., 2015; Wang et al., 2019; Li et al., 2020). By demonstrating the lowland breeds sharing the signatures of selection with other productive breeds, this indicates that they are better suited for productive qualities but do not show an adaptation for the Welsh uplands. Despite this, selection for *CDK5RAP2* seen in the lowland breeds of the resequencing study may be linked to upland adaptation as mice models with truncating mutations have lower red blood cell counts, however, this is true for white blood cells too, and so, may be linked to immunity traits as well as neurodevelopment (Barker and Bernstein, 1983).

Differences in the results when using the genotyping and resequencing data were expected and can be attested to an increased density of SNPs in the resequencing data and a different composition of populations used for each study. The former effect would be expected to lead to narrower regions being detected under selection in the resequencing dataset, which could shift the most significant SNP away from candidate genes detected from the genotyping data. Secondly, using groups of multiple breeds could mean that whilst haplotypes often remain similar amongst closely related breeds, certain point mutations that differentiate those breeds from each other may become diluted in frequency, and so, would not be considered under selection by this method when a small number of animals is used. This is likely the case where, lowland breeds show less signatures of selection than upland breeds, suggesting that they share less signatures when grouped. This postulates that Welsh lowland breeds are more diverse than the upland breeds, supported by data from Beynon et al. (2015), which could be in response to the demand for socioeconomic traits and lack of selection pressures in comparison to the upland breeds, where there seem to be a selective pressure on the same region, leading to shared signatures of selection. This further suggests that the lowland breeds have been selected for the production of socioeconomic traits, rather than adaptation to their local environment.

Lower costs of genome resequencing have allowed a deeper insight to the individual mutations that could have functional roles within a region under selection; however, this is not always a realistic approach when investigating many related breeds in a single study. Our method here has demonstrated reliability in using resequencing data from a small number of individuals of different breeds but applied to similar environments can be supported by genotyping many individuals of these breeds. This is truer with the upland Welsh breeds, which is most likely due to the higher environmental selective pressures applied when compared to the lowland breeds. This has greatly eluted candidate genes relating to hardiness and survivability in both the genotyping and resequencing data.

CONCLUSION

Here, we have seen the first investigation into signatures of selection in Welsh sheep breeds using a large number of genotyped individuals and a small number of whole-genome resequenced individuals. Statistical pipelines have shown selection in Welsh upland breeds in regions containing genes relating to adaptation to the local environment, including candidate genetic variants, as well as some genes related to the production of socioeconomic traits in the lowland breeds. In turn, this information is useful, not only for the conservation of these culturally important breeds, but also for the improved production capabilities of mountain breeds and adaptation of productive breeds through a marker-assisted selection.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are publicly available in NCBI using accession number PRJNA646642.

AUTHOR CONTRIBUTIONS

DL: leading the project, sample collection, and writing the manuscript. JS-J: running the analyses and drafting the manuscript. VL: genome sequencing of Welsh sheep samples and initial analysis. AY: analysis pipeline development. NY: sample collection and manuscript editing. MS: Welsh sheep sequencing and initial analysis. All authors edited the manuscript.

FUNDING

The collection and sequencing of Russian sheep samples used in this study were funded by the Russian Scientific Foundation grant 19-76-20026 to DL. Resources from HPC Wales and Supercomputing Wales supported the contributions of VL and MS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.612492/full#supplementary-material

REFERENCES

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7:Unit7.20. doi: 10.1002/0471142905.hg0720s76

Armstrong, E., Ciappesoni, G., Iriarte, W., Da Silva, C., Macedo, F., Navajas, E. A., et al. (2018). Novel genetic polymorphisms associated with carcass traits in grazing Texel sheep. *Meat Sci.* 145, 202–208. doi: 10.1016/j.meatsci.2018.06.014

Askew, E. W. (2002). Work at high altitude and oxidative stress: antioxidant nutrients. *Toxicology* 180, 107–119. doi: 10.1016/S0300-483X(02)00385-2

Barclay, A. N., and van den Berg, T. K. (2014). The interaction between signal regulatory protein alpha (SIRPα) and CD47: structure, function, and therapeutic target. *Annu. Rev. Immunol.* 32, 25–50. doi: 10.1146/annurev-immunol-032713-120142

Barker, J., and Bernstein, S. (1983). Hertwig's anemia: characterization of the stem cell defect. *Blood* 61, 765–769. doi: 10.1182/blood.v61.4.765.bloodjournal614765

Benavides, M. V., Sonstegard, T. S., Kemp, S., Mugambi, J. M., Gibson, J. P., Baker, R. L., et al. (2015). Identification of novel loci associated with gastrointestinal parasite resistance in a Red Maasai x Dorper backcross population. *PLoS One* 10:e0122797. doi: 10.1371/journal.pone.0122797

Bento, J. L., Palmer, N. D., Mychaleckyj, J. C., Lange, L. A., Langefeld, C. D., Rich, S. S., et al. (2004). Association of protein tyrosine phosphatase 1B gene polymorphisms with type 2 diabetes. *Diabetes* 53:3007. doi: 10.2337/diabetes.53.11.3007

Beynon, S. E., Slavov, G. T., Farré, M., Sunduimijid, B., Waddams, K., Davies, B., et al. (2015). Population structure and history of the Welsh sheep breeds determined by whole genome genotyping. *BMC Genet.* 16:65. doi: 10.1186/s12863-015-0216-x

Bochkov, Y. A., Watters, K., Ashraf, S., Griggs, T. F., Devries, M. K., Jackson, D. J., et al. (2015). Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. *Proc. Natl. Acad. Sci. U S A.* 112, 5485–5490. doi: 10.1073/pnas.1421178112

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., et al. (2010). Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186, 241–262. doi: 10.1534/genetics.104.117275

Booth, L., Roberts, J. L., and Dent, P. (2015). HSPA5/Dna K may be a useful target for human disease therapies. *DNA Cell Biol.* 34, 153–158. doi: 10.1089/dna.2015.2808

Bovo, S., Mazzoni, G., Bertolini, F., Schiavo, G., Galimberti, G., Gallo, M., et al. (2019). Genome-wide association studies for 30 haematological and blood clinical-biochemical traits in Large White pigs reveal genomic regions affecting intermediate phenotypes. *Sci. Rep.* 9:7003. doi: 10.1038/s41598-019-43297-1

Bowles, D. (2015). Recent advances in understanding the genetic resources of sheep breeds locally adapted to the UK uplands: opportunities they offer for sustainable productivity. *Front. Genet.* 6. doi: 10.3389/fgene.2015.00024

Bowles, D., Carson, A., and Isaac, P. (2014). Genetic distinctiveness of the Herdwick sheep breed and two other locally adapted hill breeds of the UK. *PLoS One* 9:e87823. doi: 10.1371/journal.pone.0087823

Chen, C. Y., Salomon, A. K., Caporizzo, M. A., Curry, S., Kelly, N. A., Bedi, K. C., et al. (2020). Depletion of vasohibin 1 speeds contraction and relaxation in failing human cardiomyocytes. *Circ. Res.* 127, e14–e27. doi: 10.1161/CIRCRESAHA.119.315947

Cheng, J. B., Jacquemin, E., Gerhardt, M., Nazer, H., Cresteil, D., Heubi, J. E., et al. (2003). Molecular genetics of 3β-hydroxy-δ5-c27-steroid oxidoreductase deficiency in 16 patients with loss of bile acid synthesis and liver disease. *J. Clin. Endocrinol. Metab.* 88, 1833–1841. doi: 10.1210/jc.2002-021580

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Davies, E. (1935). Sheep farming in Upland Wales. *Geography* 20, 97–111.

Dekker, N., Voorn-Brouwer, T., Verhoek, M., Wennekes, T., Narayan, R. S., Speijer, D., et al. (2011). The cytosolic β-glucosidase GBA3 does not influence type 1 Gaucher disease manifestation. *Blood Cells Mol. Dis.* 46, 19–26. doi: 10.1016/j.bcmd.2010.07.009

Delaneau, O., Marchini, J., and Zagury, J.-F. (2011). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181. doi: 10.1038/nmeth.1785

Elchebly, M., Payette, P., Michaliszyn, E., Cromlish, W., Collins, S., Loy, A. L., et al. (1999). Increased insulin sensitivity and obesity resistance in mice lacking the protein tyrosine phosphatase-1B gene. *Science* 283:1544. doi: 10.1126/science.283.5407.1544

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. (2013). Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics* 193, 929–941. doi: 10.1534/genetics.112.147231

Fariello, M.-I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., Consortium, I. S. G., et al. (2014). Selection signatures in worldwide sheep populations. *PLoS One* 9:e103813. doi: 10.1371/journal.pone.0103813

Fernandez-Marcos, P. J., and Auwerx, J. (2011). Regulation of PGC-1α, a nodal regulator of mitochondrial biogenesis. *Am. J. Clin. Nutr.* 93, 884S–890S. doi: 10.3945/ajcn.110.001917

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent selective sweeps in north american *Drosophila* melanogaster show signatures of soft sweeps. *PLoS Genet.* 11:e1005004. doi: 10.1371/journal.pgen.1005004

Giovannone, B., Lee, E., Laviola, L., Giorgino, F., Cleveland, K. A., and Smith, R. J. (2003). Two novel proteins that are linked to insulin-like growth factor (IGF-I) receptors by the Grb10 adapter and modulate IGF-I signalling. *J. Biol. Chem.* 278, 31564–31573. doi: 10.1074/JBC.M211572200

Grant, J. R., Arantes, A. S., Liao, X., and Stothard, P. (2011). In-depth annotation of SNPs arising from resequencing projects using NGS-SNP. *Bioinformatics* 27, 2300–2301. doi: 10.1093/bioinformatics/btr372

Grilz-Seger, G., Neuditschko, M., Ricard, A., Velie, B., Lindgren, G., Mesarič, M., et al. (2019). Genome-wide homozygosity patterns and evidence for selection in a set of European and Near Eastern horse breeds. *Genes (Basel)* 10:491. doi: 10.3390/genes10070491

Gu, X., Feng, C., Ma, L., Song, C., Wang, Y., Da, Y., et al. (2011). Genome-wide association study of body weight in chicken F2 resource population. *PLoS One* 6:e21872. doi: 10.1371/journal.pone.0021872

Heaton, M. P., Leymaster, K. A., Kalbfleisch, T. S., Kijas, J. W., Clarke, S. M., McEwan, J., et al. (2014). SNPs for parentage testing and traceability in globally diverse breeds of sheep. *PLoS One* 9:e94851. doi: 10.1371/journal.pone.0094851

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Huang, X., Otecko, N. O., Peng, M., Weng, Z., Li, W., Chen, J., et al. (2020). Genome-wide genetic structure and selection signatures for color in 10 traditional Chinese yellow-feathered chicken breeds. *BMC Genomics* 21:316. doi: 10.1186/s12864-020-6736-4

Issa, L., Mueller, K., Seufert, K., Kraemer, N., Rosenkotter, H., Ninnemann, O., et al. (2013). Clinical and cellular features in patients with primary autosomal recessive microcephaly and a novel CDK5RAP2 mutation. *Orphanet J. Rare Dis.* 8:59. doi: 10.1186/1750-1172-8-59

Kang, N., Chen, D., Wang, L., Duan, L., Liu, S., Tang, L., et al. (2008). Rwdd1, a thymus aging related molecule, is a new member of the intrinsically unstructured protein family. *Cell. Mol. Immunol.* 5, 333–339. doi: 10.1038/cmi.2008.41

Kärst, S., Cheng, R., Schmitt, A. O., Yang, H., de Villena, F. P. M., Palmer, A. A., et al. (2011). Genetic determinants for intramuscular fat content and water-holding capacity in mice selected for high muscle mass. *Mamm. Genome* 22, 530–543. doi: 10.1007/s00335-011-9342-6

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., et al. (2012). Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol.* 10:e1001258. doi: 10.1371/journal.pbio.1001258

Kim, E.-S., Elbeltagy, A. R., Aboul-Naga, A. M., Rischkowsky, B., Sayre, B., Mwacharo, J. M., et al. (2016). Multiple genomic signatures of selection in goats and sheep indigenous to a hot arid environment. *Heredity (Edinb)* 116, 255–264. doi: 10.1038/hdy.2015.94

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D.-A., Mitterecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40:e69. doi: 10.1093/nar/gks003

Kobayashi, M., Oshima, S., Maeyashiki, C., Nibe, Y., Otsubo, K., Matsuzawa, Y., et al. (2016). The ubiquitin hybrid gene UBA52 regulates ubiquitination of ribosome and sustains embryonic development. *Sci. Rep.* 6:36780. doi: 10.1038/srep36780

Kobayashi, Y., Ohshiro, N., Sakai, R., Ohbayashi, M., Kohyama, N., and Yamamoto, T. (2005). Transport mechanism and substrate specificity of human organic anion transporter 2 (hOat2 [SLC22A7]). *J. Pharm. Pharmacol.* 57, 573–578. doi: 10.1211/0022357055966

Li, C., Sun, D., Zhang, S., Yang, S., Alim, M. A., Zhang, Q., et al. (2016). Genetic effects of FASN, PPARGC1A, ABCG2 and IGF1 revealing the association with milk fatty acids in a Chinese Holstein cattle population based on a post genome-wide association study. *BMC Genet.* 17:110. doi: 10.1186/s12863-016-0418-x

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [preprint] arXiv:1303.3997

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

Li, X., Kim, S.-W., Do, K.-T., Ha, Y.-K., Lee, Y.-M., Yoon, S.-H., et al. (2011). Analyses of porcine public SNPs in coding-gene regions by re-sequencing and phenotypic association studies. *Mol. Biol. Rep.* 38, 3805–3820. doi: 10.1007/s11033-010-0496-1

Li, X., Liu, X., Qin, Z., Wei, M., Hou, X., Zhang, T., et al. (2018). A novel transcription factor Rwdd1 and its SUMOylation inhibit the expression of sqr, a key gene of mitochondrial sulfide metabolism in Urechis unicinctus. *Aquat. Toxicol.* 204, 180–189. doi: 10.1016/j.aquatox.2018.09.012

Li, X., Yang, J., Shen, M., Xie, X.-L., Liu, G.-J., Xu, Y.-X., et al. (2020). Whole-genome resequencing of wild and domestic sheep identifies genes associated with morphological and agronomic traits. *Nat. Commun.* 11:2815. doi: 10.1038/s41467-020-16485-1

Lizarraga, S. B., Margossian, S. P., Harris, M. H., Campagna, D. R., Han, A.-P., Blevins, S., et al. (2010). Cdk5rap2 regulates centrosome function and chromosome segregation in neuronal progenitors. *Development* 137, 1907–1917. doi: 10.1242/dev.040410

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q., and Simianer, H. (2015). Properties of different selection signature statistics and a new strategy for combining them. *Heredity (Edinb)* 115, 426–436. doi: 10.1038/hdy.2015.42

Marnetto, D., and Huerta-Sánchez, E. (2017). Haplostrips: revealing population structure through haplotype visualization. *Methods Ecol. Evol.* 8, 1389–1392. doi: 10.1111/2041-210X.12747

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, 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

Miyashita, H., Watanabe, T., Hayashi, H., Suzuki, Y., Nakamura, T., Ito, S., et al. (2012). Angiogenesis inhibitor vasohibin-1 enhances stress resistance of endothelial cells via induction of SOD_2 and $SIRT_1$. *PLoS ONE* 7, e46459–e46459. doi: 10.1371/journal.pone.0046459

Naval-Sanchez, M., Nguyen, Q., McWilliam, S., Porto-Neto, L. R., Tellam, R., Vuocolo, T., et al. (2018). Sheep genome functional annotation reveals proximal regulatory elements contributed to the evolution of modern breeds. *Nat. Commun.* 9:859. doi: 10.1038/s41467-017-02809-1

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* 76, 5269–5273. doi: 10.1073/pnas.76.10.5269

Olivier, M., Hsiung, C. A., Chuang, L.-M., Ho, L.-T., Ting, C.-T., Bustos, V. I., et al. (2004). Single nucleotide polymorphisms in protein tyrosine phosphatase 1beta (PTPN1) are associated with essential hypertension and obesity. *Hum. Mol. Genet.* 13, 1885–1892. doi: 10.1093/hmg/ddh196

Pasiakos, S., Berryman, C., Carrigan, C., Young, A., and Carbone, J. (2017). Muscle protein turnover and the molecular regulation of muscle mass during hypoxia. *Med. Sci. Sport. Exerc.* 49, 1340–1350. doi: 10.1249/mss.0000000000001228

Periasamy, K., Pichler, R., Poli, M., Cristel, S., Cetrá, B., Medus, D., et al. (2014). Candidate gene approach for parasite resistance in sheep–variation in immune pathway genes and association with fecal egg count. *PLoS One* 9:e88337. doi: 10.1371/journal.pone.0088337

Prades, C., Arnould, I., Annilo, T., Shulenin, S., Chen, Z. Q., Orosco, L., et al. (2002). The human ATP binding cassette gene ABCA13, located on chromosome 7p12.3, encodes a 5058 amino acid protein with an extracellular domain encoded in part by a 4.8-kb conserved exon. *Cytogenet. Genome Res.* 98, 160–168. doi: 10.1159/000069852

Pszczola, M., Strabel, T., Mucha, S., and Sell-Kubiak, E. (2018). Genome-wide association identifies methane production level relation to genetic control of digestive tract development in dairy cows. *Sci. Rep.* 8:15164. doi: 10.1038/s41598-018-33327-9

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Rahman, M. M., Bagdassarian, E., Ali, M. A. M., and McFadden, G. (2017). Identification of host DEAD-box RNA helicases that regulate cellular tropism of oncolytic Myxoma virus in human cancer cells. *Sci. Rep.* 7:15710. doi: 10.1038/s41598-017-15941-1

Ryder, M. (1964). The history of sheep breeds in Britain. *Agric. Hist. Rev.* 12, 1–12. doi: 10.2307/40273081

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi: 10.1086/502802

Stronen, A. V., Pertoldi, C., Iacolina, L., Kadarmideen, H. N., and Kristensen, T. N. (2019). Genomic analyses suggest adaptive differentiation of northern European native cattle breeds. *Evol. Appl.* 12, 1096–1113. doi: 10.1111/eva.12783

Sweet-Jones, J., Yurchenko, A. A., Igoshin, A. V., Yudin, N. S., Swain, M. T., and Larkin, D. M. (2021). Resequencing and signatures of selection scan in two Siberian native sheep breeds point to candidate genetic variants for adaptation and economically important traits. *Anim. Genet.* 52, 126–131. doi: 10.1111/age.13015

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595. doi: 10.1093/genetics/123.3.585

Takasuga, A., Sato, K., Nakamura, R., Saito, Y., Sasaki, S., Tsuji, T., et al. (2015). Non-synonymous FGD3 variant as positional candidate for disproportional tall stature accounting for a carcass weight QTL (CW-3) and skeletal dysplasia in Japanese Black cattle. *PLoS Genet.* 11:e1005433. doi: 10.1371/journal.pgen.1005433

Takeda, E., Suzuki, Y., and Sato, Y. (2016). Age-associated downregulation of vasohibin-1 in vascular endothelial cells. *Aging Cell* 15, 885–892. doi: 10.1111/acel.12497

Tollenaere, M. A. X., Tiedje, C., Rasmussen, S., Nielsen, J. C., Vind, A. C., Blasius, M., et al. (2019). GIGYF1/2-driven cooperation between ZNF598 and TTP in posttranscriptional regulation of inflammatory signaling. *Cell Rep.* 26, 3511–3521.e4. doi: 10.1016/j.celrep.2019.03.006

Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv* [preprint] doi: 10.1101/005165 bioRxiv 005165.

Wang, W., Zhang, X., Zhou, X., Zhang, Y., La, Y., Zhang, Y., et al. (2019). Deep Genome resequencing reveals artificial and natural selection for visual deterioration, plateau adaptability and high prolificacy in Chinese domestic sheep. *Front. Genet.* 10:300. doi: 10.3389/fgene.2019.00300

Weatherhead, B., and Logan, A. N. N. (1981). Interaction of α-melanocyte-stimulating hormone, melatonin, cyclin AMP and cyclic GMP in the control of melanogenesis in hair follicle melanocytes in vitro. *J. Endocrinol.* 90, 89–96. doi: 10.1677/joe.0.0900089

Wei, C., Wang, H., Liu, G., Wu, M., Cao, J., Liu, Z., et al. (2015). Genome-wide analysis reveals population structure and selection in Chinese indigenous sheep breeds. *BMC Genomics* 16:194. doi: 10.1186/s12864-015-1384-9

Weir, B. S., and Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution (N. Y)* 38, 1358–1370. doi: 10.2307/2408641

Williams-Davies, J. (1981). *Welsh Sheep and Their Wool*, 1st Edn. Dyfed: Gormer Press.

Yan, K., Gao, L.-N., Cui, Y.-L., Zhang, Y., and Zhou, X. (2016). The cyclic AMP signalling pathway: Exploring targets for successful drug discovery (Review). *Mol. Med. Rep.* 13, 3715–3723. doi: 10.3892/mmr.2016.5005

Yang, G.-L., Fu, D.-L., Lang, X., Wang, Y.-T., Cheng, S.-R., Fang, S.-L., et al. (2013). Mutations in MC1R gene determine black coat color phenotype in Chinese sheep. *Sci. World J.* 2013:675382. doi: 10.1155/2013/675382

Yang, J., Li, W.-R., Lv, F.-H., He, S.-G., Tian, S.-L., Peng, W.-F., et al. (2016). Whole-Genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* 33, 2576–2592. doi: 10.1093/molbev/msw129

Yuan, Z., Li, W., Li, F., and Yue, X. (2019). Selection signature analysis reveals genes underlying sheep milking performance. *Arch. Anim. Breed.* 62, 501–508. doi: 10.5194/aab-62-501-2019

Yurchenko, A. A., Deniskova, T. E., Yudin, N. S., Dotsev, A. V., Khamiruev, T. N., Selionova, M. I., et al. (2019). High-density genotyping reveals signatures of selection related to acclimation and economically important traits in 15 local sheep breeds from Russia. *BMC Genomics* 20:294. doi: 10.1186/s12864-019-5537-0

Zeder, M. A. (2008). Domestication and early agriculture in the Mediterranean Basin: origins, diffusion, and impact. *Proc. Natl. Acad. Sci.* 105, 11597–11604. doi: 10.1073/pnas.0801317105

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 22 June 2021
doi: 10.3389/fgene.2021.610774

Disruptive Selection of Human Immunostimulatory and Immunosuppressive Genes Both Provokes and Prevents Rheumatoid Arthritis, Respectively, as a Self-Domestication Syndrome

Natalya V. Klimova[1], Evgeniya Oshchepkova[1], Irina Chadaeva[1], Ekaterina Sharypova[1], Petr Ponomarenko[1], Irina Drachkova[1], Dmitry Rasskazov[1], Dmitry Oshchepkov[1], Mikhail Ponomarenko[1,2], Ludmila Savinkova[1], Nikolay A. Kolchanov[1] and Vladimir Kozlov[2]*

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (ICG SB RAS), Novosibirsk, Russia, [2] Research Institute of Fundamental and Clinical Immunology (RIFCI SB RAS), Novosibirsk, Russia

OPEN ACCESS

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

Reviewed by:
Ranajit Das,
Yenepoya University, India
Elvira Galieva,
Novosibirsk State University, Russia
Anatoliy Ivashchenko,
Al-Farabi Kazakh National University, Kazakhstan

***Correspondence:**
Mikhail Ponomarenko
pon@bionet.nsc.ru

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

Received: 27 September 2020
Accepted: 17 May 2021
Published: 22 June 2021

Citation:
Klimova NV, Oshchepkova E, Chadaeva I, Sharypova E, Ponomarenko P, Drachkova I, Rasskazov D, Oshchepkov D, Ponomarenko M, Savinkova L, Kolchanov NA and Kozlov V (2021) Disruptive Selection of Human Immunostimulatory and Immunosuppressive Genes Both Provokes and Prevents Rheumatoid Arthritis, Respectively, as a Self-Domestication Syndrome. Front. Genet. 12:610774. doi: 10.3389/fgene.2021.610774

Using our previously published Web service SNP_TATA_Comparator, we conducted a genome-wide study of single-nucleotide polymorphisms (SNPs) within core promoters of 68 human rheumatoid arthritis (RA)-related genes. Using 603 SNPs within 25 genes clinically associated with RA-comorbid disorders, we predicted 84 and 70 candidate SNP markers for overexpression and underexpression of these genes, respectively, among which 58 and 96 candidate SNP markers, respectively, can relieve and worsen RA as if there is a neutral drift toward susceptibility to RA. Similarly, we predicted natural selection toward susceptibility to RA for 8 immunostimulatory genes (e.g., *IL9R*) and 10 genes most often associated with RA (e.g., *NPY*). On the contrary, using 25 immunosuppressive genes, we predicted 70 and 109 candidate SNP markers aggravating and relieving RA, respectively (e.g., *IL1R2* and *TGFB2*), suggesting that natural selection can simultaneously additionally yield resistance to RA. We concluded that disruptive natural selection of human immunostimulatory and immunosuppressive genes is concurrently elevating and reducing the risk of RA, respectively. So, we hypothesize that RA in human could be a self-domestication syndrome referring to evolution patterns in domestic animals. We tested this hypothesis by means of public RNA-Seq data on 1740 differentially expressed genes (DEGs) of pets vs. wild animals (e.g., dogs vs. wolves). The number of DEGs in the domestic animals corresponding to worsened RA condition in humans was significantly larger than that in the related wild animals (10 vs. 3). Moreover, much less DEGs in the domestic animals were accordant to relieved RA condition in humans than those in the wild animals (1 vs. 8 genes). This indicates that the anthropogenic environment, in contrast to a natural one, affects gene expression across the whole genome (e.g., immunostimulatory and immunosuppressive genes) in a manner that likely contributes to RA. The difference in gene numbers is

statistically significant as confirmed by binomial distribution ($p < 0.01$), Pearson's χ^2 ($p < 0.01$), and Fisher's exact test ($p < 0.05$). This allows us to propose RA as a candidate symptom within a self-domestication syndrome. Such syndrome might be considered as a human's payment with health for the benefits received during evolution.

Keywords: rheumatoid arthritis, human, gene, promoter, TATA box, candidate SNP marker, self-domestication syndrome, RNA-seq verification

INTRODUCTION

Rheumatoid arthritis (RA) is an autoimmune disease involving autoantibodies (e.g., anti-citrullinated protein antibodies) and proinflammatory cytokines (e.g., TNF-α and IL-6) that participate in the induction of chronic synovitis and bone erosion, followed by deformity (Smolen and Aletaha, 2015), which are some of the most prevalent causes of disability (Koller and Nobauer-Huhmann, 2009). Currently, it is widely accepted that RA immunopathogenesis is mostly mediated by the mechanisms involving a breakdown of immune tolerance to self antigens that is characterized by an increase in the activity of effector T cells causing RA symptoms (Sarkander et al., 2016). Two subpopulations of helper cells, Th1 and Th17, are mostly responsible for the increase in the activity of effector T cells (Frey et al., 2010). Additionally, memory cells of adaptive immunity in a given individual keep information on all the diseases survived; these cells will increase resistance to these diseases in the future (Sarkander et al., 2016). On the contrary, low activity of regulatory cells [e.g., regulatory T cells (Tregs) and myeloid suppressor cells] is often seen in RA (Alsaed et al., 2018). The immunosuppressive-activity deficit is one of the central features of RA pathogenesis (Lawson et al., 2006). Moreover, in patients with RA, an increase in the resistance of effector T cells to the suppressive action of Tregs is detectable too (Malemud, 2018). There are numerous factors related to disturbances in the mechanisms regulating immunocompetent cells, e.g., primarily, cytokines, and their receptors as well as transcription factors.

Rheumatoid arthritis cannot be completely cured (Smolen et al., 2016) in terms of autoimmunity because memory immune cells in the blood are able to retain information about antigens for a long time after successful treatment (Sarkander et al., 2016). For this reason, clinicians usually talk only about RA remission (Smolen and Aletaha, 2015). Because the immune system is more potent in women, their risk of RA is three times higher than that of men (Krasselt and Baerwald, 2017) and increases with the level of sex hormones after menopause and during pregnancy (Ho and Weinshilboum, 2017) but decreases during lactation (Karlson et al., 2004). A retrospective clinical and pharmacological meta-analysis of 29,880 RA patients in comparison with 73,758 relatively healthy volunteers identified 42 loci and 98 genes as candidate therapeutic targets in RA within the framework of predictive, preventive personalized, and participatory (4P) medicine (Okada et al., 2014).

Conventionally, RA risk is dependent on genetic factors and the lifestyle approximately equally (Nair et al., 2017), namely, the microbiome (reflecting a diet) (Sato et al., 2017), previous illnesses (Scott et al., 2010), environmental pollution (Klareskog et al., 2006), and bad habits (Malm et al., 2016) such as smoking (Erlandsson et al., 2016), physical inactivity, and overeating (Somers et al., 2014). That is why RA fits well the main idea of post-genomic 4P medicine, thus giving a chance to people to reduce their disease risks by correction of the lifestyle in line with their individual sequenced genomes (Trovato, 2014). This is important because at late RA stages, fibroblast-like synoviocytes are capable of hyperproliferation, which can cause leukemia in children with leukopenia (Jones et al., 2006) and, in adulthood, may lead to synovial hyperplasia as an impairment of the shape and mobility of joints (Lim and Bae, 2011).

The keystone of 4P medicine is the top scientific project of the 21st century, "1000 Genomes" (Telenti et al., 2016), due to which hundreds of thousands of individual human genomes are sequenced referenced in the variome data. This Big Data set contains consensus human genome sequences and hundreds of millions of single-nucleotide polymorphisms (SNPs) publicly available within databases Ensembl (Zerbino et al., 2015), the UCSC Genome Browser (Haeussler et al., 2015), and dbSNP (Day, 2010). Additionally, databases ClinVar (Landrum et al., 2014) and OMIM (Amberger et al., 2015) document, systematize, and prioritize only clinically proven and experimentally studied human disease SNP markers, respectively, whose allele frequencies significantly differ between cohorts of patients and conventionally healthy volunteers as a mandatory criterion (Varzari et al., 2019). Finally, the dbWGFP database (Wu et al., 2016) does the same in whole-genome mode for all the 10 billion potential SNPs in humans. If we assume that each SNP can affect at least 1 of known human 55,000 diseases (Pocai, 2019), such genetic load will be too high to survive in evolution. So, both Kimura's theory (1968) and Haldane's dilemma (1957) lead to the conclusion about neutrality of the vast majority of human SNPs. These neutral SNPs might be bioinformatically identified and discarded without time-consuming clinical testing. Although the current accuracy of bioinformatic calculations is not above the threshold of clinical applicability yet (Yoo et al., 2015), this accuracy grows each year (Putlyaeva et al., 2018; Zorlu et al., 2019).

Most of the SNPs documented in the OMIM database (Amberger et al., 2015) are within protein-coding regions of human genes and correspond to aberrations in protein structure and therefore function (Mitsuyasu et al., 1998). Indeed, these damages are uniform within any tissue and thus are easily detectable but cannot be corrected either therapeutically or *via* lifestyle changes. In contrast, SNPs within regulatory gene regions (Deplancke et al., 2016) have pathogenic manifestations

correctable both by medication and by lifestyle changes within the framework of 4P medicine because these manifestations are limited to alterations of gene expression levels without any protein damage; the latter is negligibly rare among experimentally studied regulatory SNPs (Amberger et al., 2015). Certainly, many factors can independently modulate expression levels of the majority of genes; this situation complicates interpreting the expression patterns of these genes as partial contributions of such modulators as SNPs, somatic mutations, stressors, silencers, inhibitors, and activators. Actually, exogenous recombinant activated coagulation factor VII (*F7*), as an adjunctive therapy, can successfully help to urgently stop internal bleeding caused by acquired hemophilia as an autoimmune complication of RA, which is treated with immunosuppressive therapy at the same time (Drobiecki et al., 2013). The best-studied regulatory SNPs are mostly within 70-bp regions upstream of transcription start sites (Bhuiyan and Timmers, 2019) and affect gene expression levels proportionally with effect of these SNPs on the binding affinity of TATA-binding protein (TBP) for TBP-sites in these promoter regions (Mogno et al., 2010) when TATA box is canonical (Ponomarenko et al., 2013). According to the EPD database (Dreos et al., 2017), only ∼15% of eukaryotic gene promoters contain canonical TATA boxes as TBP-sites (Bucher, 1990), whereas genome-wide chromatin immunoprecipitation experiments (ChIP-seq) have detected such sites upstream of all the transcription start sites within eukaryotic genomes (Rhee and Pugh, 2012). Indeed, the binding of TBP to TBP-sites of genes shifts the equilibrium from transcriptionally inactive packing of these genes to pre-initiation complexes necessary to initiate the expression of these genes (Godde et al., 1995) as proven in TBP knockout mice (Martianov et al., 2002).

We developed our Web service SNP_TATA_Comparator[1] (Ponomarenko et al., 2015), whose input consists of two promoter DNA sequences representing ancestral and minor alleles of the SNP being examined; the software generates TBP-binding affinity estimates for these promoter alleles (± standard error) and significance α of their difference with Fisher's *Z*-test (Waardenberg et al., 2015). We applied it from SNP to SNP to predict their contribution to diseases [e.g., chronopathologies (Ponomarenko et al., 2016)] and selectively verified the obtained results using F1-hybrid mice (Chadaeva I. et al., 2019), real-time polymerase chain reaction (Oshchepkov et al., 2019), RNA-Seq data (Vasiliev et al., 2021), gel retardation assay, stopped-flow spectrometry, biosensors, or bioluminescence, as reviewed (Ponomarenko et al., 2017). SNP_TATA_Comparator (Ponomarenko et al., 2015) is already used in independent clinical studies [e.g., in a pulmonary tuberculosis case-control study (Varzari et al., 2018)]. In the present work, at the stage of comprehensive experimental validation, we tested whole-genome sequence-based predictions for RA-associated candidate biomedical SNP markers *in vivo* using publicly available RNA-Seq data (Albert et al., 2012; Hekman et al., 2018; Yang et al., 2018). Finally, we discuss the results of the verification of the SNP_TATA_Comparator predictions vis-à-vis the semiquantitative RNA-Seq data for the next step: further comprehensive experimental verification of our biomedical predictions using SNP_TATA_Comparator compared with genome-wide data on quantitative trait loci, QTLs [e.g., in human cardiopathology (Koopmann et al., 2014)].

[1]http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan/start.pl

MATERIALS AND METHODS

DNA Sequences

We retrieved 1896 SNPs of 68 human genes from the dbSNP database, build No. 151, and DNA sequences from Ensembl (reference genome assembly GRCh38/hg38) using the UCSC Genome Browser.

Analysis of DNA Sequences

We used the web tool SNP_TATA_Comparator (Ponomarenko et al., 2015), the input of which are two proximal promoter sequences carrying either an ancestral (wt) or a minor (min) allele of an SNP being analyzed, as shown within two textboxes, "Basic sequence" and "Editable sequence," respectively (**Figure 1E**). The double-headed open arrows (⇔) between **Figures 1B,C,E** explain how SNP_TATA_Comparator uses the Bioperl toolkit (Stajich et al., 2002) to retrieve the ancestral variant of the DNA sequence of the human *MMP12* promoter from database Ensembl (Zerbino et al., 2015) in an automated mode. The solid arrows between **Figures 1A,D,E** show the input of a minor variant of this sequence into SNP_TATA_Comparator according to its description within the database dbSNP (Day, 2010) visualized using the UCSC Genome Browser (Haeussler et al., 2015).

Using the "Calculate" option, we first estimated two ($-\ln(K_D) \pm \delta$) pairwise value sets of the highest observed estimate of TBP-promoter affinity ± its standard error according to our three-step approximation of their complex formation on each of these sequences independently, as depicted in **Figure 1F** and described in depth in **Supplementary File 1** entitled Section "Supplementary DNA Sequence Analysis." To this end, we took into account non-specific TBP-DNA affinity (Hahn et al., 1989), the position-weight matrix of TBP-sites (Bucher, 1990), minor-groove width of B-helical DNA (Karas et al., 1996), DNA melting during its bending, which fixes the TBP-promoter complex (Flatters and Lavery, 1998), and abundance levels of TA-rich dinucleotides (Ponomarenko et al., 1999) in the sequences analyzed. Then, we calculated Fisher's *Z*-score and, finally, converted it into its *p*-value of statistical significance taken from standard software R (Waardenberg et al., 2015), as one can see in **Figure 1F**. Eventually, the "Result" textbox (**Figure 1E**) shows all the intermediate and final results, namely, $-\ln(K_D{}^{(wt)}) \pm \delta_{(wt)}$ and $-\ln(K_D{}^{(min)}) \pm \delta_{(min)}$ in lines 1 and 2, respectively; the prediction made using the terms "deficiency," "excess," "significant," and "insignificant" in line 3 as proven experimentally (Mogno et al., 2010); and the *Z*-score and *p*-value in line 4.

Thus, we examined the SNPs one by one independently from the others and, as a result, either discarded those with insignificant effects on the TBP-promoter binding affinity according to our predictions (data not shown) or presented

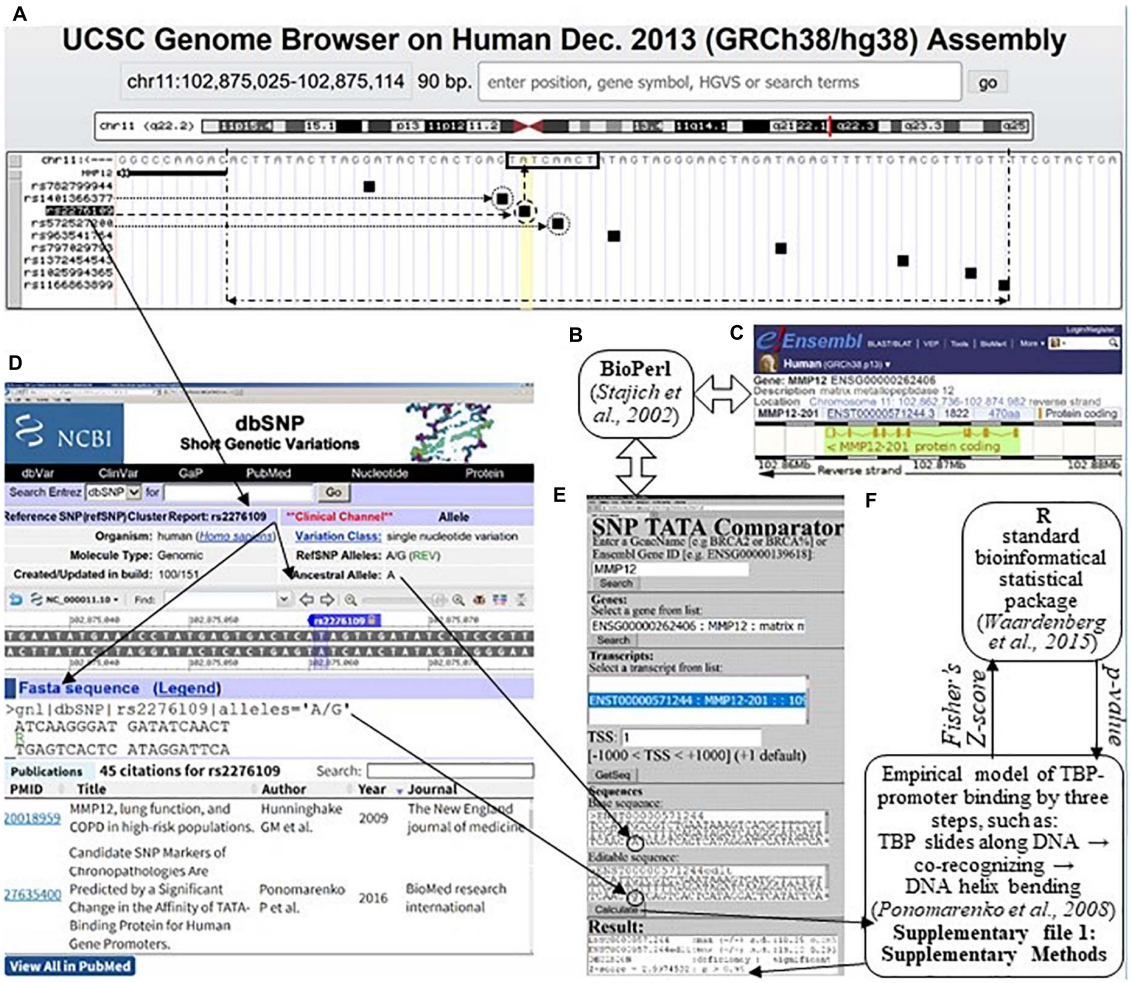


FIGURE 1 | The result produced by SNP_TATA_Comparator (Ponomarenko et al., 2015) in the case of a clinically proven SNP marker (rs2276109) of an asthma risk reduction due to MMP12 downregulation (Hunninghake et al., 2009), which is associated with a reduced risk of rheumatoid arthritis, RA (Liu et al., 2004). **(A)** The UCSC Genome Browser (Haeussler et al., 2015) visualizes a 70-bp promoter (double-headed dash-and-dot arrow) of a given human gene (here: MMP12, row 2, left column), where there are a TBP-binding site (TBP-site; framed, □) and SNPs retrieved from dbSNP (Day, 2010); their IDs constitute the left-hand column. **(B)** The Bioperl toolkit (Stajich et al., 2002). **(C)** The Ensembl database (Zerbino et al., 2015). **(D)** The current build (No. 151) of the dbSNP database describes the SNP under study (rs2276109 in this example). **(E)** Our prediction of MMP12 downregulation (textbox "Result": row 3) caused by SNP rs2276109 (circled) by means of its both alleles (i.e., ancestral and minor ones as input data within two textboxes "Basic sequence" and "Editable sequence," respectively). Double-headed open arrows (⇔) depict how SNP_TATA_Comparator retrieved the ancestral DNA sequence of the human *MMP12* promoter from the Ensembl database (Zerbino et al., 2015) using the Bioperl toolkit (Stajich et al., 2002). **(F)** Our bioinformatics model of the TBP-promoter binding *via* three steps, as follows: (1) TBP sliding along DNA ↔ (2) TBP retarded by the TBP-site ↔ the TBP-promoter complex is fixed due to the bending of the DNA double helix at right angles (Ponomarenko et al., 2008). This model is based on the standard bioinformatical package of R (Waardenberg et al., 2015), as described in **Supplementary File 1: Supplementary Methods**. Solid arrows: data flows, when one predicts the manifestation of a given SNP using SNP_TATA_Comparator; dashed arrow: the SNP in this example (rs2276109); dotted arrows: two more SNPs, rs572527200 and rs1401366377 (**Figure 3A**), which correspond to either underexpression or overexpression of *MMP12*, respectively, as predicted in this work.

the SNPs in **Supplementary Tables 1–4** (hereinafter: see **Supplementary File 2** section "**Supplementary Results**").

Keyword Searches in the PubMed Database

For each SNP that can statistically significantly alter the expression of the studied human genes according to our prediction above, we handmade a standard keyword search in the PubMed database (Lu, 2011) as depicted in **Supplementary Figure 1** (hereinafter: see **Supplementary File 3** entitled section "**Supplementary Keyword Search**").

In vivo Validation of Our Current Predictions Using Public RNA-Seq Data

For *in vivo* validation of our predictions made in this work (**Supplementary Tables 1–4**) using public RNA-Seq data, we compiled all the 68 human genes whose effects on RA in humans have been estimated here, as shown in

Figure 2 (Step-1), and we present them in **Supplementary Table 5** (hereinafter, see **Supplementary File 2** entitled section "**Supplementary Results**").

We verified our computer-based prediction (that RA can be a self-domestication syndrome) by means of previously published publicly available whole-genome RNA-Seq data, as described in **Table 1** and depicted in **Figure 2** (Step-2). For minimizing the effects of false–positive errors, we selected only the differentially expressed genes (DEGs) that were statistically significant according to Fisher's *Z*-test, with those corrections on multiple comparisons ($P_{ADJ} < 0.05$) that have been published by their authors (Albert et al., 2012; Hekman et al., 2018;

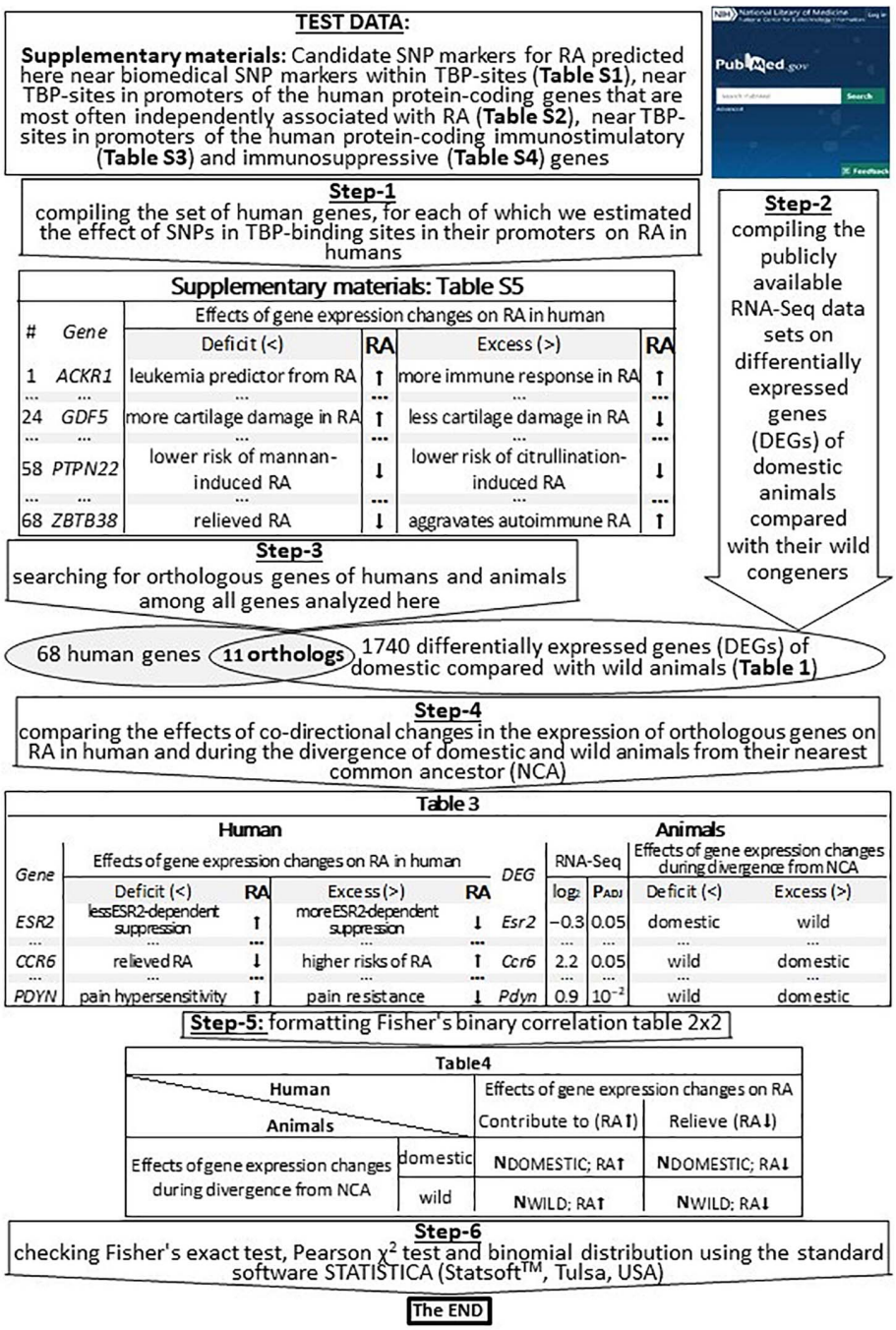


FIGURE 2 | An algorithmic flowchart of *in vivo* validation of our current predictions by means of publicly available RNA-Seq data on differentially expressed genes (DEGs) of domestic vs. wild animals. RA, rheumatoid arthritis; effect on RA, contributes to (↑), relieves (↓); NCA, nearest common ancestor; P_{ADJ}, significance [Fisher's *Z*-test with those corrections on multiple comparisons that have been published by Albert et al. (2012), Hekman et al. (2018), and Yang et al. (2018)].

Yang et al., 2018). Therefore, these data included DEGs in the frontal cortex of guinea pigs (*Cavia porcellus*) vs. cavies (*C. aperea*; 883 DEGs), pigs vs. boars (*Sus scrofa*; 30 DEGs), domesticated vs. wild rabbits (*Oryctolagus cuniculus*; 17 DEGs), tame vs. aggressive rats (*Rattus norvegicus*; 20 DEGs), and dogs (*Canis familiaris*) vs. wolves (*C. lupus*; 13 DEGs); these data were retrieved from another study (Albert et al., 2012). Additionally, we used 450 DEGs in the blood of dogs vs. wolves retrieved from Yang et al. (2018). Besides, we analyzed 327 DEGs within anterior pituitary tissues of adult male foxes (*Vulpes vulpes*) of two unique outbred lines artificially selected for domestication (tameness) or aggressiveness (Hekman et al., 2018). Thus, the total number of DEGs analyzed in this study was 1740.

For the RNA-Seq data analysis, we employed the conventional concept of "divergence from the nearest common ancestor" (Samet, 1985), which we applied to domestic and wild animals whose DEGs were compared with orthologous human genes studied here, as suggested recently (Vasiliev et al., 2021). This procedure yielded 11 pairs of orthologous genes in humans and animals, as shown in **Figure 2** (Step-3 and a Venn diagram). In this context, we compared the effects of codirected changes in the expression of orthologous genes on RA in humans and during the divergence of domestic and wild animals from their NEAREST COMMON ANCESTOR, as depicted in **Figure 2** (Step-4). Accordingly, we filled out Fisher's table 2 × 2 (**Figure 2**: Step-5) and, finally, tested this binary correlation table for statistical significance using standard package STATISTICA (StatSoft^{TM}, Tulsa, OK, United States), as depicted in **Figure 2** (Step-6).

Statistical Analysis

Using options "Statistics" → "Non-parametrics" → "2 × 2 Table" in STATISTICA (StatSoft, Tulsa, OK, United States), we verified our predictions (about the candidate SNP markers of worse RA as a self-domestication syndrome) with the 1740 DEGs of domestic vs. wild animals (**Table 1**) by three criteria, namely, binomial distribution, Pearson's χ^2, and Fisher's exact test.

RESULTS AND DISCUSSION

To test our Web service SNP_TATA_Comparator for robustness vis-à-vis the annual increase in the SNP count, we compared its predictions between two cases, namely, (1) a previous build (No. 147) of the dbSNP database dated 2016, and after that, (2) its current build (No. 151) dated 2017. As one can see in rows 3 and 4 of **Table 2**, during the period considered, the progress of biomedicine yielded a 1.5-fold increase (from 18 to 27) in the number of human genes with biomedical SNP markers in known TBP-sites and an almost threefold increase (from 225 to 646) in the number of SNPs within the gene regions in question. That is why here we first analyzed SNPs within only the human genes that were analyzed in our previous study on RA (Chadaeva I. V. et al., 2019), to compare the results obtained using the current build (No. 151) of database dbSNP (Day, 2010) with those obtained by means of the previous build (No. 147) of this database. Below, we present the results on some genes.

Candidate SNP Markers of RA Near Clinical SNP Markers of Diseases at TBP-Sites

The human *MMP12* gene encodes macrophage elastase and carries an SNP, rs2276109. Rs2276109 is a clinically proven SNP marker of an asthma risk reduction due to MMP12 deficiency (Hunninghake et al., 2009), as one can see in **Figures 1A,D**. **Figure 1E** presents how, using SNP_TATA_Comparator (Ponomarenko et al., 2015), we predicted that this SNP causes MMP12 deficiency (i.e., the "Decision" line of the "Result" textbox) because this SNP damages the TBP-site within the promoter of this gene, as shown by the dashed arrows in **Figure 1A**. This match between our prediction and the clinical data in question (Hunninghake et al., 2009) indicates suitability of SNP_TATA_Comparator for biomedical studies *in silico*, as highlighted in bold in the first row of **Supplementary Table 1**. In the two rightmost columns of this table, just below the citation of these clinical data, readers can see other data on *MMP12*

TABLE 1 | The investigated genome-wide RNA-Seq transcriptomes of domestic animals with their wild congeners publicly available in database PubMed.

No.	Domestic animals	Wild animals	Number of DEGs	Tissue	References
1	Guinea pigs (*Cavia porcellus*): three females and three males	Cavy (*C. aperea*): three females and three males	883	Frontal cortex	Albert et al., 2012
2	Pigs (*Sus scrofa*): five females	Boars (*S. scrofa*): five females	30		
3	Domesticated rabbits (*Oryctolagus cuniculus domesticus*): three females and 3 males	Wild rabbits (*Oryctolagus cuniculus*): three females and three males	17		
4	Tame rats (*Rattus norvegicus*): three females and three males	Aggressive rats (*R. norvegicus*): three females and three males	20		
5	Dogs (*C. familiaris*): three females and three males	Wolves (*C. lupus*): three females and one male	13		
6	Dogs (*Canis familiaris*): one female and one male	Wolves (*C. lupus*): two females and one male	450	Blood	Yang et al., 2018
7	Tame foxes (*Vulpes vulpes*): six males	Aggressive foxes (*V. vulpes*): six males	327	Pituitary	Hekman et al., 2018
Total	Six domestic animal species: 17 females and 19 males	Six wild animal species: 18 females and 17 males	1740	3 tissues	

downregulation as a clinically proven physiological marker of RA (Liu et al., 2004); we found these data by means of keywords in the PubMed database. This finding allows us to predict that the clinical SNP marker (rs2276109) of the reduced risk of asthma is a candidate SNP marker of a reduced risk of RA too, as we highlighted in *italics* in **Supplementary Table 1**.

As one can see in **Figure 1A**, there are eight more SNPs within the analyzed 70-bp proximal promoter region depicted by a double-headed dash-and-dot arrow; two of them can either decrease (rs572527200) or increase (rs1401366377) *MMP12* expression (dotted arrows and circles) according to our predictions, as exemplified by **Figure 3A**. With this in mind, on the basis of the same clinical data (Liu et al., 2004), we predicted two more candidate SNP markers of either decreased (rs572527200) or increased (rs1401366377) risk of RA, as shown in **Supplementary Table 1**.

Likewise, we one-by-one updated the calculation results for those 13 human genes (*ACKR1*, *APOA1*, *DHFR*, *F3*, *F7*, *HBB*, *HBD*, *IL1B*, *INS*, *MBL2*, *NOS2*, *SOD1*, and *TPI1*) that we have already examined earlier and described in detail in our previous article on RA (Chadaeva I. V. et al., 2019); the updated results on these genes are found in **Supplementary Table 1**. Finally, after the publication of our previous article on RA (Chadaeva I. V. et al., 2019), within PubMed (Lu, 2011), we found clinically proven SNP disease markers located within 70-bp proximal promoters of the 11 human genes (*ADH7*, *CETP*, *COMT*, *ESR2*, *FGFR2*, *HSD17B1*, *HTR2C*, *MLH1*, *PDYN*, *RET*, and *TGFBR2*), as described below for the first time.

Human gene *PDYN* (i.e., prodynorphin as a basic building block of endorphins inhibiting pain and causing euphoria as hormones/neurotransmitters of joy) has a known SNP marker (rs886056538) of spinocerebellar ataxia when PDYN is in excess (**Supplementary Table 1**) according to the ClinVar database. Our keyword search in PubMed pointed to a rat model of human diseases, where the pain sensitivity threshold in RA positively correlates with this hormone's abundance (Zheng et al., 2014). With this in mind, we predicted that rs886056538 is a candidate SNP marker of relieved RA as presented in **Supplementary Table 1**. As readers can see in this table, in the vicinity of rs886056538, we selected nine SNPs (e.g., rs1195765727), each of which can reduce the PDYN level and thus may be a candidate SNP marker of relieved RA too (**Supplementary Table 1**).

The human *COMT* gene encoding catechol-O-methyltransferase bears two SNP markers, rs370819229 and rs777650793, implicated by the ClinVar database in dilated cardiomyopathy and other cardiovascular diseases, respectively, because of COMT downregulation and upregulation calculated in our study (**Supplementary Table 1**). Within PubMed, we found a short clinical communication (Finan and Zautra, 2013)

TABLE 2 | Candidate SNP markers of rheumatoid arthritis (RA) that are located near TBP-sites of human gene promoters as predicted in this work, in comparison with genome-wide patterns.

Data: GRCh38 (Zerbino et al., 2015) dbSNP build 151 (Day, 2010)				Result	Neutral drift (Haldane, 1957; Kimura, 1968; Kasowski et al., 2010)			H_0: ↑↓-equivalence		
No.	**SNPs**	N_G	N_S	N_R	$N_>$	$N_<$	$p(H_0: N_> < N_<)$	$N_↑$	$N_↓$	$p(H_0: N_↑ \equiv N_↓)$
1	Whole-genome norm for SNPs of TBP-sites (The 1000 Genomes Project Consortium (GPC) et al., 2012)	10^4	10^5	10^3	200	800	>0.99	–	–	–
2	Clinical SNP markers of diseases at TBP-sites (Ponomarenko et al., 2015)	33	203	51	14	37	>0.99	–	–	–
3	Candidate SNP markers of RA near clinical SNP markers of diseases at TBP-sites (Chadaeva I. V. et al., 2019)	18	225	42	12	30	>0.99	32	10	$<10^{-4}$
4	Candidate SNP markers of RA near clinical SNP markers of diseases at TBP-sites (this work)	25	603	154	84	70	>0.15	96	58	$<10^{-2}$
5	Candidate SNP markers of RA near TBP-sites of the genes most often associated with RA (this work)	10	466	69	46	23	$<10^{-2}$	42	27	<0.05
6	Candidate SNP markers of RA near TBP-sites of immunostimulatory genes (this work)	8	479	114	71	43	<0.01	71	43	$<10^{-2}$
7	Candidate SNP markers of RA near TBP-sites of immunosuppressive genes (this work)	25	928	179	104	75	<0.025	70	109	$<10^{-2}$
8	**Total**	**68**	**1896**	**516**	–	–	–	–	–	–

Rheumatoid arthritis (RA): contribute to (↑) and relieve (↓). N_G and N_S: total numbers of the human genes and of their SNPs meeting the criteria of this study. N_R: the total number of the candidate SNP markers increasing ($N_>$) or decreasing ($N_<$) the affinity of TATA-binding protein (TBP) for these promoters studied in this work. $N_↑$ and $N_↓$: total numbers of the candidate SNP markers that can contribute to or relieve RA, respectively. $P(H_0)$: the estimate of probability for the acceptance of this H_0 hypothesis, according to a binomial distribution. TBP-site: TBP-binding site.

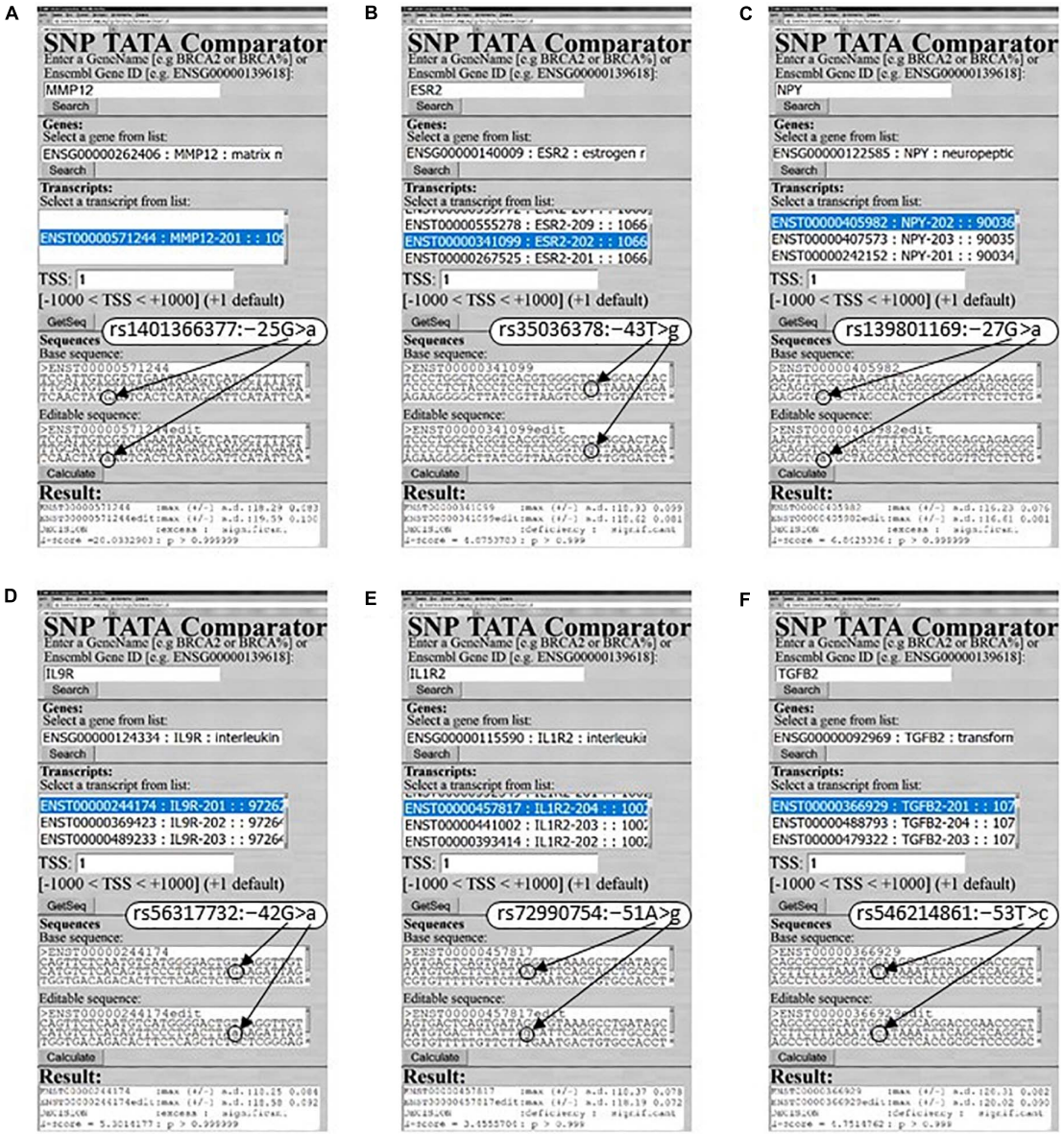


FIGURE 3 | Examples of our predictions for human RA-related genes in this work. **(A)** *MMP12*: rs1401366377; **(B)** *ESR2*: rs35036378; **(C)** *NPY*: rs139801169; **(D)** *IL9R*: rs56317732; **(E)** *IL1R2*: rs72990754; **(F)** *TGFB2*: rs546214861.

on a negative correlation between chronic pain sensitivity in RA and COMT. In accordance with these clinical observations, we proposed two known SNP markers of cardiovascular pathologies (rs370819229 and rs777650793) as candidate SNP markers of worsened and alleviated RA, respectively, as shown in **Supplementary Table 1**. In total, nearby we predicted two more candidate SNP markers (rs901020754 and rs779542396) worsening RA and 13 more candidate SNP markers relieving RA (e.g., rs748298389), all of which are given in **Supplementary Table 1**.

Human gene *RET* codes for the Ret proto-oncogene, whereas database ClinVar shows two SNPs, rs10900297 and rs10900296, located within its 70-bp proximal promoter, because they occur in patients with either pheochromocytoma or renal dysplasia (**Supplementary Table 1**). According to our predictions (**Supplementary Table 1**), these SNPs can surprisingly cause over- and under-expression of this gene, respectively, as confirmed by two clinical studies (Bridgewater et al., 2008; Sarin et al., 2014) retrieved by our keyword search in PubMed. Moreover, in this way, we found a clinical case report (Townsend et al., 1994) where both pheochromocytoma and RA were mutually complicating diagnoses. For this reason, we suggest two candidate SNP markers (rs10900297 and rs10900296) of complications in RA diagnosis and the same for four other SNPs, which can either significantly reduce the RET level (i.e., rs551321384 and rs1191017949) or significantly elevate it

(i.e., rs1237152255 and rs1372293149), as readers can see in **Supplementary Table 1**.

Human gene *MLH1* encodes DNA mismatch repair protein MLH1; in this gene's promoter, ClinVar lists two SNP markers (rs63750527 and rs756099600) of colon cancer that elevate the MLH1 level, thus preventing cancer cell apoptosis during cancer chemotherapy and an immune response (**Supplementary Table 1**). We found a retrospective association (Jeong et al., 2017) (comorbidity) of colon cancer and RA. Accordingly, we predicted two candidate SNP markers (rs63750527 and rs756099600) of high risk of comorbidities in RA (**Supplementary Table 1**). Additionally, near rs63750527 and rs756099600, we identified three more SNPs also increasing both the MLH1 amount and colon cancer risk (e.g., rs752622244) and hence elevating the risk of an RA comorbidity (**Supplementary Table 1**). In the table, the reader can see three other SNPs that are located within the *MLH1* promoter (e.g., rs864622145) and reduce an expression of this gene, thereby possibly accelerating the progression of RA to cancer because of reduced DNA repair (Kullmann et al., 2000).

Human gene *ADH7* encoding alcohol dehydrogenase 7 carries a biomedically identified SNP marker (rs17537595) of esophageal cancer; the minor allele of this SNP can reduce *ADH7* expression (Abbas et al., 2006) in agreement with our prediction presented in **Supplementary Table 1**. Our keyword search in PubMed revealed that both a deficit and an excess of this enzyme occur quite often in digestive-tract cancers (Jelski et al., 2008) that are comorbid with RA (Hemminki et al., 2012). That is why, for increased risk of comorbidities of RA, we predicted candidate SNP marker rs17537595 as well as three more markers nearby (rs372329931, rs755152695, and rs1238877951) having the same effect on this gene's expression (**Supplementary Table 1**).

The promoter of human gene *HSD17B1* contains SNP rs201739205, which decreases the amount of hydroxysteroid (17-β) dehydrogenase 1 encoded by this gene (**Supplementary Table 1**), as revealed in patients with hereditary breast cancer (Peltoketo et al., 1994), which is discordant to RA (Chen et al., 2019) according to the results of our keyword search in database PubMed. Therefore, we predicted a candidate SNP marker (rs201739205) of low risk of RA (**Supplementary Table 1**). Additionally, around it, we chose two more SNPs (rs748743528 and rs779674159) that decrease the HSD17B1 level and therefore are candidate SNP markers of an RA risk reduction too (**Supplementary Table 1**). Finally, within the promoter under study, we predicted three SNPs able to cause HSD17B1 overexpression (e.g., rs1332869256), each of which can also cause breast cancer (He et al., 2016) according to our PubMed search and therefore may be a candidate SNP marker of decreased risk of RA (**Supplementary Table 1**).

The human *ESR2* gene coding for estrogen receptor 2 (β) carries a clinically proven SNP marker (rs35036378) of an ESR2-deficient primary pT1 tumor of the mammary gland (Philips et al., 2012), for which we also correctly predicted ESR2 underexpression, as shown in both **Figure 3B** and **Supplementary Table 1**. Our PubMed keyword search yielded a mouse model of human diseases with ESR2-dependent suppression of inflammation (Armstrong et al., 2013). These data allowed us to predict a candidate SNP marker (rs35036378) of reduced suppression of inflammation in RA (**Supplementary Table 1**). Nearby, we identified the only SNP (rs766797386) that can decrease the ESR2 level and consequently is a candidate SNP marker of reduced suppression of inflammation in RA (**Supplementary Table 1**).

The human *FGFR2* gene (fibroblast growth factor receptor 2) is annotated in the ClinVar database showing two biomedical SNP markers in its promoter: rs886046768 for craniosynostosis and rs387906677 for bent bone dysplasia, which correspond, respectively, to an excess and deficit of this receptor according to our predictions (**Supplementary Table 1**). As for RA, using our keyword search in PubMed, we found a physiological *in vitro* model of human angiogenesis, where neovascularization increases with an FGFR2 concentration increase and vice versa (Brown et al., 1996). This finding permits us to predict that all SNPs in this gene's promoter, which can significantly elevate (e.g., rs886046768) and reduce (e.g., rs387906677) this receptor's level, are candidate SNP markers of elevated and reduced neovascularization in RA, respectively (**Supplementary Table 1**).

The promoter of human gene *TGFBR2* has a known SNP marker (rs138010137) of aortic thoracic aneurysm according to database ClinVar. Our prediction for this SNP is underexpression of transforming growth factor beta receptor 2 encoded by this gene (**Supplementary Table 1**). We learned that a TGFBR2 deficit disrupts regulatory T-cell homeostasis in RA (Wang et al., 2018). This allows us to suggest a candidate SNP marker (rs138010137) of higher risk of RA as well as nearby, one more candidate SNP marker (rs1300366819) of the same RA parameter because of a decrease in the TGFBR2 level, according to our prediction (**Supplementary Table 1**). Finally, in the same way, we found out that synovial-fibroblast proliferation increases with an increase in the TGFBR2 level (Bira et al., 2005). Consequently, yet another SNP (rs1310294304) located within this promoter can be a candidate SNP marker of worsened RA because of TGFBR2 upregulation (**Supplementary Table 1**).

The human *CETP* gene codes for plasma lipid transfer protein and has a widely used SNP marker (rs1427119663) of hyperalphalipoproteinemia, which is an 18-bp deletion including the TBP-site in the promoter region of this gene; this mutation causes CETP underexpression (Plengpanich et al., 2011), as described in **Supplementary Table 1**. Because there is a known phenomenon of CETP-deficient mortality in RA (Ferraz-Amaro et al., 2013), in line with our keyword search in PubMed, we predicted rs1427119663 as a candidate SNP marker of complicated RA (**Supplementary Table 1**). In its vicinity, we identified five SNPs (e.g., rs17231520), all of which can increase this gene's expression as a known risk factor of RA (Kim et al., 2016), hence our suggestion to regard them as candidate SNP markers of higher risk of RA (**Supplementary Table 1**).

Human gene *HTR2C* coding for serotonin receptor 2C is mentioned in ClinVar owing to its clinically documented SNP marker (rs3813929) of obesity as a complication of olanzapine-based antipsychotic treatment; this marker is associated with an HTR2C excess according to our calculations (**Supplementary Table 1**). As for RA, using our keyword search *via* PubMed, we revealed that HTR2C overexpression elevates the risk of RA with obstructive

sleep apnea (Jagannathan et al., 2017) in contrast to HTR2C underexpression, which reduces adipogenesis and thus the obesity-related risk of RA (Priyadarshini et al., 2018). Keeping this in mind, we predicted that rs3813929 and five more SNPs nearby (e.g., rs886838672) are candidate SNP markers of the obesity-related complications in RA, as readers can see in **Supplementary Table 1**.

Predictions Based Upon dbSNP: Build No. 151 Versus Build No. 147

Row 4 of **Table 2** sums up the aforementioned predictions on the basis of dbSNP, build No. 151, about the 603 SNPs within 25 genes' promoters containing biomedical SNP markers, namely, a total number of candidate SNP markers is 154, that is, ∼fourfold higher than 42 for build No. 147 (row 3). In the three rightmost columns, readers can see the statistically significant predominance (a greater number) of candidate SNP markers contributing to RA (i.e., 32 in row 3 and 96 in row 4) over candidate SNP markers preventing RA (i.e., 10 in row 3 and 58 in row 4). This result supports the robustness of our predictions in both compared cases. This predominance of predisposition over resistance to RA fits the notion that diversity of adaptive-immunity memory cells in an individual grows with an increase in the number of diseases patient survived, thus elevating the disease resistance in the future (Sarkander et al., 2016) and the risk of false activation of these cells according to Ashby's law (Schuldberg, 2015).

Besides, the three central columns illustrate how we verified the robustness of our predictions with the annual growth of dbSNP from the standpoint of both Kimura's theory (1968) and Haldane's dilemma (1957), which highlight neutral drift of most SNPs as the cornerstone of the human genome as a whole. As a bioinformatics criterion of neutral drift within regulatory regions of the genome, some authors (Kasowski et al., 2010) first noted empirically and next proposed heuristically to estimate an excess in the number of SNPs damaging protein-binding sites over those improving them. In row 1, readers can see that the genome-wide pattern of SNPs in humans that was identified using ChIP-seq data (The 1000 Genomes Project Consortium (GPC) et al., 2012) fits this neutral drift criterion at a probability rate (P) of >0.99 and actually follows a binomial distribution, indeed. Row 2 presents the genome-wide pattern of the clinically proven biomedical SNP markers near the TBP-sites within human gene promoters according to our predictions made by means of SNP_TATA_Comparator, and this pattern also matches the same criterion of neutral drift. This finding reflects the essence of Haldane's dilemma (1957), who hypothesized an under-threshold level of the human genetic load. In row 3, there are relevant results of our previous research (Chadaeva I. V. et al., 2019) on the candidate SNP markers of RA within a previous build of dbSNP No. 147; again, they seem to be in agreement with the neutral drift criterion used. As for the candidate SNP markers of RA predicted using build No. 151 of this database (row 4), contrary to all of the above, there is a lower, not higher, number of SNPs damaging TBP-sites than those improving them; this result is insignificant ($P > 0.15$) and does not allow to rule out their neutral drift completely. This indicates that almost a fourfold greater number of SNPs within the current build of dbSNP relative to the previous one may allow us to detect other genome-wide patterns of candidate SNP markers of RA besides their neutral drift, which is unquestionable (Haldane, 1957; Kimura, 1968). Therefore, we next examined the human genes that are most often associated with RA according to independent sources, as presented in **Supplementary Table 2** and described below.

Candidate SNP Markers of RA Near TBP-Sites in Promoters of the Human Protein-Coding Genes That Are Most Often Independently Associated With RA

The human *NPY* gene (neuropeptide Y) has four SNPs causing an excess of this protein (e.g., **Figure 3C**: rs139801169) and four SNPs diminishing its level (e.g., rs1223788416), which correspond to either higher or lower risk of obesity that is comorbid with RA (Stofkova et al., 2009), as found in PubMed. Thus, we predicted eight RA-related candidate SNP markers (**Supplementary Table 2**).

The human *CCR6* promoter contains two SNPs rs1433814180 and rs1047738754 that both can decrease the amount of C–C motif chemokine receptor 6 encoded by this gene, as shown in **Supplementary Table 2** presenting our predictions. Our keyword search within the PubMed database revealed a CCR6-deficient mouse model of human RA with reduced autoimmunity (Bonelli et al., 2018). This result allows us to propose two candidate SNP markers of decreased risk of RA (**Supplementary Table 2**).

Human gene *CTLA4* (cytotoxic T-lymphocyte associated protein 4) carries four SNPs causing an overabundance this protein (e.g., rs127192924), which is a molecular biomarker of the overlapping autoimmunity proven by a case-control study (AlFadhli, 2013), as indicated in **Supplementary Table 2**. In addition, within its proximal promoters, there are four other SNPs whose minor alleles reduce the CTLA4 level according to our predictions (**Supplementary Table 2**: e.g., rs561368432), while there are CTLA4-inducible knockout mice as a laboratory animal model of human RA with an elevated autoimmune response (Alissafi et al., 2017). Considering all of the above, we predicted eight candidate SNP markers of worsened RA.

In total, within 10 human genes *CCR6*, *CTLA4*, *HLA-A*, *IL23R*, *IRF5*, *NPY*, *PADI4*, *PTPN22*, *STAT4*, and *TRAF1*, we predicted 42 and 27 candidate SNP markers, respectively, corresponding to a high and low risk of RA (**Supplementary Table 2**). On the other hand, 46 and 23 of all of them can, respectively, improve and damage TBP-sites within the promoters of the RA-related human genes examined, as shown in row 5 of **Table 2**. Thus, thanks to a fourfold increased number of SNPs within the current build No. 151 of dbSNP, besides the neutral drift toward predisposition to RA in humans (**Table 2**: rows 3 and 4), we first detected the natural selection against underexpression of the human genes often associated with RA; this selection could indeed elevate the risk of RA, as explained above. In hopes of further detailing this trend of natural selection, we next investigated human immunostimulatory and immunosuppressive genes independently from one another,

TABLE 3 | Comparing the effects of the orthologous-gene expression changes on rheumatoid arthritis (RA) in humans and during the divergence of domestic and wild animals from their nearest common ancestor.

Humans					Animals				
Gene	**Effect of gene expression change (Δ) on RA, i.e., either contributes to (↑) or relieves (↓)**				***DEG***	***RNA-Seq***		**Δ During divergence from the nearest common ancestor**	
	Deficit	**RA**	**Excess**	**RA**		**log_2**	**P_{ADJ}**	**Deficit**	**Excess**
					Tame vs. aggressive foxes (Hekman et al., 2018)				
ESR2	Less ESR2-dependent suppression (Armstrong et al., 2013)	↑	More ESR2-dependent suppression (Armstrong et al., 2013)	↓	*Esr2*	−0.3	0.05	Domestic	Wild
IL1R2	More inflammation (Ocsko et al., 2018)	↑	Less inflammation (Ocsko et al., 2018)	↓	*Il1r2*	−0.4	0.05	Domestic	Wild
IL9R	Less inflammation judging by fibroblast-like synoviocytes (Raychaudhuri et al., 2018)	↓	More inflammation judging by fibroblast-like synoviocytes (Raychaudhuri et al., 2018)	↑	*Il9r*	0.4	0.05	Wild	Domestic
NPY	Lower risk of obesity-caused RA (Stofkova et al., 2009)	↓	Higher risk of obesity-caused RA (Stofkova et al., 2009)	↑	*Npy*	0.4	10^{-2}	Wild	Domestic
TGFB2	Better recovery from RA (Um et al., 2018)	↓	Inhibited bone repair in RA (Um et al., 2018)	↑	*Tgfb2*	0.5	10^{-2}	Wild	Domestic
					Domesticated vs. wild rabbits (Albert et al., 2012)				
F7	Higher risk of hemorrhagic forms of RA (Thornorsteinsson et al., 2004)	↑	Recombinant F7 is a drug in hemophilia comorbid with RA (Drobiecki et al., 2013)	↓	*F7*	−2.7	0.05	Domestic	Wild
					Guinea pigs vs. cavies (Albert et al., 2012)				
CCR6	Relieved RA (Bonelli et al., 2018)	↓	Higher risk of RA (Jatczak-Pawlik et al., 2020)	↑	*Ccr6*	2.2	0.05	Wild	Domestic
CETP	CETP deficiency promotes mortality in RA (Ferraz-Amaro et al., 2013)	↑	Higher risk of RA (Kim et al., 2016)	↑	*Cetp*	2.1	10^{-3}	Wild	Domestic
IL1B	Relieved RA (Rzepecka et al., 2015)	↓	Circadian pain in RA (Olkkonen et al., 2015)	↑	*Il1b*	2.3	10^{-2}	Wild	Domestic
PDYN	Pain hypersensitivity (Zheng et al., 2014)	↑	Pain resistance (Zheng et al., 2014)	↓	*Pdyn*	0.9	10^{-2}	Wild	Domestic
					Dog vs. wolf (blood) (Yang et al., 2018)				
HBB	Thalassemia-related osteoporosis worsens RA (Giakoumi et al., 2005)	↑	Hemolytic-origin extracellular HBB releases thrombogenic heme that adds to RA-related thrombogenesis (Bisoendial et al., 2010; Gall et al., 2018)	↑	*Hbbl*	−5.9	10^{-8}	Domestic	Wild

log_2: differential expression of a gene of pets normalized to that in wild animals (log_2 units).

as presented in **Supplementary Tables 3**, **4**, respectively, and discussed below.

Candidate SNP Markers of RA Near TBP-Sites in Promoters of Human Protein-Coding Immunostimulatory Genes

Human gene *IL9R* codes for interleukin 9 receptor and includes two SNPs rs56317732 and rs945044791 corresponding to its overexpression and underexpression, respectively, according to our predictions exemplified by **Figure 3D** as well as to high and low both inflammation and fibroblast-like synoviocyte proliferation in RA (Raychaudhuri et al., 2018). Therefore, we proposed them as candidate SNP markers for aggravated and alleviated RA (**Supplementary Table 3**).

Looking through **Supplementary Table 3**, readers can see a significant number of candidate SNP markers of improved TBP-sites than candidate SNP markers of damaged TBP-sites (71 vs. 43) within promoters of eight human immunostimulatory genes *ATF3*, *CCR7*, *IL3RA*, *IL9R*, *IL25*, *LCK*, *NFKB1*, and *ZBTB38* ($p < 0.01$, binomial distribution), as summarized in row 6 of **Table 2**. This corresponds to a significant predominance of 71 candidate SNP markers increasing RA-related risks over 43 such markers reducing these risks ($p < 0.01$), as presented *ibid*. Therefore, we can conclude that there is the pressure of natural selection on human immunostimulatory genes, and its direction both prevents their underexpression and elevates the human predisposition to RA, thus fitting the trends in the abovementioned rows 3, 4, and 5.

Candidate SNP Markers of RA Near TBP-Sites in the Promoters of Human Protein-Coding Immunosuppressive Genes

Human *IL1R2* gene promoters contain two SNPs rs960068265 and rs946299576 able to increase the level of interleukin 2 receptor subunit β encoded by this gene as well as two other SNPs rs72990754 and rs960678696 capable of decreasing this level, as exemplified in **Figure 3E**. Using a keyword search in the PubMed database, we found a murine laboratory model of human diseases (Ocsko et al., 2018) where epigenetic silencing of this anti-inflammatory gene supported RA. Accordingly, **Supplementary Table 4** presents four RA-related candidate SNP markers within this gene's promoters, as predicted here.

The human gene *TGFB2* promoter has only one SNP, rs546214861, that can lower the production of transforming growth factor beta 2 (synonym: glioblastoma-derived T-cell suppressor factor) encoded by this gene (**Figure 3F**), thereby improving the healing of bone-related tissues in inflammatory RA (Um et al., 2018) as a candidate SNP marker of RA alleviation (**Supplementary Table 4**).

If we look through **Supplementary Table 4**, there are more (104 vs. 75) candidate SNP markers corresponding to enhancement of (than damage to) TBP-sites of 25 human immunosuppressive genes *BCL6*, *CD4*, *CNMD*, *EBI3*, *FGF21*, *GAS6*, *GDF5*, *DUSP1*, *FGF22*, *FOXP3*, *IL1R2*, *IL2RA*, *IL2RB*, *IL4*, *IL10*, *IL10RA*, *IL10RB*, *IRF2*, *IRF4*, *IRF8*, *PDCD1*, *PIAS1*, *SOCS3*, *TGFB2*, and *TNFRSF8* (**Table 2**, row 7: $p < 0.025$, binomial distribution). This means natural-selection pressure directed against underexpression of the human immunosuppressive genes, consistently with the abovementioned rows 6 and 7 of **Table 2**; these rows correspond to genes often associated with RA and immunostimulatory genes, respectively. As for predisposition to RA, contrary to all the abovementioned rows 3, 4, 5, and 6 of **Table 2**, when summarizing **Supplementary Table 4**, we were surprised by the significant predominance (109 over 70) of candidate SNP markers corresponding to alleviation over aggravation of RA ($p < 0.01$) binomial distribution rather than vice versa, as presented in row 7 of **Table 2**. This result indicates that natural-selection pressure on the human immunosuppressive genes is directed toward resistance to RA rather than predisposition to RA presented in rows 3, 4, 5, and 6 of this table. First, within row 7 compared with rows 3 and 4 of **Table 2**, readers can see two statistically significant genome-wide patterns opposite to each other, namely: (1) neutral drift increasing RA-related risks and (2) natural selection decreasing them; superposition of these patterns can stabilize them and thus establish the normal level of RA-related risks as a human trait. Finally, rows 5, 6, and 7 indicate the only consistently bidirectional natural selection at the whole-genome scale [either (1) for RA resistance in case of immunosuppressive genes or (2) for RA predisposition in case of immunostimulatory and all remaining genes] that can disrupt the RA norm in humans as if self-domestication (Theofanopoulou et al., 2017) has occurred with its disruptive natural selection (Belyaev, 1979). According to the mainstream point of view on Human Origins,

TABLE 4 | Correlations between the effects of codirected changes in the expression of orthologous genes on rheumatoid arthritis (RA) in humans and during the divergence of domestic and wild animals from their nearest common ancestor.

Animals	Humans	Effect of gene expression changes on RA		Binomial distribution	χ^2-test		Fisher's exact test, *p*
		Contribute to (↑)	Relieve (↓)		χ^2	*p*	
Effect of gene expression changes during divergence from the nearest common ancestor	Domestic	10	1	$<10^{-2}$	9	10^{-2}	0.05
	Wild	3	8	>0.1			

there is not enough scientific evidence that this could actually happen. In this work, we verified the bidirectional natural selection on the human genome-wide scale simultaneously for resistance and predisposition to RA during their comparison with publicly available data on DEGs in pets vs. wild animals, as presented below.

In vivo Validation of Our Predictions Using DEGs Within Pets Versus Wild Animals

Here, we compared 68 human genes within whose promoters there are RA-related candidate SNP markers predicted in this work (**Supplementary Tables 1–4**, the essence of which is **Supplementary Table 5**) with 1740 DEGs of pets vs. wild animals (Albert et al., 2012; Hekman et al., 2018; Yang et al., 2018). Their description is given in **Table 1**, as depicted in **Figure 2**. The obtained results are presented in **Table 3**.

First of all, as readers can see in **Table 3**, underexpression of three human genes *ESR2*, *IL1R2*, and *F7* corresponds to attenuated ESR2-dependent suppression (Armstrong et al., 2013), increased inflammation (Ocsko et al., 2018), and hemorrhagic forms of RA (Thornorsteinsson et al., 2004); this pattern is consistent with underexpression of their orthologous animal genes during domestication and *vice versa*. Besides, overexpression of five human genes *CCR6*, *IL9R*, *NPY*, *IL1B*, and *TGFB2* elevates the risk of RA (Jatczak-Pawlik et al., 2020), inflammation (Raychaudhuri et al., 2018), obesity (Stofkova et al., 2009), and circadian pain (Olkkonen et al., 2015) and inhibits bone repair (Um et al., 2018), respectively; these data match the pattern of overexpression of their orthologous genes during animal domestication, and *vice versa*. Additionally, both underexpression and overexpression of two human genes *CETP* and *HBB* contribute to RA, as shown in **Table 3**. Finally, overexpression of only human gene *PDYN* in our gene set reduces pain sensitivity (Zheng et al., 2014), thus relieving RA, and in only this case in our study corresponds to the orthologous gene's overexpression during the guinea pig domestication and *vice versa* (**Table 3**).

Table 4 sums up the findings of the comparative analysis of the above orthologous genes from humans and animals, namely, 10 and 1 of these domestic-animal DEGs were found to correspond to human gene-markers of aggravated and relieved RA, and the same is true for three and eight DEGs in the wild animals. Therefore, the DEGs in domestic animals are significantly consistent with their human orthologous genes that contribute to RA, according to Pearson's χ^2 test ($p < 0.01$), Fisher's exact test ($p < 0.05$), and binomial distribution ($p < 0.01$). Finally, **Table 4** indicates that the DEGs of wild animals correspond equally to human orthologous genes that aggravate and relieve RA ($p > 0.1$, binomial distribution) in agreement with the conventional wild-type norm.

These findings mean that during domestication, the anthropogenic environment, in contrast to a natural environment, may alter gene expression in the animals on the genome-wide scale (e.g., immunostimulatory and immunosuppressive genes) in a manner that more often contributes to RA than not.

CONCLUSION

Because it is best to study TBP-sites genome-wide, we created SNP_TATA_Comparator and applied it to build No. 147 dbSNP of 2016 to predict candidate SNP markers of RA (Chadaeva I. V. et al., 2019) as a cause of disability (Koller and Nobauer-Huhmann, 2009). The robustness of this tool was verified here with the growth of the SNP number, as seen in dbSNP build No. 151 of 2017. Here we analyzed a fourfold higher SNP number allowing us to detect previously unknown whole-genome SNP patterns besides neutral drift (Haldane, 1957; Kimura, 1968) in relation to RA predisposition because diversity of adaptive-immunity memory cells grows with an increase in the number of diseases experienced, thereby enhancing both disease resistance (Sarkander et al., 2016) and the risk of false activation of these cells according to Ashby's law of requisite variety (Schuldberg, 2015). That is why we additionally investigated both immunostimulatory genes and genes quite often linked to RA, and we noted natural selection against underexpression of these genes in the same direction, toward predisposition to RA. Likewise, we analyzed immunosuppressive genes and surprisingly observed the same natural selection against underexpression of these genes, which nevertheless acts in the opposite direction, toward resistance to RA. Overall, the natural-selection pressure seems bidirectional, e.g.: (1) on immunosuppressive genes toward RA alleviation and (2) on immunostimulatory gene toward RA aggravation, suggesting that self-domestication events (Theofanopoulou et al., 2017) have happened in humans because of the disruptive natural selection (Belyaev, 1979) found in our study. It is common knowledge that the use of gene promoters alone is not enough for the analysis of modes of evolution as a whole. Belyaev (1979) defined the domestication-related disruptive selection as "... what may be selected for are changes in the regulation of genes — that is, in the timing and the amount of gene expression rather than changes in individual structural genes" (Belyaev, 1979). This is a microevolution event when domestic and wild populations of the same species diverge from their nearest common ancestor, to which it is appropriate to compare our human-promoter-based predictions *in silico* (Ponomarenko et al., 2015) and the DEGs of domestic vs. wild animals *in vivo*, as demonstrated recently (Vasiliev et al., 2021).

In accordance with the mainstream opinion that scientific evidence for the human self-domestication is insufficient, we tested our predictions of relief and aggravation in RA using public data on 1740 DEGs in pets vs. wild animals (Albert et al., 2012; Hekman et al., 2018; Yang et al., 2018) as a bioinformatic animal model of human diseases. Among the DEGs examined, this approach yielded 10 and 1 DEGs that correspond to alleviation and aggravation of RA in pets, in contrast to three and 8 DEGs in the case of wild animals. Consequently, during domestication, the anthropogenic habitat

conditions in comparison with a natural environment may change gene expression in the animals on the whole-genome scale (e.g., immunostimulatory and immunosuppressive genes) and thus contribute more often to RA according to three independent statistical criteria, such as Pearson's χ^2 test ($p < 0.01$), Fisher's exact test ($p < 0.05$), and the binomial distribution test ($p < 0.01$). This finding allows us to propose RA as a candidate symptom within a self-domestication syndrome (Theofanopoulou et al., 2017). Such syndrome might be considered as a human's payment with health for the benefits received during evolution.

Besides, the RA-related candidate SNP markers predicted here, which are expected to survive obligatory clinical "case-control" studies in the future, may become useful for physicians for optimizing the treatment of patients according to their individual sequenced genome reducing RA-related risks.

The presented verification of SNP_TATA_Comparator's predictions vis-à-vis the semiquantitative RNA-Seq data is statistically significant. Therefore, the next step of further comprehensive experimental verification of SNP_TATA_Comparator's biomedical predictions by means of genome-wide data on QTLs [e.g., in human cardiopathology (Koopmann et al., 2014)] seems to be justified and timely.

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

NAK and VK contributed to study conception. MP wrote the manuscript. EO, IC, PP, DO, and DR contributed to the development and optimization of the software. NVK, ES, ID, and LS contributed to the data analysis. All authors contributed to the article and approved the submitted version.

FUNDING

The concept was supported by the Russian Federal Science & Technology Program for the Development of Genetic Technologies (for NAK and VK). The software development, manuscript writing, and data analysis were supported by Project No. 0259-2021-0009 from the Russian Government Budget (for EO, IC, PP, DO, DR, NVK, ES, ID, and LS). All authors declare that these funding bodies did not play roles in the design of the study; in the collection, analysis, and interpretation of the data; and in the writing of the manuscript.

ACKNOWLEDGMENTS

We are thankful to Shevchuk Editing (http://www.shevchuk-editing.com) (Brooklyn, NY, United States) for English editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.610774/full#supplementary-material

Supplementary File 1 : Supplementary Methods | An estimate of the affinity of TATA-binding protein (TBP) for a 70 bp proximal promoter in human genes.

Supplementary File 2 : Supplementary Results | Supplementary Tables 1–5. Candidate SNP markers of RA predicted in this work near TBP-binding sites within a promoter of human protein-coding genes.

Supplementary File 3 : Supplementary Keyword Search | Supplementary Figure 1. A decision logic of the keyword search for RA-related studies in the PubMed database (Lu, 2011).

REFERENCES

Abbas, A., Lechevrel, M., and Sichel, F. (2006). Identification of new single nucleotid polymorphisms (SNP) in alcohol dehydrogenase class IV ADH7 gene within a French population. *Arch. Toxicol.* 80, 201–205. doi: 10.1007/s00204-005-0031-7

Albert, F. W., Somel, M., Carneiro, M., Aximu-Petri, A., Halbwax, M., Thalmann, O., et al. (2012). A comparison of brain gene expression levels in domesticated and wild animals. *PLoS Genet.* 8:e1002962. doi: 10.1371/journal.pgen.1002962

AlFadhli, S. (2013). Overexpression and secretion of the soluble CTLA-4 splice variant in various autoimmune diseases and in cases with overlapping autoimmunity. *Genet. Test. Mol. Biomarkers* 17, 336–341. doi: 10.1089/gtmb.2012.0391

Alissafi, T., Banos, A., Boon, L., Sparwasser, T., Ghigo, A., Wing, K., et al. (2017). Tregs restrain dendritic cell autophagy to ameliorate autoimmunity. *J. Clin. Invest.* 127, 2789–2804. doi: 10.1172/JCI92079

Alsaed, O., Hadwan, N., Khanjar, I., and Al-Allaf, A.-W. (2018). Seronegative bilateral symmetrical inflammatory polyarthritis: think twice before starting immunosuppression. *Eur. J. Case Rep. Intern. Med.* 5:000895. doi: 10.12890/2018_000895

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: online Mendelian Inheritance in Man (OMIM§), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Armstrong, C. M., Billimek, A. R., Allred, K. F., Sturino, J. M., Weeks, B. R., and Allred, C. D. (2013). A novel shift in estrogen receptor expression occurs as estradiol suppresses inflammation-associated colon tumor formation. *Endocr. Relat. Cancer* 20, 515–525. doi: 10.1530/erc-12-0308

Belyaev, D. K. (1979). The Wilhelmine E. Key 1978 invitational lecture. Destabilizing selection as a factor in domestication. *J. Hered.* 70, 301–308. doi: 10.1093/oxfordjournals.jhered.a109263

Bhuiyan, T., and Timmers, H. T. M. (2019). Promoter recognition: putting TFIID on the spot. *Trends Cell Biol.* 29, 752–763. doi: 10.1016/j.tcb.2019.06.004

Bira, Y., Tani, K., Nishioka, Y., Miyata, J., Sato, K., Hayashi, A., et al. (2005). Transforming growth factor beta stimulates rheumatoid synovial fibroblasts via the type II receptor. *Mod. Rheumatol.* 15, 108–113. doi: 10.1007/s10165-004-0378-2

Bisoendial, R. J., Levi, M., Tak, P. P., and Stroes, E. S. (2010). The prothrombotic state in rheumatoid arthritis: an additive risk factor for adverse cardiovascular events. *Semin. Thromb. Hemost.* 36, 452–457. doi: 10.1055/s-0030-1254054

Bonelli, M., Puchner, A., Goschl, L., Hayer, S., Niederreiter, B., Steiner, G., et al. (2018). CCR6 controls autoimmune but not innate immunity-driven experimental arthritis. *J. Cell. Mol. Med.* 22, 5278–5285. doi: 10.1111/jcmm.13783

Bridgewater, D., Cox, B., Cain, J., Lau, A., Athaide, V., Gill, P. S., et al. (2008). Canonical WNT/beta-catenin signaling is required for ureteric branching. *Dev. Biol.* 317, 83–94. doi: 10.1016/j.ydbio.2008.02.010

Brown, K. J., Maynes, S. F., Bezos, A., Maguire, D. J., Ford, M. D., and Parish, C. R. (1996). A novel in vitro assay for human angiogenesis. *Lab. Invest.* 75, 539–555.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212, 563–578. doi: 10.1016/0022-2836(90)90223-9

Chadaeva, I., Ponomarenko, P., Rasskazov, D., Sharypova, E., Kashina, E., Kleshchev, M., et al. (2019). Natural selection equally supports the human tendencies in subordination and domination: a genome-wide study with in silico confirmation and in vivo validation in mice. *Front. Genet.* 10:73. doi: 10.3389/fgene.2019.00073

Chadaeva, I. V., Rasskazov, D. A., Sharypova, E. B., Drachkova, I. A., Oshchepkova, E. A., Savinkova, L. K., et al. (2019). Candidate SNP-markers of rheumatoid arthritis that can significantly alter the affinity of the TATA-binding protein for human gene promoters. *Vavilov. Zh. Genet. Selektsii.* 23, 1047–1058. doi: 10.18699/vj19.586

Chen, H. H., Lin, C. H., Chen, D. Y., Chao, W. C., Chen, Y. H., Hung, W. T., et al. (2019). Risk of major autoimmune diseases in female breast cancer patients: a nationwide, population-based cohort study. *PLoS One* 14:e0222860. doi: 10.1371/journal.pone.0222860

Day, I. N. (2010). dbSNP in the detail and copy number complexities. *Hum. Mutat.* 31, 2–4. doi: 10.1002/humu.21149

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. *Cell* 166, 538–554. doi: 10.1016/j.cell.2016.07.012

Dreos, R., Ambrosini, G., Groux, R., Perier, R. C., and Bucher, P. (2017). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res.* 45, D51–D55. doi: 10.1093/nar/gkw1069

Drobiecki, A., Pasiarski, M., Hus, I., Sokolowska, B., and Wątek, M. (2013). Acquired hemophilia in the patient suffering from rheumatoid arthritis: case report. *Blood Coagul. Fibrinolysis* 24, 874–880. doi: 10.1097/mbc.0b013e3283646635

Erlandsson, M. C., Doria Medina, R., Toyra Silfversward, S., and Bokarewa, M. I. (2016). Smoking functions as a negative regulator of IGF1 and impairs adipokine network in patients with rheumatoid arthritis. *Mediators Inflamm.* 2016:3082820. doi: 10.1155/2016/3082820

Ferraz-Amaro, I., Gonzalez-Gay, M. A., Garcia-Dopico, J. A., and Diaz-Gonzalez, F. (2013). Cholesteryl ester transfer protein in patients with rheumatoid arthritis. *J. Rheumatol.* 40, 1040–1047. doi: 10.3899/jrheum.121507

Finan, P. H., and Zautra, A. J. (2013). Rheumatoid arthritis: stress affects rheumatoid arthritis, but via what mechanisms? *Nat. Rev. Rheumatol.* 9, 569–570. doi: 10.1038/nrrheum.2013.139

Flatters, D., and Lavery, R. (1998). equence-dependent dynamics of TATA-Box binding sites. *Biophys. J.* 75, 372–381. doi: 10.1016/S0006-3495(98)77521-6

Frey, O., Meisel, J., Hutloff, A., Bonhagen, K., Bruns, L., Kroczek, R. A., et al. (2010). Inducible costimulator (ICOS) blockade inhibits accumulation of polyfunctional T helper 1/T helper 17 cells and mitigates autoimmune arthritis. *Ann. Rheum. Dis.* 69, 1495–1501. doi: 10.1136/ard.2009.119164

Gall, T., Petho, D., Nagy, A., Hendrik, Z., Mehes, G., Potor, L., et al. (2018). Heme induces endoplasmic reticulum stress (HIER stress) in human aortic smooth muscle cells. *Front. Physiol.* 9:1595. doi: 10.3389/fphys.2018.01595

Giakoumi, X., Tsironi, M., Floudas, C., Polymeropoylos, E., Papalambros, E., and Aessopos, A. (2005). Rheumatoid arthritis in thalassemia intermedia: coincidence or association? *Isr. Med. Assoc. J.* 7, 667–669.

Godde, J. S., Nakatani, Y., and Wolffe, A. P. (1995). The amino-terminal tails of the core histones and the translational position of the TATA box determine TBP/TFIIA association with nucleosomal DNA. *Nucleic Acids Res.* 23, 4557–4564. doi: 10.1093/nar/23.22.4557

Haeussler, M., Raney, B. J., Hinrichs, A. S., Clawson, H., Zweig, A. S., Karolchik, D., et al. (2015). Navigating protected genomics data with UCSC Genome Browser in a Box. *Bioinformatics* 31, 764–766. doi: 10.1093/bioinformatics/btu712

Hahn, S., Buratowski, S., Sharp, P., and Guarente, L. (1989). Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* 86, 5718–5722. doi: 10.1073/pnas.86.15.5718

Haldane, J. B. S. (1957). The cost of natural selection. *J. Genet.* 55, 511–524. doi: 10.1007/bf02984069

He, W., Gauri, M., Li, T., Wang, R., and Lin, S. X. (2016). Current knowledge of the multifunctional 17β-hydroxysteroid dehydrogenase type 1 (HSD17B1). *Gene* 588, 54–61. doi: 10.1016/j.gene.2016.04.031

Hekman, J. P., Johnson, J. L., Edwards, W., Vladimirova, A. V., Gulevich, R. G., Ford, A. L., et al. (2018). Anterior pituitary transcriptome suggests differences in ACTH release in tame and aggressive foxes. *G3 (Bethesda)* 8, 859–873. doi: 10.1534/g3.117.300508

Hemminki, K., Liu, X., Ji, J., Sundquist, J., and Sundquist, K. (2012). Effect of autoimmune diseases on mortality and survival in subsequent digestive tract cancers. *Ann. Oncol.* 23, 2179–2184. doi: 10.1093/annonc/mdr590

Ho, M. F., and Weinshilboum, R. M. (2017). Immune mediator pharmacogenomics: TCL1A SNPs and estrogen-dependent regulation of inflammation. *J. Nat. Sci.* 3:e416.

Hunninghake, G., Cho, M., Tesfaigzi, Y., Soto-Quiros, M., Avila, L., Lasky-Su, J., et al. (2009). MMP12, lung function, and COPD in high-risk populations. *N. Engl. J. Med.* 361, 2599–2608. doi: 10.1056/nejmoa0904006

Jagannathan, R., Seixas, A., St-Jules, D., Jagannathan, L., Rogers, A., Hu, L., et al. (2017). Systems biology genetic approach identifies serotonin pathway as a possible target for obstructive sleep apnea: results from a literature search review. *Sleep Disord.* 2017:6768323. doi: 10.1155/2017/6768323

Jatczak-Pawlik, I., Wolinski, P., KsiAZek-Winiarek, D., Pietruczuk, M., and Glabinski, A. (2020). CCR6 blockade on regulatory T cells ameliorates experimental model of multiple sclerosis. *Cent. Eur. J. Immunol.* 45, 256–266. doi: 10.5114/ceji.2020.101241

Jelski, W., Chrostek, L., Zalewski, B., and Szmitkowski, M. (2008). Alcohol dehydrogenase (ADH) isoenzymes and aldehyde dehydrogenase (ALDH) activity in the sera of patients with gastric cancer. *Dig. Dis. Sci.* 53, 2101–2105. doi: 10.1007/s10620-007-0135-4

Jeong, H., Baek, S. Y., Kim, S. W., Eun, Y. H., Kim, I. Y., Kim, H., et al. (2017). Comorbidities of rheumatoid arthritis: results from the Korean National Health and Nutrition Examination Survey. *PLoS One* 12:e0176260. doi: 10.1371/journal.pone.0176260

Jones, O. Y., Spencer, C. H., Bowyer, S. L., Dent, P. B., Gottlieb, B. S., and Rabinovich, C. E. (2006). A multicenter case-control study on predictive factors distinguishing childhood leukemia from juvenile rheumatoid arthritis. *Pediatrics* 117, e840–4. doi: 10.1542/peds.2005-1515

Karas, H., Knuppel, R., Schulz, W., Sklenar, H., and Wingender, E. (1996). Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput. Appl. Biosci.* 12, 441–446. doi: 10.1093/bioinformatics/12.5.441

Karlson, E. W., Mandl, L. A., Hankinson, S. E., and Grodstein, F. (2004). Do breast-feeding and other reproductive factors influence future risk of rheumatoid arthritis? Results from the Nurses' Health Study. *Arthritis Rheum.* 50, 3458–3467. doi: 10.1002/art.20621

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., et al. (2010). Variation in transcription factor binding among humans. *Science* 328, 232–235. doi: 10.1126/science.1183621

Kim, J. Y., Lee, E. Y., Park, J. K., Song, Y. W., Kim, J. R., and Cho, K. H. (2016). Patients with rheumatoid arthritis show altered lipoprotein profiles with dysfunctional high-density lipoproteins that can exacerbate inflammatory and atherogenic process. *PLoS One* 11:e0164564. doi: 10.1371/journal.pone.0164564

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626. doi: 10.1038/217624a0

Klareskog, L., Padyukov, L., Lorentzen, J., and Alfredsson, L. (2006). Mechanisms of disease: genetic susceptibility and environmental triggers in the development of rheumatoid arthritis. *Nat. Clin. Pract. Rheumatol.* 2, 425–433. doi: 10.1038/ncprheum0249

Koller, M., and Nobauer-Huhmann, I. (2009). Early arthritis: action desired - treatment required. *Wien. Med. Wochenschr.* 159, 66–69. doi: 10.1007/s10354-009-0653-0

Koopmann, T. T., Adriaens, M. E., Moerland, P. D., Marsman, R. F., Westerveld, M. L., Lal, S., et al. (2014). Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. *PLoS One* 9:e97380. doi: 10.1371/journal.pone.0097380

Krasselt, M., and Baerwald, C. (2017). Sex, symptom severity, and quality of life in rheumatology. *Clin. Rev. Allergy Immunol.* 56, 346–361. doi: 10.1007/s12016-017-8631-6

Kullmann, F., Widmann, T., Kirner, A., Justen, H. P., Wessinghage, D., Dietmaier, W., et al. (2000). Microsatellite analysis in rheumatoid arthritis synovial fibroblasts. *Ann. Rheum. Dis.* 59, 386–389. doi: 10.1136/ard.59.5.386

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi: 10.1093/nar/gkt1113

Lawson, C. A., Brown, A. K., Bejarano, V., Douglas, S. H., Burgoyne, C. H., Greenstein, A. S., et al. (2006). Early rheumatoid arthritis is associated with a deficit in the CD4+CD25high regulatory T cell population in peripheral blood. *Rheumatology* 45, 1210–1217. doi: 10.1093/rheumatology/kel089

Lim, D. S., and Bae, Y. S. (2011). Metastatic lymph node 51 and fibroblast-like synoviocyte hyperproliferation in rheumatoid arthritis pathogenesis. *Rheumatol. Int.* 31, 843–847. doi: 10.1007/s00296-011-1818-x

Liu, M., Sun, H., Wang, X., Koike, T., Mishima, H., Ikeda, K., et al. (2004). Association of increased expression of macrophage elastase (matrix metalloproteinase 12) with rheumatoid arthritis. *Arthritis Rheum.* 50, 3112–3117. doi: 10.1002/art.20567

Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011:baq036. doi: 10.1093/database/baq036

Malemud, C. J. (2018). Defective T-cell apoptosis and t-regulatory cell dysfunction in rheumatoid arthritis. *Cells* 7:223. doi: 10.3390/cells7120223

Malm, K., Bremander, A., Arvidsson, B., Andersson, M. L., Bergman, S., and Larsson, I. (2016). The influence of lifestyle habits on quality of life in patients with established rheumatoid arthritis-A constant balancing between ideality and reality. *Int. J. Qual. Stud. Health Well-being* 11:30534. doi: 10.3402/qhw.v11.30534

Martianov, I., Viville, S., and Davidson, I. (2002). RNA polymerase II transcription in murine cells lacking the TATA binding protein. *Science* 298, 1036–1039. doi: 10.1126/science.1076327

Mitsuyasu, H., Izuhara, K., Mao, X., Gao, P., Arinobu, Y., Enomoto, T., et al. (1998). Ile50Val variant of IL4R alpha upregulates IgE synthesis and associates with atopic asthma. *Nat. Genet.* 19, 119–120. doi: 10.1038/472

Mogno, I., Vallania, F., Mitra, R. D., and Cohen, B. A. (2010). TATA is a modular component of synthetic promoters. *Genome Res.* 20, 1391–1397. doi: 10.1101/gr.106732.110

Nair, N., Wilson, A. G., and Barton, A. (2017). DNA methylation as a marker of response in rheumatoid arthritis. *Pharmacogenomics* 18, 1323–1332. doi: 10.2217/pgs-2016-0195

Ocsko, T., Toth, D. M., Hoffmann, G., Tubak, V., Glant, T. T., and Rauch, T. A. (2018). Transcription factor Zbtb38 downregulates the expression of anti-inflammatory IL1r2 in mouse model of rheumatoid arthritis. *Biochim. Biophys. Acta Gene Regul. Mech.* 1861, 1040–1047. doi: 10.1016/j.bbagrm.2018.09.007

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506, 376–381. doi: 10.1038/nature12873

Olkkonen, J., Kouri, V. P., Hynninen, J., Konttinen, Y. T., and Mandelin, J. (2015). Differentially Expressed in Chondrocytes 2 (DEC2) Increases the Expression of IL-1β and is abundantly present in synovial membrane in rheumatoid arthritis. *PLoS One* 10:e0145279. doi: 10.1371/journal.pone.0145279

Oshchepkov, D., Ponomarenko, M., Klimova, N., Chadaeva, I., Bragin, A., Sharypova, E., et al. (2019). A rat model of human behavior provides evidence of natural selection against underexpression of aggressiveness-related genes in humans. *Front. Genet.* 10:1267. doi: 10.3389/fgene.2019.01267

Peltoketo, H., Piao, Y., Mannermaa, A., Ponder, B., Isomaa, V., Poutanen, M., et al. (1994). A point mutation in the putative TATA box, detected in nondiseased individuals and patients with hereditary breast cancer, decreases promoter activity of the 17 beta-hydroxysteroid dehydrogenase type 1 gene 2 (EDH17B2) in vitro. *Genomics* 23, 250–252. doi: 10.1006/geno.1994.1487

Philips, S., Richter, A., Oesterreich, S., Rae, J. M., Flockhart, D. A., Perumal, N. B., et al. (2012). Functional characterization of a genetic polymorphism in the promoter of the ESR2 gene. *Horm. Cancer* 3, 37–43. doi: 10.1007/s12672-011-0086-2

Plengpanich, W., Le Goff, W., Poolsuk, S., Julia, Z., Guerin, M., and Khovidhunkit, W. (2011). CETP deficiency due to a novel mutation in the CETP gene promoter and its effect on cholesterol efflux and selective uptake into hepatocytes. *Atherosclerosis* 216, 370–373. doi: 10.1016/j.atherosclerosis.2011.01.051

Pocai, B. (2019). The ICD-11 has been adopted by the World Health Assembly. *World Psychiatry* 18, 371–372. doi: 10.1002/wps.20689

Ponomarenko, M., Mironova, V., Gunbin, K., and Savinkova, L. (2013). "Hogness Box," in *Brenner's Encyclopedia of Genetics*, 2nd Edn, eds S. Maloy and K. Hughes (Cambridge: Academic Press), 491–494. doi: 10.1016/B978-0-12-374984-0.00720-8

Ponomarenko, M., Ponomarenko, J., Frolov, A., Podkolodny, N., Savinkova, L., Kolchanov, N., et al. (1999). Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins. *Bioinformatics* 15, 687–703. doi: 10.1093/bioinformatics/15.7.687

Ponomarenko, M., Rasskazov, D., Arkova, O., Ponomarenko, P., Suslov, V., Savinkova, L., et al. (2015). How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed. Res. Int.* 2015:359835. doi: 10.1155/2015/359835

Ponomarenko, M., Rasskazov, D., Chadaeva, I., Sharypova, E., Ponomarenko, P., Arkova, O., et al. (2017). SNP_TATA_Comparator: genomewide landmarks for preventive personalized medicine. *Front. Biosci.* 9, 276–306. doi: 10.2741/s488

Ponomarenko, P., Rasskazov, D., Suslov, V., Sharypova, E., Savinkova, L., Podkolodnaya, O., et al. (2016). Candidate SNP markers of chronopathologies are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Biomed. Res. Int.* 2016:8642703. doi: 10.1155/2016/8642703

Ponomarenko, P. M., Savinkova, L. K., Drachkova, I. A., Lysova, M. V., Arshinova, T. V., Ponomarenko, M. P., et al. (2008). A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl. Biochem. Biophys.* 419, 88–92. doi: 10.1134/S1607672908020117

Priyadarshini, S., Pradhan, B., Griebel, P., and Aich, P. (2018). Cortisol regulates immune and metabolic processes in murine adipocytes and macrophages through HTR2c and HTR5a serotonin receptors. *Eur. J. Cell Biol.* 97, 483–492. doi: 10.1016/j.ejcb.2018.07.004

Putlyaeva, L. V., Demin, D. E., Korneev, K. V., Kasyanov, A. S., Tatosyan, K. A., Kulakovskiy, I. V., et al. (2018). Potential markers of autoimmune diseases, alleles rs115662534(T) and rs548231435(C), disrupt the binding of transcription factors STAT1 and EBF1 to the regulatory elements of human CD40 gene. *Biochemistry* 83, 1534–1542. doi: 10.1134/S0006297918120118

Raychaudhuri, S. K., Abria, C., Maverakis, E. M., and Raychaudhuri, S. P. (2018). IL-9 receptor: regulatory role on FLS and pannus formation. *Cytokine* 111, 58–62. doi: 10.1016/j.cyto.2018.08.001

Rhee, H. S., and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* 483, 295–301. doi: 10.1038/nature10799

Rzepecka, J., Pineda, M. A., Al-Riyami, L., Rodgers, D. T., Huggan, J. K., Lumb, F. E., et al. (2015). Prophylactic and therapeutic treatment with a synthetic analogue of a parasitic worm product prevents experimental arthritis and inhibits IL-1β production via NRF2-mediated counter-regulation of the inflammasome. *J. Autoimmun.* 60, 59–73. doi: 10.1016/j.jaut.2015.04.005

Samet, H. (1985). A top-down quadtree traversal algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* 7, 94–98. doi: 10.1109/tpami.1985.4767622

Sarin, S., Boivin, F., Li, A., Lim, J., Svajger, B., Rosenblum, N. D., et al. (2014). β-Catenin overexpression in the metanephric mesenchyme leads to renal dysplasia genesis via cell-autonomous and non-cell-autonomous mechanisms. *Am. J. Pathol.* 184, 1395–1410. doi: 10.1016/j.ajpath.2014.01.018

Sarkander, J., Hojyo, S., and Tokoyoda, K. (2016). Vaccination to gain humoral immune memory. *Clin. Transl. Immunol.* 5:e120. doi: 10.1038/cti.2016.81

Sato, K., Takahashi, N., Kato, T., Matsuda, Y., Yokoji, M., Yamada, M., et al. (2017). Aggravation of collagen-induced arthritis by orally administered Porphyromonas gingivalis through modulation of the gut microbiota and gut immune system. *Sci. Rep.* 7:6955. doi: 10.1038/s41598-017-07196-7

Schuldberg, D. (2015). What is optimum variability? *Nonlinear Dynamics Psychol. Life Sci.* 19, 553–568.

Scott, D. L., Wolfe, F., and Huizinga, T. W. (2010). Rheumatoid arthritis. *Lancet* 376, 1094–1108. doi: 10.1016/S0140-6736(10)60826-4

Smolen, J. S., and Aletaha, D. (2015). Rheumatoid arthritis therapy reappraisal: strategies, opportunities and challenges. *Nat. Rev. Rheumatol.* 11, 276–289. doi: 10.1038/nrrheum.2015.8

Smolen, J. S., Aletaha, D., and McInnes, I. B. (2016). Rheumatoid arthritis. *Lancet* 388, 2023–2038. doi: 10.1016/S0140-6736(16)30173-8

Somers, T. J., Wren, A. A., Blumenthal, J. A., Caldwell, D., Huffman, K. M., and Keefe, F. J. (2014). Pain, physical functioning, and overeating in obese rheumatoid arthritis patients: do thoughts about pain and eating matter ? *J. Clin. Rheumatol.* 20, 244–250. doi: 10.1097/RHU.0000000000000124

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: perl modules for the life sciences. *Genome Res.* 12, 1611–1618. doi: 10.1101/gr.361602

Stofkova, A., Skurlova, M., Kiss, A., Zelezna, B., Zorad, S., and Jurcovicova, J. (2009). Activation of hypothalamic NPY, AgRP, MC4R, AND IL-6 mRNA levels in young Lewis rats with early-life diet-induced obesity. *Endocr. Regul.* 43, 99–106.

Telenti, A., Pierce, L. C., Biggs, W. H., di Iulio, J., Wong, E. H., Fabani, M. M., et al. (2016). Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U. S. A.* 113, 11901–11906. doi: 10.1073/pnas.1613365113

The 1000 Genomes Project Consortium (GPC), Abecasis, G., Auton, A., Brooks, L., DePristo, M., Durbin, R., et al. (2012). An integrated map of genetic variation from 1.092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Theofanopoulou, C., Gastaldon, S., O'Rourke, T., Samuels, B. D., Martins, P. T., Delogu, F., et al. (2017). Self-domestication in Homo sapiens: insights from comparative genomics. *PLoS One* 12:e0185306. doi: 10.1371/journal.pone.0185306

Thornorsteinsson, V., Magnusson, S., Hellman-Erlingsson, S., Gutmundsdottir, B. R., and Arnason, A. (2004). Congenital deficiency of coagulation factor VII in an Icelandic family. *Laeknabladid* 90, 385–388.

Townsend, R. R., McGinnis, P. A., Tuan, W. M., and Thrasher, K. (1994). Case report: bilateral adrenal pheochromocytoma. *Am. J. Med. Sci.* 308, 123–125. doi: 10.1097/00000441-199408000-00013

Trovato, G. M. (2014). Sustainable medical research by effective and comprehensive medical skills: overcoming the frontiers by predictive, preventive and personalized medicine. *EPMA J.* 5:14. doi: 10.1186/1878-5085-5-14

Um, S., Lee, J. H., and Seo, B. M. (2018). TGF-β2 downregulates osteogenesis under inflammatory conditions in dental follicle stem cells. *Int. J. Oral Sci.* 10:29. doi: 10.1038/s41368-018-0028-8

Varzari, A., Deyneko, I. V., Vladei, I., Grallert, H., Schieck, M., Tudor, E., et al. (2019). Genetic variation in TLR pathway and the risk of pulmonary tuberculosis in a Moldavian population. *Infect. Genet. Evol.* 68, 84–90. doi: 10.1016/j.meegid.2018.12.005

Varzari, A., Tudor, E., Bodrug, N., Corloteanu, A., Axentii, E., and Deyneko, I. V. (2018). Age-specific association of CCL5 gene polymorphism with pulmonary tuberculosis: a case-control study. *Genet. Test. Mol. Biomarkers* 22, 281–287. doi: 10.1089/gtmb.2017.0250

Vasiliev, G., Chadaeva, I., Rasskazov, D., Ponomarenko, P., Sharypova, E., Drachkova, I., et al. (2021). A bioinformatics model of human diseases on the basis of differentially expressed genes (of domestic versus wild animals) that are orthologs of human genes associated with reproductive-potential changes. *Int. J. Mol. Sci.* 22:2346. doi: 10.3390/ijms22052346

Waardenberg, A. J., Basset, S. D., Bouveret, R., and Harvey, R. P. (2015). CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments. *BMC Bioinformatics* 16:275. doi: 10.1186/s12859-015-0701-2

Wang, L., Wang, C., Jia, X., and Yu, J. (2018). Circulating exosomal miR-17 inhibits the induction of regulatory t cells via suppressing TGFBR II expression in rheumatoid arthritis. *Cell. Physiol. Biochem.* 50, 1754–1763. doi: 10.1159/000494793

Wu, J., Wu, M., Li, L., Liu, Z., Zeng, W., and Jiang, R. (2016). dbWGFP: a database and web server of human whole-genome single nucleotide variants and their functional predictions. *Database* 2016:baw024. doi: 10.1093/database/baw024

Yang, X., Zhang, H., Shang, J., Liu, G., Xia, T., Zhao, C., et al. (2018). Comparative analysis of the blood transcriptomes between wolves and dogs. *Anim. Genet.* 49, 291–302. doi: 10.1111/age.12675

Yoo, S. S., Jin, C., Jung, D. K., Choi, Y. Y., Choi, J. E., Lee, W. K., et al. (2015). Putative functional variants of XRCC1 identified by RegulomeDB were not associated with lung cancer risk in a Korean population. *Cancer Genet.* 208, 19–24. doi: 10.1016/j.cancergen.2014.11.004

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). The Ensembl regulatory build. *Genome Biol.* 16:56. doi: 10.1186/s13059-015-0621-5

Zheng, B., Hu, L., Song, X., Wu, Z., Cai, R., He, L., et al. (2014). Analgesic effect of different moxibustion durations in rheumatoid arthritis rats. *J. Tradit. Chin. Med.* 34, 90–95. doi: 10.1016/s0254-6272(14)60060-1

Zorlu, N., Hoffjan, S., Haghikia, A., Deyneko, I. V., and Epplen, J. T. (2019). Evaluation of variation in genes of the arylhydrocarbon receptor pathway for an association with multiple sclerosis. *J. Neuroimmunol.* 334:576979. doi: 10.1016/j.jneuroim.2019.576979

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.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 05 July 2021
doi: 10.3389/fgene.2021.662770

A Catalog of Human Genes Associated With Pathozoospermia and Functional Characteristics of These Genes

Elena V. Ignatieva[1,2], Alexander V. Osadchuk[1], Maxim A. Kleshchev[1], Anton G. Bogomolov[1] and Ludmila V. Osadchuk[1]*

[1] The Federal Research Center Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, [2] Department of Natural Science, Novosibirsk State University, Novosibirsk, Russia

Genetic causes of the global decline in male fertility are among the hot spots of scientific research in reproductive genetics. The most common way to evaluate male fertility in clinical trials is to determine semen quality. Lower semen quality is very often accompanied by subfertility or infertility, occurs in many diseases and can be caused by many factors, including genetic ones. The following forms of lowered semen quality (pathozoospermia) are known: azoospermia, oligozoospermia, asthenozoospermia, teratozoospermia, and some combined forms. To systematize information about the genetic basis of impaired spermatogenesis, we created a catalog of human genes associated with lowered semen quality (HGAPat) and analyzed their functional characteristics. The catalog comprises data on 126 human genes. Each entry of the catalog describes an association between an allelic variant of the gene and a particular form of lowered semen quality, extracted from the experimental study. Most genes included into the catalog are located on autosomes and are associated with such pathologies as non-obstructive azoospermia, oligozoospermia or asthenozoospermia. Slightly less than half of the included genes (43%) are expressed in the testes in a tissue-specific manner. Functional annotation of genes from the catalog showed that spermatogenic failure can be associated with mutations in genes that control biological processes essential for spermiogenesis (regulating DNA metabolism, cell division, formation of cellular structures, which provide cell movement) as well as with mutations in genes that control cellular responses to unfavorable conditions (stress factors, including oxidative stress and exposure to toxins).

Keywords: male fertility, infertility, spermatogenesis, pathozoospermia, genetic polymorphism, catalog of genes

OPEN ACCESS

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

Reviewed by:
Liina Nagirnaja, Oregon Health & Science University, United States
Kajal Khodamoradi, University of Miami, United States

****Correspondence:***
Elena V. Ignatieva
eignat@bionet.nsc.ru

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

Received: *01 February 2021*
Accepted: *26 April 2021*
Published: *05 July 2021*

Citation:
Ignatieva EV, Osadchuk AV, Kleshchev MA, Bogomolov AG and Osadchuk LV (2021) A Catalog of Human Genes Associated With Pathozoospermia and Functional Characteristics of These Genes. Front. Genet. 12:662770. doi: 10.3389/fgene.2021.662770

INTRODUCTION

Currently, the global demographic crisis in industrialized countries, including Russia, is accompanied by a decrease in the reproductive potential of human populations. Over the past 60 years, a temporary trend of declining semen quality has been observed, which was expressed as a decrease in sperm counts, increasing prevalence of male infertility and incidence of some urological

diseases and congenital abnormalities of the male reproductive system (Joffe, 2010; Levine et al., 2017; Krausz and Riera-Escamilla, 2018).

Understanding all the factors, influencing on male fertility, is of great importance in reproductive medicine for diagnostics and treatment. In clinical practice and population studies, semen analysis is the corner stone of work-up, diagnosis, and treatment of men in the subfertile or infertile couples. The recent WHO laboratory manual for the examination and processing of human semen identified and recommended a set of methods for performing semen analysis (World Health Organization., 2010). The terms used to describe semen quality, as well as the reference values for semen parameters were given in the WHO laboratory manual (World Health Organization., 2010) and are partially presented in **Supplementary Table 1** of this paper. Important steps of microscopic investigation of semen include evaluation of sperm concentration, motility, and morphological characteristics. A decrease in the quantitative and/or qualitative semen indicators (one or more) below the reference values is recognized as a lowered semen quality and is most often the cause of male infertility. For example, 755 men suffering from infertility or disorders of the reproductive system were examined at the Moscow Medical and Genetic Research Center, and lowered semen quality was detected in 89% of cases (Andreeva et al., 2017). Lowered semen quality (called pathozoospermia) exists in various forms, including azoospermia, oligozoospermia, asthenozoospermia, teratozoospermia, and some combinations thereof. Azoospermia (absence of spermatozoa in the ejaculate) and severe oligozoospermia (sperm concentration is less than 5 million/ml) represent the most severe forms of male infertility, and men with azoospermia have the highest risk of genetic disorders (25%) (Krausz et al., 2018).

Currently, the following most common genetic causes of male infertility are clearly documented: microdeletions in AZFa, AZFb, and AZFc regions of the AZF locus of the Y chromosome (Kamp et al., 2001; Fernandes et al., 2002, 2006; Krausz et al., 2015; Kim et al., 2017), androgen receptor (*AR*) mutations, including CAG repeat polymorphism in exon 1 (Xiao et al., 2016), chromosomal abnormalities (Alves et al., 2002; Aston and Conrad, 2013; Madureira et al., 2014), cystic fibrosis gene (*CFTR*) polymorphism (Vallières and Elborn, 2014; Jiang et al., 2017). The frequencies of the most common genetic abnormalities associated with male infertility are as follows: chromosomal abnormalities 5–7%, Y-chromosome deletions 5–10%, *CFTR* gene mutations 5%, *AR* mutations 2–3% (Ferlin et al., 2006; Chernykh, 2009). The standard genetic diagnostic examination of infertile men includes karyotype analysis and screening for AZF microdeletions. Karyotype abnormalities and AZF microdeletions are most often detected in men with a sperm concentration less than 5 million/ml (Chernykh, 2009). Screening for *CFTR* mutations in infertile men is carried out after identification of congenital absence of the vas deferens, since mutations in the *CFTR* gene can cause both a severe hereditary disease of cystic fibrosis and aplasia of the vas deferens (Havasi et al., 2010; Ahmad et al., 2013; Jiang et al., 2017). At present, the analysis for the CAG polymorphism of the *AR* gene is not recommended to use in routine diagnostic practice; however, in some cases, it is performed before androgen replacement therapy of patients (Francomano et al., 2013). Moreover, sperm apoptosis markers (Almeida et al., 2005, 2013), DNA fragmentation test (Sá et al., 2015), estimation of sperm DNA methylation (Marques et al., 2004, 2010) and testicular histology (Sousa et al., 2002) are important for male fertility assessment.

The above-mentioned genetic abnormalities contribute significantly to the understanding the nature of male infertility; however, in humans, a set of gene loci associated with reproductive disorders caused by impaired spermatogenesis is much wider. The use of candidate gene approach and some modifications of the whole genome approach (whole-genome and whole-exome sequencing) made it possible to identify other genetic loci (in addition to *AR* and *CFTR* genes, as well as AZF deletions) associated with impaired spermatogenic function (Ferrás et al., 2007; Esteves, 2013; Zhang et al., 2013; Krausz et al., 2015, 2018; Pereira et al., 2015, 2017, 2019; He et al., 2019; Oud et al., 2019; Araujo et al., 2020).

Thus, a large number of genes associated with specific forms of lowered semen quality are currently known. Data on associations are presented both in scientific publications and in non-specialized databases on phenotype-genotype associations: OMIM[1], ClinVar[2], HGMD[3], PheGenI[4], EGA[5], GAD[6] and dbGaP[7]. There is also an Internet portal of the project "Male Fertility Gene Atlas CRU Male Germ Cells" (MFGA)[8], which accumulates data on genes associated with various disorders of the male reproductive system, including some forms of lowered semen quality. However, as it turned out, each of these databases contains a far from complete set of genes associated with lowered semen quality (**Supplementary Table 2**). In addition, information may be inaccurate (for example, some genes are incorrectly associated with the trait, or the study of the relationship between the gene and the trait was carried out on a model species (see the comments to **Supplementary Table 2**). Thus, even with a number of databases on phenotype-genotype associations, additional efforts are needed to extract data on genes associated with specific forms of lowered semen quality and integrate these data, bringing them to a unified format.

The aim of this study was to create a set of genes with polymorphic loci associated with non-syndromic forms of lowered semen quality, as the frequent genetic reason of male infertility. We conducted a search for such genes and their polymorphic loci in scientific publications, and the collected data were presented in the form of a catalog. The functional annotation of the genes from our catalog was then carried out. As a result, we identified a set of biological processes, the genetic control of which can be impaired in individuals with lowered

[1]https://www.ncbi.nlm.nih.gov/omim
[2]https://www.clinicalgenome.org/data-sharing/clinvar/
[3]http://www.hgmd.cf.ac.uk/ac/index.php
[4]https://www.ncbi.nlm.nih.gov/gap/phegeni
[5]https://ega-archive.org/
[6]https://geneticassociationdb.nih.gov/
[7]https://www.ncbi.nlm.nih.gov/gap/
[8]https://mfga.uni-muenster.de/projectInfo.html

semen quality. The genetic data systematized as a catalog might help the development of genetic panels for the diagnosis of male infertility.

MATERIALS AND METHODS

Extracting Data From Publications

The data was extracted from the publications manually. To find such articles in PubMed[9] we performed queries to a number of informational resources (databases and information systems). We used search terms referring to various conditions manifested in lowered semen quality (**Supplementary Table 1**). Four terms (*non-obstructive azoospermia, cryptozoospermia, oligozoospermia, and severe oligozoospermia*) designated conditions associated with a decrease in the number of spermatozoa in the ejaculate. The term *asthenozoospermia* denoted a pathology manifested as an impairment of sperm motility. The terms *teratozoospermia* and its subtype *globozoospermia* denoted conditions characterized by the presence of sperm with abnormal morphology. Since sperm morphological abnormalities are often accompanied by a decrease in sperm motility (Maettner et al., 2014), and both abnormalities in morphology and/or decreased motility can be observed against the background of a decrease in sperm concentration (Guzick et al., 2001; Menkveld et al., 2001), the vocabulary included terms denoting combined forms of lowered semen quality (*asthenoteratozoospermia, oligoasthenozoospermia, oligoteratozoospermia, oligoasthenoteratozoospermia*). The term *azoospermia* was not included in the number of keywords. This was due to the fact that the term *azoospermia* denotes two different conditions: (1) obstructive azoospermia, when spermatogenesis was not disturbed, but an absence of sperm in the ejaculate occurred as a result of mechanical obstruction in genital paths; (2) non-obstructive azoospermia, when an absence of sperm in the ejaculate was due to impaired spermatogenesis, with normal state of the seminal vesicles and vas deferens (Gamidov et al., 2015). Obstructive azoospermia, as a rule, is caused by infections or inflammation of the reproductive tract, trauma and other external factors, and less often by gene mutations (mutations in the *CFTR* gene), and in this case, azoospermia may be present in the clinical picture of the disease as a minor sign. Therefore, only the term *non-obstructive azoospermia* was included in the keywords.

All these forms of lowered semen quality listed in **Supplementary Table 1** will hereinafter be referred to as forms of pathozoospermia.

The following databases on phenotype-genotype associations were used as sources of information: (1) OMIM (see text footnote 1); (2) ClinVar (see text footnote 2); (3) HGMD (see text footnote 3); (4) PheGenI (see text footnote 4); (5) EGA (see text footnote 5); (6) dbGaP (see text footnote 7); (7) MFGA (see text footnote 8). We also used ANDSystem[10] (Ivanisenko et al., 2019) as an additional source of data. ANDSystem contains data on associations between genes and diseases obtained through automatic text-mining analysis of texts collected in PubMed.

Data Processing

The description of genes was provided with identifiers, official symbols and data on localization on the chromosome obtained from the Entrez Gene database[11].

Each form of lowered semen quality was assigned to one of three categories (manifestations): (1) low sperm count; (2) reduced sperm motility; (3) abnormal sperm morphology. Combined forms were assigned to two or three categories respectively. The relations between each pathology and an appropriate category (categories) are presented in **Supplementary Table 1**.

The location of the variant in the gene region (exon, intron, etc.) was extracted from the research paper or from databases (UCSC genome browser)[12] or dbSNP[13]. Variant genomic location was extracted from dbSNP. If dbSNP rs identifier was not specified in the article, it was identified by queries to databases: (1) to UCSC genome browser if the DNA sequence harboring polymorphic locus was known; (2) to dbSNP or ClinVar[14] if several variant names (aliases) were found in the research paper (**Supplementary Figure 1**).

Web Interface of the Catalog

We used an open-source relational database management system MariaDB 10.2.12 (MariaDB Corporation AB)[15] to provide access to data from catalog. The web interface was developed with PHP 5.3.3 and it is accessible at https://www.sysbio.ru/hgap/.

Identification of the Testis-Specific and Testis-Enriched Genes

Tissue-specific genes, that are expressed in the testes, were identified using the TSEA tool (Wells et al., 2015). TSEA tool[16] uses data on tissue enrichment score of gene expression products (SI) and the corresponding pSI values obtained from the analysis of RNA-seq data across 45 tissue types from the healthy, adult human body (Melé et al., 2015). Depending on the pSI value, calculated for a given tissue, the transcript was considered to be expressed in a tissue-specific manner for this tissue (at $pSI < 0.01$) or to be expressed in a tissue-enriched manner (at $pSI < 0.05$).

Assignment of Genes to Functional Categories

The Database for Annotation, Visualization and Integrated Discovery web-based Functional Annotation Tool (DAVID tool) was applied to the sets of genes from the catalog (Huang et al., 2009). The DAVID tool allowed us to detect

[9] https://pubmed.ncbi.nlm.nih.gov/

[10] http://www-bionet.sscc.ru/and/cell/

[11] https://www.ncbi.nlm.nih.gov/gene

[12] https://genome.ucsc.edu/

[13] https://www.ncbi.nlm.nih.gov/snp/

[14] https://www.ncbi.nlm.nih.gov/clinvar/

[15] https://mariadb.org/

[16] http://genetics.wustl.edu/jdlab/tsea/

the number of genes annotated to each GO term and to identify GO terms that are highly associated with a given gene set (overrepresented GO terms). FDR = 0.05 was used as a threshold criterion characterizing the statistical significance of the excess (enrichment) of the observed number of associations of genes with a specific GO term in comparison with the expected number of associations. The overrepresented GO terms from the Biological Processes vocabulary (GOTERM_BP, GOTERM_BP_5) and Cellular Component vocabulary (GOTERM_CC_5) were considered in our study.

The identification of representative terms characterizing the set of significantly overrepresented GO terms was carried out using the REViGO software[17] (Supek et al., 2011).

RESULTS

The Catalog of Genes Associated With Different Forms of Lowered Semen Quality Caused by Impaired Spermatogenesis (HGAPat): Web Presentation and Information Content

Using the selected terms (**Supplementary Table 1**) as the keywords and the data collection method described above, we found 126 genes with polymorphic loci associated with various forms of pathozoospermia.

These data are available on the Internet as a relational database. The database consists of four tables: *Genes, Disease, PubGenes*, and *GeneticVariant* (**Supplementary Figure 2** shows the database scheme). The *Genes* table contains identifiers, symbols, full names, chromosomal localization, expression pattern and GO annotation of genes. The *Disease* table contains forms of pathozoospermia and their manifestations. The *PubGenes* table contains the name of the gene given in the article, if it differs from the official gene symbol given in Entrez Gene database. The *GeneticVariant* table is the main table that contains data on associations between allelic variants of genes and specific forms of pathozoospermia. So the data are presented in 25 information fields, including 6 identifiers.

The web-interface of HGAPat allows users to observe the collected data and use simple filters: select data by gene, by form of pathozoospermia, by ethnic group, by odds ratio.

Now the catalog HGAPat contains data on 126 genes and 260 variants extracted from 111 publications (**Figure 1A**). These publications present the results of studies carried out on samples of individuals from 47 different human populations. The largest number of genes was found in Chinese (48 genes) and European (47 genes) populations, as well as in Japanese (11 genes) one.

Most of the genes (89.6% or 113 genes) are located on autosomes, 4.0% and 2.4% (5 and 3 genes) are located on the X and Y-chromosomes, respectively, and 4.0% (5 genes) are located on mitochondrial DNA (**Figure 1B**). The largest number of genes is associated with such forms of pathozoospermia as non-obstructive azoospermia, oligozoospermia, and asthenozoospermia (**Figure 1C**).

About sixty percent of the variants (145) described in catalog have dbSNP rs identifiers (**Figure 1A**) and in half of the cases (74) the identifiers were determined as a result of additional data processing (**Supplementary Figure 1**). Three quarters of variants are located in exons, 13% are located in introns, 2% are located in UTRs, 2% are located in 5′-near gene or in promoter regions, the rest variants have other location or data on location is not presented.

According to the TSEA tool (see text footnote 16), 43% of genes (54 out of 126) are expressed in the testes in a tissue-specific manner (assigned by the TSEA tool to the testis-specific and testis-enriched categories (**Figure 1D** and **Supplementary Table 4**). The remaining 72 genes (57% of the total) are not testis-specific or testis-enriched (in **Figure 1D** this group of genes is denoted as *Other).*

Functional Annotation of Genes

The DAVID tool was used to identify biological processes, the genetic control of which may be impaired in patients with pathozoospermia. Using DAVID tool we detected GO terms associated with genes from the catalog more frequently than it was expected by chance. We found the following overrepresented terms from the *Cellular Component* vocabulary: (1) terms denoting the chromosomal localization of proteins (*condensed chromosome, chromosomal part, nuclear chromosome part, etc.*); (2) terms indicating association with the synaptonemal complex (*synaptonemal complex*), cilia (*cilium, motile cilium, ciliary plasm*), axoneme (*axoneme, axonemal dynein complex, dynein complex*), sperm-specific voltage-gated calcium channel (*CatSper complex*) (**Supplementary Table 5**).

Among the GO terms denoting biological processes, 63 overrepresented (FDR < 0.05) terms were identified (**Supplementary Table 6**). Using the REViGO program (Supek et al., 2011), we identified sixteen representative terms (**Figure 2**). These GO terms denoted: (1) development and formation of gametes (*gamete generation, cell development, reproductive system development, gonad development, cell maturation, positive regulation of male gonad development*); (2) metabolic processes associated with DNA (*DNA metabolic process, DNA methylation, DNA modification*); (3) processes associated with cell division (*nuclear division, nuclear chromosome segregation, DNA recombination, saynaptonemal complex organization, asymmetric stem cell division*); (4) motility of the cell (*sperm motility*); (5) hormonal regulation (*response to steroid hormone*).

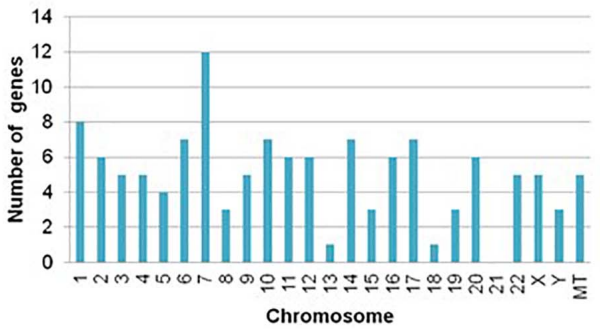
Then, using the DAVID tool, we analyzed the list of genes from the catalog (72 genes) that were not categorized as testis-specific or testis-enriched (in the **Figure 1D** these genes were referred to the category *Other*) and found more than seventy over-represented (FDR < 0.05) GO terms denoting biological processes. Among them twenty nine terms were representative according to REViGO (**Supplementary Table 7**). Most of the representative GO terms found at this step were associated either with the development and formation of cellular organelles (*cellular developmental process, positive regulation of cell differentiation, cellular component*

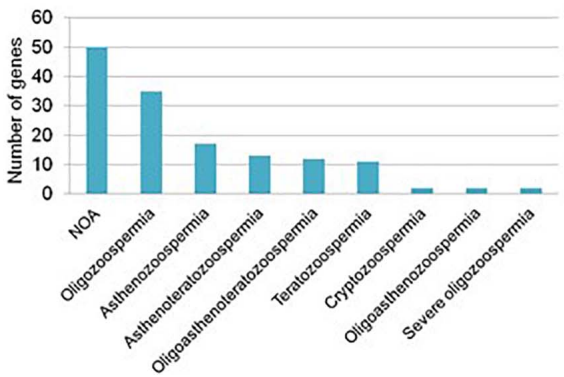
[17] http://revigo.irb.hr/

A

Data type	Number
Genes	126
Infertility phenotypes (forms of sperm abnormalities)	10
Variants	260
Variants with dbSNP rs identifiers	145[a]
Variants without dbSNP rs identifiers	115
Populations	47
Publications	111

[a] - 71 IDs were obtained from research papers and 74 IDs were determined by queries to databases





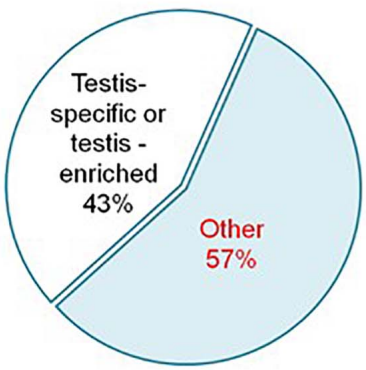


FIGURE 1 | Informational content of the catalog HGAPat. **(A)** HGAPat statistics. Queries to databases were performed according to the scheme presented in **Supplementary Figure 1**. **(B)** Chromosomal localization of genes. *MT* stands for genes, located in mitochondrial DNA. **(C)** The number of genes associated with a specific form of pathozoospermia (NOA, *non-obstructive azoospermia*). When calculating the number of genes associated with *Teratozoospermia*, genes associated with *Globozoospermia* were taken into account. **(D)** Proportion of genes, expressed in a testis-specific or testis-enriched manner.

assembly, cellular component organization, or biogenesis, etc.) or with DNA processing (*reciprocal DNA recombination, DNA metabolic process*) or with cell movement (*locomotion, cell motility, localization of cell*). Thus, the set of GO terms identified at this stage specified approximately the same functional characteristics of genes as the set of GO terms identified for the complete list of genes from the catalog (**Supplementary Table 6** and **Figure 2**). In addition, we identified GO terms that denoted the response to unfavorable conditions (*response to oxidative stress, response to oxygen-containing compound, response to stress, response to toxic substance*). These GO terms were associated with 30 genes (in **Supplementary Table 7** these genes and GO terms are shown in red).

To compile a complete list of genes related to the response to unfavorable conditions we looked again at the results obtained by DAVID tool for all 126 genes from the catalog and found sets of genes associated with these four GO terms (**Supplementary Table 8**). In this way we found 44 genes, fourteen of which were testis-specific or testis-enriched (**Figure 3**).

DISCUSSION

The Catalog of Genes Associated With Pathozoospermia

The goal of the current study was to characterize the genetic basis of various non-syndromic forms of pathozoospermia. To achieve this goal, we collected data on genes and their polymorphic loci associated with lowered semen quality caused by impaired spermatogenesis and presented data in the form of a catalog that is publicly available via Internet. Thus, a specialized information resource HGAPat was created that contained more data on the indicated topic than any known open information resource. As an example, **Supplementary Table 2** presents the number

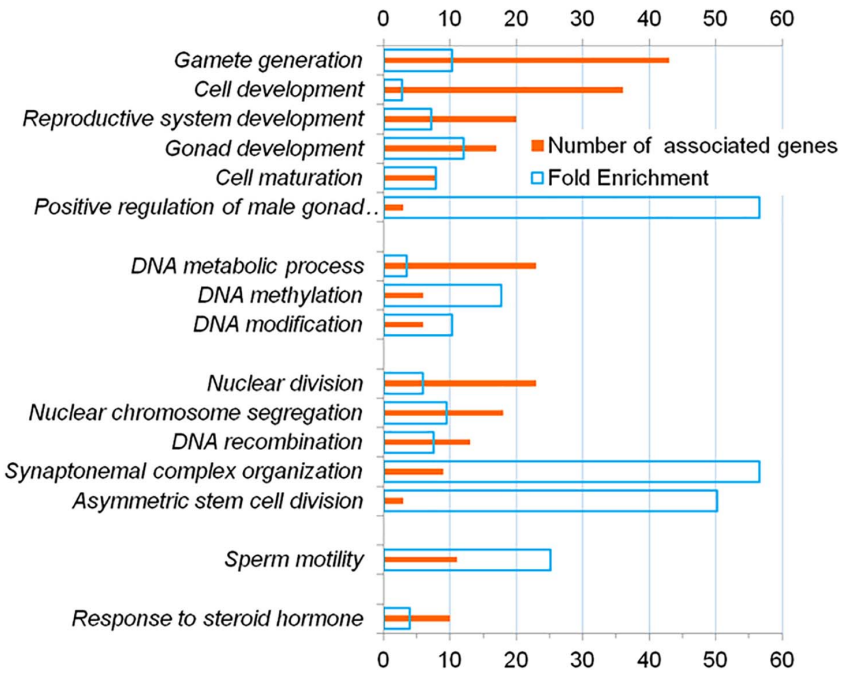


FIGURE 2 | Representative terms from the Gene Ontology database which, according to DAVID, are overrepresented (FDR $< 5 * 10^{-2}$) in the annotations of genes from the catalog HGAPat.

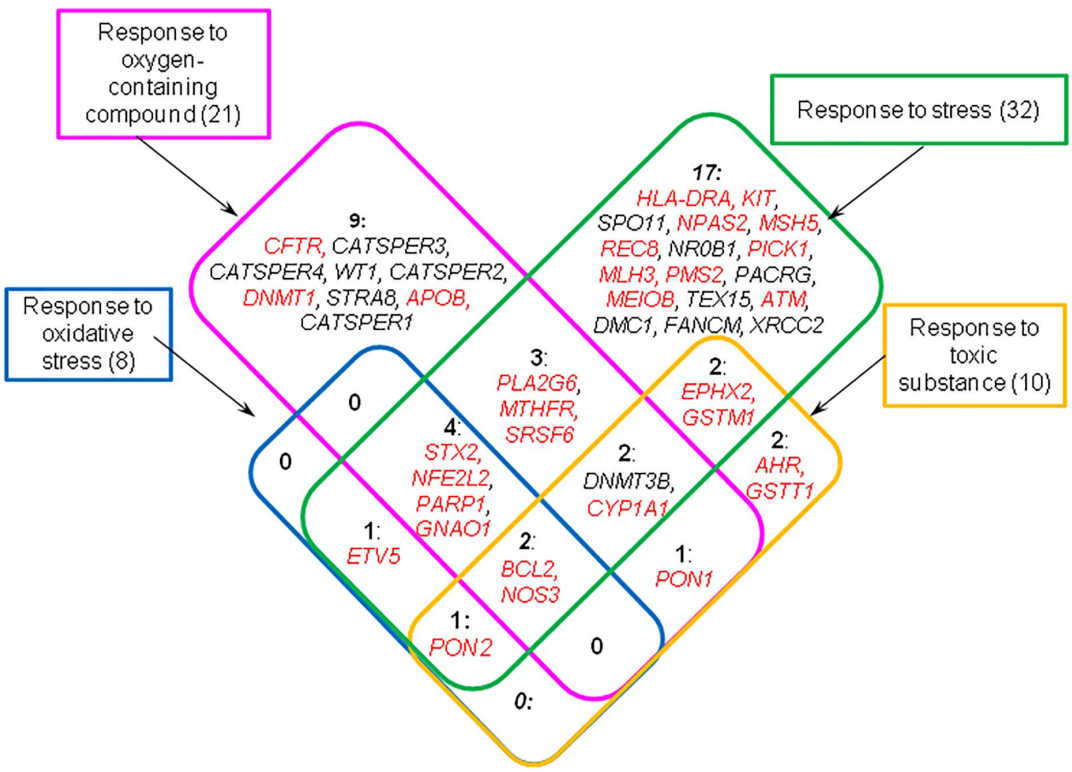


FIGURE 3 | Genes from the catalog HGAPat associated with GO terms designating response to unfavorable conditions (these terms are presented in rectangular boxes). The numbers indicate the number of genes in each group. Genes that are expressed in testis-enriched or testis-specific manner are shown in black; the *other* genes are shown in red.

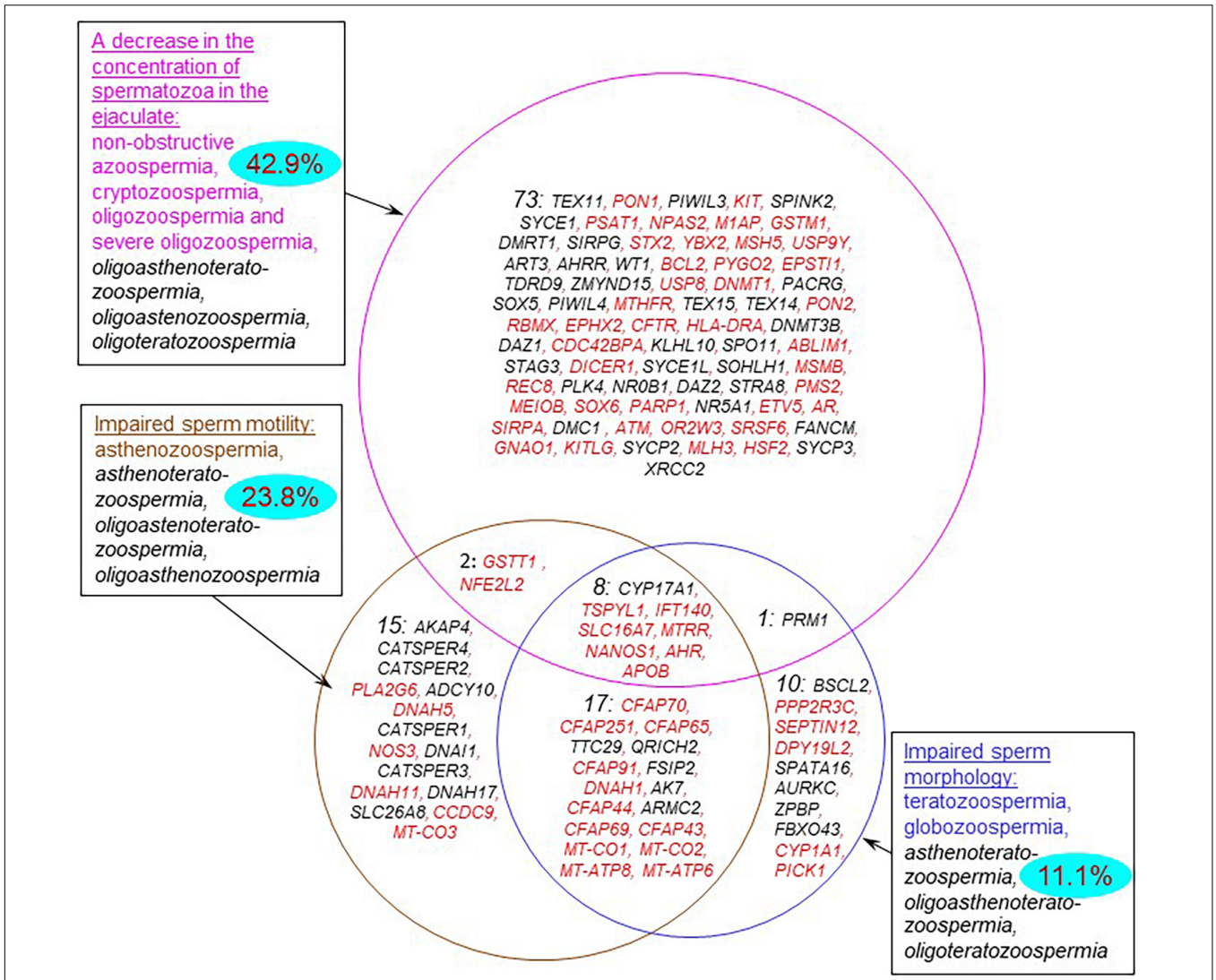


FIGURE 4 | Genes from the catalog associated with various forms of sperm abnormalities. The forms of pathozoospermia are presented in rectangular frames. Combined forms of pathozoospermia are shown in black and *italics*. The numbers in each part of the circles indicate the numbers of genes in each group. The blue ovals show the percentage of genes that control the response to adverse factors (*response to oxidative stress, response to oxygen-containing compound, response to stress, response to toxic substance)*. Genes that are expressed in testis-enriched or testis-specific manner are shown in black; the other genes are shown in red.

of genes contained in the catalog and associated with two uncombined forms of pathozoospermia (asthenozoospermia and teratozoospermia), as well as with the considered subtype of azoospermia (non-obstructive azoospermia). The volumes of these data exceed those available in other well-known databases. Likewise HGAPat contains more genes associated with non-syndromic forms of pathozoospermia than the most recent publication (Oud et al., 2019) summarizing knowledge on 521 male infertility genes. We revealed that only 59 genes presented by Oud et al. (2019) (**Supplementary Figure 3**) met the criteria used in our study (had at least one potentially pathogenic variant described and had relationships with non-syndromic forms of pathologies listed in **Supplementary Table 1**).

The data we systematized can serve as a basis for the development of prognostic criteria for subfertility and infertility in men based on genetic screening, which is an important but not resolved problem of reproductive medicine.

Semen abnormalities associated with impaired spermatogenesis (**Supplementary Table 1**) manifest themselves as: (1) a decrease in sperm concentration; (2) abnormal sperm morphology; and (3) decreased sperm motility. There are the combined forms. In accordance with these manifestations of pathozoospermia, the genes from the catalog can be classified into three groups (**Figure 4**). The largest number of genes ($n = 84$) was associated with the forms of pathozoospermia associated with a decrease in sperm concentration. Fewer genes were associated with the forms of pathozoospermia due to disturbances in sperm morphology and motility (36 and 42, respectively). Thus, it was found a different degree of genetic heterogeneity of these manifestations of pathozoospermia. Perhaps this can

be explained by the different intensity of the study of these manifestations, since the assessment of sperm morphology is a more complex clinical test than the measurement of sperm concentration or the assessment of sperm motility.

Most of genes (77%) were associated with an impairment of one semen parameter (a decrease in sperm concentration, or normal morphology, or motility). At the same time, the groups of genes associated with impaired sperm morphology and motility contained a significant proportion of genes associated with combined forms of pathozoospermia (72% and 64%), especially with asthenoteratozoospermia (25 genes). This finding was in a good agreement with the fact that impaired morphology is one of the possible reasons for a decrease in sperm motility (Maettner et al., 2014).

Functional Annotation of Genes

The subsequent functional annotation of genes from the catalog allowed us to identify a set of biological processes, the genetic control of which may be disturbed in pathozoospermia.

With the help of DAVID tool, we revealed a tight association of genes from the catalog with a number of biological processes important for the formation of spermatozoa (*DNA recombination, nuclear chromosome segregation, synaptonemal complex organization, sperm motility, etc.* (**Supplementary Table 6** and **Figure 2**) and with the basic cellular structures of spermatozoa (*condensed chromosome, synaptonemal complex, motile cilium* etc.; **Supplementary Table 5**). The results obtained are in good agreement with the idea that the cellular processes mentioned above (DNA replication, chromosome segregation, formation of a synaptonemal complex) are essential for the formation of a mature sperm cell (Hann et al., 2011; Fowler et al., 2019) and at the same time a complex cellular apparatus is formed that ensures the movement of spermatozoa (O'Donnell, 2014; Pereira et al., 2017; Cannarella et al., 2020). In particular, proteins involved to synaptonemal complex assembly are crucial for spermatogenesis (Wellard et al., 2020). It was shown that infertile men were characterized by increased levels of sperm aneuploidy likely due to increased errors in meiotic recombination and chromosome synapsis (Tempest, 2011).

We found that only 43% of genes are expressed in a testis-specific or testis-enriched manner (**Figure 1D** and **Supplementary Table 4**), and the rest of genes do not belong to this category. Our data are in agreement with those of the other researchers who showed that only 933 of 3,580 genes differentially expressed in the testes of infertile and fertile men were undetectable in 45 embryonic and adult non-testicular tissues (Chalmel et al., 2012). The results obtained confirm the idea that spermatogenesis is closely related with health of the whole organism (Omu, 2013).

In order to identify the most significant biological processes associated with the rest part of genes (which we categorized as *others*, **Figure 1D**), we performed the additional analysis using the DAVID tool. We found 29 representative biological processes associated with genes from this group (**Supplementary Table 7**). Many of the processes found at this step were directly related to the formation of spermatozoa and their cellular functions (*cellular developmental process, locomotion, cell motility, DNA metabolic process*), which was in good agreement with the results obtained for the complete list of genes. Along with this, we identified a group of 44 genes (one third of the total number of genes) associated with the response to adverse environmental factors (**Figure 4** and **Supplementary Table 8**). Among them there were the genes encoding proteins, the protective role of which is very well studied. For example, *GSTM1, GSTT1, CYP1A1*, and *EPHX2* encode xenobiotic-metabolizing enzymes (Pande et al., 2008; Decker et al., 2009). *AHR* encodes ligand-activated transcription factor involved in the regulation of biological responses to aromatic hydrocarbons and oxidative stress (Dietrich, 2016). *PON1* and *PON2* encode enzymes that modulate oxidative stress and inflammation (Furlong et al., 2016).

The identification of such a group of genes indicates that the genes, which control the response to adverse factors, can play an important role in maintaining male reproductive function, namely, in maintaining adequate sperm quality. The largest fraction of such genes (42.9%) was found in the set of genes associated with a low sperm count (**Figure 4**). Smaller fractions were found among the sets of genes associated with reduced sperm motility and abnormal morphology (23.8 and 11.1%, respectively). However, the identification of genes controlling the response to unfavorable conditions among the genes of all three groups suggests that three manifestations of pathozoospermia (a decrease in sperm count, abnormal sperm morphology and motility) may be associated with unfavorable environmental conditions. Negative external factors including environmental pollution, smoking, oxidative stress etc. are known to affect male fertility, resulting in decreasing sperm count, sperm motility and percentage of morphologically normal sperm as well as sperm DNA damage (Omu, 2013; Lafuente et al., 2016). Our results are in agreement with these evidences, demonstrating that genetically determined sustainability to environmental factors is crucial to male fertility.

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

EI performed search for publications and manual extraction of data from the articles, processed data, created the general scheme of the catalog HGAPat, interpreted the data, and wrote the manuscript. AO interpreted the data and corrected the manuscript. MK performed manual extraction of data on genetic variants from the articles and interpreted the data. AB developed the software, wrote, and corrected the manuscript. LO provided overall supervision of the project, interpreted the data, and corrected the manuscript. All authors read and approved submission of this manuscript.

FUNDING

This study was supported by grant from the Russian Science Foundation (No. 19-15-00075).

ACKNOWLEDGMENTS

The general scheme of the HGAPat catalog was done, using the Bioinformatics Shared Access Center supported by the budget project no. 0259-2021-0009. The software and web-interface for the HGAPat catalog were developed using resources of the Common Use Center for Microscopy of Biologic Objects, supported by the budget project no. 0259-2021-0011.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.662770/full#supplementary-material

REFERENCES

Ahmad, A., Ahmed, A., and Patrizio, P. (2013). Cystic fibrosis and fertility. *Curr. Opin. Obstet. Gynecol.* 25, 167–172. doi: 10.1097/GCO.0b013e32835f1745

Almeida, C., Cardoso, M. F., Sousa, M., Viana, P., Gonçalves, A., Silva, J., et al. (2005). Quantitative study of caspase-3 activity in semen and after swim-up preparation in relation to sperm quality. *Hum. Reprod.* 20, 1307–1313. doi: 10.1093/humrep/deh727

Almeida, C., Correia, S., Rocha, E., Alves, A., Ferraz, L., Silva, J., et al. (2013). Caspase signalling pathways in human spermatogenesis. *J. Assis. Reprod. Gene.* 30, 487–495. doi: 10.1007/s10815-013-9938-8

Alves, C., Carvalho, F., Cremades, N., Sousa, M., and Barros, A. (2002). Unique (Y;13) translocation in a male with oligozoospermia: cytogenetic and molecular studies. *EJHG* 10, 467–474. doi: 10.1038/sj.ejhg.5200835

Andreeva, M. V., Khayat, S. S. H., Sorokina, T. M., Chernykh, V. B., Schileyko, L. V., Shtaut, M. I, et al. (2017). Types of pathozoospermia in men with infertility in marriage and/or disorders of reproductive system. *Androl. Genit. Surg.* 18, 33–38. doi: 10.17650/2070-9781-2017-18-2-33-38

Araujo, T. F., Friedrich, C., Grangeiro, C. P., Martelli, L. R., Grzesiuk, J. D., Emich, J., et al. (2020). Sequence analysis of 37 candidate genes for male infertility: challenges in variant assessment and validating genes. *Andrology* 8, 434–441. doi: 10.1111/andr.12704

Aston, K. I., and Conrad, D. F. (2013). A review of genome-wide approaches to study the genetic basis for spermatogenic defects. *Spermatogenesis* 927, 397–410. doi: 10.1007/978-1-62703-038-0_34

Cannarella, R., Condorelli, R. A., Mongioì, L. M., La Vignera, S., and Calogero, A. E. (2020). Molecular biology of spermatogenesis: novel targets of apparently idiopathic male infertility. *Int. J. Mol. Sci.* 21:1728. doi: 10.3390/ijms21051728

Chalmel, F., Lardenois, A., Evrard, B., Mathieu, R., Feig, C., Demougin, P., et al. (2012). Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. *Hum. Reprod.* 27, 3233–3248. doi: 10.1093/humrep/des301

Chernykh, V. B. (2009). AZF deletions are a common genetic cause of male infertility: the current state of research. *Prob. Reprod.* 1, 10–15.

Decker, M., Arand, M., and Cronin, A. (2009). Mammalian epoxide hydrolases in xenobiotic metabolism and signalling. *Arch. Toxicol.* 83, 297–318. doi: 10.1007/s00204-009-0416-0

Dietrich, C. (2016). Antioxidant functions of the aryl hydrocarbon receptor. *Stem Cells Int.* 2010:7943495. doi: 10.1155/2016/7943495

Esteves, S. C. (2013). A clinical appraisal of the genetic basis in unexplained male infertility. *J. Hum. Reprod. Sci.* 6:176. doi: 10.4103/0974-1208.121419

Ferlin, A., Arredi, B., and Foresta, C. (2006). Genetic causes of male infertility. *Reproduc. Toxicol.* 22, 133–141. doi: 10.1016/j.reprotox.2006.04.016

Fernandes, A. T., Fernandes, S., Gonçalves, R., Sá, R., Costa, P., Rosa, A., et al. (2006). DAZ gene copies: evidence of Y chromosome evolution. *Mol. Hum. Reprod.* 12, 519–523. doi: 10.1093/molehr/gal051

Fernandes, S., Huellen, K., Goncalves, J., Dukal, H., Zeisler, J., Rajpert De , E., et al. (2002). High frequency of DAZ1/DAZ2 gene deletions in patients with severe oligozoospermia. *Mol. Hum. Reprod.* 8, 286–298. doi: 10.1093/molehr/8.3.286

Ferrás, C., Zhou, X. L., Sousa, M., Lindblom, A., and Barros, A. (2007). DNA mismatch repair gene hMLH3 variants in meiotic arrest. *Fertil. Steril.* 88, 1681–1684. doi: 10.1016/j.fertnstert.2007.01.063

Fowler, K. E., Mandawala, A. A., and Griffin, D. K. (2019). The role of chromosome segregation and nuclear organisation in human subfertility. *Biochem. Soc. Transac.* 47, 425–432. doi: 10.1042/BST20180231

Francomano, D., Greco, E. A., Lenzi, A., and Aversa, A. (2013). CAG repeat testing of androgen receptor polymorphism: is this necessary for the best clinical management of hypogonadism? *J. Sex. Med.* 10, 2373–2381. doi: 10.1111/jsm.12268

Furlong, C. E., Marsillach, J., Jarvik, G. P., and Costa, L. G. (2016). Paraoxonases-1, -2 and -3: what are their functions? *Chem. Biol. Interact.* 259, 51–62. doi: 10.1016/j.cbi.2016.05.036

Gamidov, S. I, Popova, A. Y. U., and Ovchinnikov, R. I (2015). Nonobstructive azoospermia - clinical recommendations. *Russkiy. Meditsinskiy. Zhurnal.* 11:595.

Guzick, D. S., Overstreet, J. W., Factor-Litvak, P., Brazil, C. K., Nakajima, S. T., Coutifaris, C., et al. (2001). Sperm morphology, motility, and concentration in fertile and infertile men. *N. Eng. J. Med.* 345, 1388–1393. doi: 10.1056/NEJMoa003005

Hann, M. C., Lau, P. E., and Tempest, H. G. (2011). Meiotic recombination and male infertility: from basic science to clinical reality? *Asian J. Androl.* 13:212. doi: 10.1038/aja.2011.1

Havasi, V., Rowe, S. M., Kolettis, P. N., Dayangac, D., Sahin, A., Grangeia, A., et al. (2010). Association of cystic fibrosis genetic modifiers with congenital bilateral absence of the vas deferens. *Fertil. Steril.* 94, 2122–2127. doi: 10.1016/j.fertnstert.2009.11.044

He, X., Li, W., Wu, H., Lv, M., Liu, W., Liu, C., et al. (2019). Novel homozygous CFAP69 mutations in humans and mice cause severe asthenoteratospermia with multiple morphological abnormalities of the sperm flagella. *J. Med. Gene.* 56, 96–103. doi: 10.1136/jmedgenet-2018-105486

Huang, D, W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

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, 5–15. doi: 10.1186/s12859-018-2567-6

Jiang, L., Jin, J., Wang, S., Zhang, F., Dai, Y., Shi, L., et al. (2017). CFTR gene mutations and polymorphism are associated with non-obstructive azoospermia: from case-control study. *Gene* 626, 282–289. doi: 10.1016/j.gene.2017.04.044

Joffe, M. (2010). What has happened to human fertility? *Hum. Reprod.* 25, 295–307. doi: 10.1093/humrep/dep390

Kamp, C., Huellen, K., Fernandes, S., Sousa, M., Schlegel, P. N., Mielnik, A., et al. (2001). High deletion frequency of the complete AZFa sequence in men with Sertoli-cell-only syndrome. *Mol. Hum. Reprod.* 10, 987–994. doi: 10.1093/molehr/7.10.987

Kim, S. Y., Kim, H. J., Lee, B. Y., Park, S. Y., Lee, H. S., and Seo, J. T. (2017). Y chromosome microdeletions in infertile men with non-obstructive azoospermia and severe oligozoospermia. *J. Reprod. Infertil.* 18, 307–315.

Krausz, C., Cioppi, F., and Riera-Escamilla, A. (2018). Testing for genetic contributions to infertility: potential clinical impact. *Exp. Rev. Mol. Diagn.* 18, 331–346. doi: 10.1080/14737159.2018.1453358

Krausz, C., Escamilla, A. R., and Chianese, C. (2015). Genetics of male infertility: from research to clinic. *Reproduction* 150, R159–R174. doi: 10.1530/REP-15-0261

Krausz, C., and Riera-Escamilla, A. (2018). Genetics of male infertility. *Nat. Rev. Urol.* 15, 369–384. doi: 10.1038/s41585-018-0003-3

Lafuente, R., García-Blàquez, N., Jacquemin, B., and Checa, M. A. (2016). Outdoor air pollution and sperm quality. *Fertil. Steril* 106, 880–896. doi: 10.1016/j.fertnstert.2016.08.022

Levine, H., Jørgensen, N., Martino-Andrade, A., Mendiola, J., Weksler-Derri, D., Mindlis, I., et al. (2017). Temporal trends in sperm count: a systematic review and meta-regression analysis. *Hum. Reprod. Update* 23, 646–659. doi: 10.1093/humupd/dmx022

Madureira, C., Cunha, M., Sousa, M., Neto, A. P., Pinho, M. J., Viana, P., et al. (2014). Treatment by testicular sperm extraction and intracytoplasmic sperm injection of 65 azoospermic patients with non-mosaic Klinefelter syndrome with birth of 17 healthy children. *Andrology* 2, 623–631. doi: 10.1111/j.2047-2927.2014.00231.x

Maettner, R., Sterzik, K., Isachenko, V., Strehler, E., Rahimi, G., Alabart, J. L., et al. (2014). Quality of human spermatozoa: relationship between high-magnification sperm morphology and DNA integrity. *Andrologia* 46, 547–555. doi: 10.1111/and.12114

Marques, C. J., Carvalho, F., Sousa, M., and Barros, A. (2004). Genomic imprinting in disruptive spermatogenesis. *Lancet* 363, 1700–1702. doi: 10.1016/S0140-6736(04)16256-9

Marques, C. J., Francisco, T., Sousa, S., Carvalho, F., Barros, A., and Sousa, M. (2010). Methylation defects of imprinted genes in human testicular spermatozoa. *Fertil. Steril.* 94, 585–594. doi: 10.1016/j.fertnstert.2009.02.051

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., et al. (2015). The human transcriptome across tissues and individuals. *Science* 348, 660–665. doi: 10.1126/science.aaa0355

Menkveld, R., Wong, W. Y., Lombard, C. J., Wetzels, A. M., Thomas, C. M., Merkus, H. M., et al. (2001). Semen parameters, including WHO and strict criteria morphology, in a fertile and subfertile population: an effort towards standardization of in-vivo thresholds. *Hum. Reprod.* 16, 1165–1171. doi: 10.1093/humrep/16.6.1165

O'Donnell, L. (2014). Mechanisms of spermiogenesis and spermiation and how they are disturbed. *Spermatogenesis* 4:e979623. doi: 10.4161/21565562.2014.979623

Omu, A. E. (2013). Sperm parameters: paradigmatic index of good health and longevity. *Med. Princip. Pract.* 22, 30–42. doi: 10.1159/000354208

Oud, M. S., Volozonoka, L., Smits, R. M., Vissers, L. E., Ramos, L., and Veltman, J. A. (2019). A systematic review and standardized clinical validity assessment of male infertility genes. *Hum. Reprod.* 34, 932–941. doi: 10.1093/humrep/dez022

Pande, M., Amos, C. I., Osterwisch, D. R., Chen, J., Lynch, P. M., Broaddus, R., et al. (2008). Genetic variation in genes for the xenobiotic-metabolizing enzymes CYP1A1, EPHX1, GSTM1, GSTT1, and GSTP1 and susceptibility to colorectal cancer in Lynch syndrome. *Cancer Epidemiol. Prev. Biomark.* 17, 2393–2401. doi: 10.1158/1055-9965.EPI-08-0326

Pereira, R., Oliveira, J., Ferraz, L., Barros, A., Santos, R., and Sousa, M. (2015). Mutation analysis in patients with total sperm immotility. *J. Assis. Reprod. Genet.* 32, 893–902. doi: 10.1007/s10815-015-0474-6

Pereira, R., Oliveira, M. E., Santos, R., Oliveira, E., Barbosa, T., Santos, T., et al. (2019). Characterization of CCDC103 expression profiles: further insights in primary ciliary dyskinesia and in human reproduction. *J. Assis. Reprod. Gene.* 36, 1683–1700. doi: 10.1007/s10815-019-01509-7

Pereira, R., Sá, R., Barros, A., and Sousa, M. (2017). Major regulatory mechanisms involved in sperm motility. *Asian J. Androl.* 19, 5–14. doi: 10.4103/1008-682X.167716

Sá, R., Cunha, M., Rocha, E., Barros, A., and Sousa, M. (2015). Sperm DNA fragmentation is related to sperm morphological staining patterns. *Reprod. Biomed.* 31, 506–515. doi: 10.1016/j.rbmo.2015.06.019

Sousa, M., Cremades, N., Silva, J., Oliveira, C., Ferraz, L., Teixeira da Silva, J., et al. (2002). Predictive value of testicular histology in secretory azoospermic subgroups and clinical outcome after microinjection of fresh and frozen-thawed sperm and spermatids. *Hum. Reprod.* 17, 1800–1810. doi: 10.1093/humrep/17.7.1800

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* 6:e21800. doi: 10.1371/journal.pone.0021800

Tempest, H. G. (2011). Meiotic recombination errors, the origin of sperm aneuploidy and clinical recommendations. *Syst. Biol. Reprod. Med.* 57, 93–101. doi: 10.3109/19396368.2010.504879

Vallières, E., and Elborn, J. S. (2014). Cystic fibrosis gene mutations: evaluation and assessment of disease severity. *Adv. Genom. Gene.* 4, 161–172. doi: 10.2147/AGG.S53768

Wellard, S. R., Schindler, K., and Jordan, P. W. (2020). Aurora B and C kinases regulate chromosome desynapsis and segregation during mouse and human spermatogenesis. *J. Cell Sci.* 133:jcs248831. doi: 10.1242/jcs.248831

Wells, A., Kopp, N., Xu, X., O'Brien, D. R., Yang, W., Nehorai, A., et al. (2015). The anatomical distribution of genetic associations. *Nucl. Acids Res.* 43, 10804–10820. doi: 10.1093/nar/gkv1262

World Health Organization. (2010). *WHO Laboratory Manual for the Examination and Processing of Human Semen.* 5-th edition. URL: https://apps.who.int/iris/bitstream/handle/10665/44261/9789241547789_eng.pdf.

Xiao, F., Lan, A., Lin, Z., Song, J., Zhang, Y., Li, J., et al. (2016). Impact of CAG repeat length in the androgen receptor gene on male infertility–a meta-analysis. *Reprod. Biomed. Online* 33, 39–49. doi: 10.1016/j.rbmo.2016.03.012

Zhang, Y., Zhong, L., Xu, B., Yang, Y., Ban, R., Zhu, J., et al. (2013). SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucl. acids Res.* 41, D1055–D1062. doi: 10.1093/nar/gks1186

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.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 July 2021
doi: 10.3389/fgene.2021.674783

Gene Loss, Pseudogenization in Plastomes of Genus *Allium* (*Amaryllidaceae*), and Putative Selection for Adaptation to Environmental Conditions

Victoria A. Scobeyeva[1,2], Ilya V. Artyushin[3], Anastasiya A. Krinitsina[4], Pavel A. Nikitin[5], Maxim I. Antipin[6], Sergei V. Kuptsov[6], Maxim S. Belenikin[2], Denis O. Omelchenko[7], Maria D. Logacheva[8], Evgenii A. Konorov[9], Andrey E. Samoilov[10] and Anna S. Speranskaya[4,10]*

[1] Department of Evolution, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, [2] Department of Molecular and Biological Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Russia, [3] Department of Vertebrate Zoology, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, [4] Department of Higher Plants, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, [5] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [6] Botanical Garden, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, [7] Laboratory of Plant Genomics, Institute for Information Transmission Problems, Moscow, Russia, [8] Center of Life Sciences, Skolkovo Institute of Science and Technology, Moscow, Russia, [9] Laboratory of Animal Genetics, Vavilov Institute of General Genetics, Russian Academy of Science (RAS), Moscow, Russia, [10] Group of Genomics and Postgenomic Technologies, Central Research Institute of Epidemiology, Moscow, Russia

OPEN ACCESS

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

Reviewed by:
Nikolai Friesen, University of Osnabrück, Germany
Mikhail P. Ponomarenko, Institute of Cytology and Genetics, Russian Academy of Sciences (RAS), Russia
Deng-Feng Xie, Sichuan University, China

***Correspondence:**
Victoria A. Scobeyeva
skobei-khanum@yandex.ru

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

Received: *02 March 2021*
Accepted: *15 June 2021*
Published: *08 July 2021*

Citation:
Scobeyeva VA, Artyushin IV, Krinitsina AA, Nikitin PA, Antipin MI, Kuptsov SV, Belenikin MS, Omelchenko DO, Logacheva MD, Konorov EA, Samoilov AE and Speranskaya AS (2021) Gene Loss, Pseudogenization in Plastomes of Genus Allium (Amaryllidaceae), and Putative Selection for Adaptation to Environmental Conditions. Front. Genet. 12:674783. doi: 10.3389/fgene.2021.674783

Amaryllidaceae is a large family with more than 1,600 species, belonging to 75 genera. The largest genus—*Allium*—is vast, comprising about a thousand species. *Allium* species (as well as other members of the *Amaryllidaceae*) are widespread and diversified, they are adapted to a wide range of habitats from shady forests to open habitats like meadows, steppes, and deserts. The genes present in chloroplast genomes (plastomes) play fundamental roles for the photosynthetic plants. Plastome traits could thus be associated with geophysical abiotic characteristics of habitats. Most chloroplast genes are highly conserved and are used as phylogenetic markers for many families of vascular plants. Nevertheless, some studies revealed signatures of positive selection in chloroplast genes of many plant families including *Amaryllidaceae*. We have sequenced plastomes of the following nine *Allium* (tribe *Allieae* of *Allioideae*) species: *A. zebdanense*, *A. moly*, *A. victorialis*, *A. macleanii*, *A. nutans*, *A. obliquum*, *A. schoenoprasum*, *A. pskemense*, *A. platyspathum*, *A. fistulosum*, *A. semenovii*, and *Nothoscordum bivalve* (tribe *Leucocoryneae* of *Allioideae*). We compared our data with previously published plastomes and provided our interpretation of *Allium* plastome genes' annotations because we found some noteworthy inconsistencies with annotations previously reported. For *Allium* species we estimated the integral evolutionary rate, counted SNPs and indels per nucleotide position as well as compared pseudogenization events in species of three main phylogenetic lines of genus *Allium* to estimate whether they are potentially important for plant physiology or just follow the

phylogenetic pattern. During examination of the 38 species of *Allium* and the 11 of other *Amaryllidaceae* species we found that *rps16*, *rps2*, *infA*, *ccsA* genes have lost their functionality multiple times in different species (regularly evolutionary events), while the pseudogenization of other genes was stochastic events. We found that the "normal" or "pseudo" state of *rps16*, *rps2*, *infA*, *ccsA* genes correlates well with the evolutionary line of genus the species belongs to. The positive selection in various NADH dehydrogenase (*ndh*) genes as well as in *matK*, *accD*, and some others were found. Taking into account known mechanisms of coping with excessive light by cyclic electron transport, we can hypothesize that adaptive evolution in genes, coding subunits of NADH-plastoquinone oxidoreductase could be driven by abiotic factors of alpine habitats, especially by intensive light and UV radiation.

Keywords: *Allium*, plastome, sequence, evolution, pseudogenization

INTRODUCTION

Vascular plants inhabit various ecology niches which may be distinguished by sets of environmental factors (e.g., irradiation level, atmospheric and soil humidity, temperature), all of them may affect photosynthesis. Mountain altitudinal gradient as well as redundantly lit or (as opp.) constantly shadowed habitats seem to be the most powerful "natural geophysical" pressure for evolutionary modification of genes of photosynthetic apparatus. The genes (approximately 120–130 genes) present in chloroplast genomes (plastomes) encode the core proteins of photosynthetic complexes: Photosystem I (*psaA*, *B*, *C*, *I*, and *J*) and Photosystem II (*psbA-F*, *H-N*, *T*, and *Z*), Cytochrome b6f (*petA*, *B*, *D*, *G*, *L*, and *N*)[Hu et al., 2015), NADH dehydrogenase (*ndhA-K*), ATP synthase (*atpA*, *B*, *E*, *F*, *H*, and *I*), the large RUBISCO subunit (*rbcL*), chloroplast ribosomal proteins of large and small subunits (*rpl* and *rps*), polymerase subunits (*rpoA*, *B*, *C1*, and *C2*), ATP-dependent protease (*clpP*), cytochrome c biogenesis (*ccsA*), membrane protein (*cemA*), translation initiation factor I (*infA*), maturase (*matK*) and some proteins of known or unknown functions of (*ycf1_short*, *ycf1_long*, *ycf2*, *ycf3*, *ycf4*), as well as four ribosomal RNAs and various tRNAs (Wicke et al., 2011; Daniell et al., 2016). The core set of plastom's genes is retained from cyanobacteria ancestors; most of them are required for the light reactions of photosynthesis or functions connected with transcription and translation (Sanchez-Puerta et al., 2005; Green, 2011). The evolution of plastome genes could be under pressure of geophysical abiotic factors of plant habitats (Hanaoka et al., 2012; Hu et al., 2015; Zhang et al., 2016; Macadlo et al., 2020). Among geophysical abiotic factors species vertical range limits seem to be more niche-dependent than horizontal (Hargreaves et al., 2014) and vertical limits are often species niche limits (Lee-Yaw et al., 2016).

Species adaptations to diverse environments are accompanied by mutations under positive selection which may be confirmed using analysis of Kn/Ks ratio. Numerous genes were proved to be under positive selection in taxons which are myco-heterotrophic plants, and also in plants that are partly or entirely non-photosynthetic (Zeng et al., 2017; Dong et al., 2018). The sets of genes which are under positive selection in plants with adaptation to various ecological niches seem to be incoherent. Positive selection in some genes, e.g., *rbcL,* is widespread in most lineages of land plants (Yao et al., 2019). Some authors propose they are associated with adaptation to various ecological niches, including dry or wet habitats (Kapralov and Filatov, 2006), high altitudes in mountain regions (Hu et al., 2015). *matK* is reported to be a highly variable sequence which was often used for phylogeny analysis of many plant taxons (Hilu et al., 2003; Tamura et al., 2004). Different sets of genes are found under positive selection in species of various taxonomic groups. For example, in *Oryza* species (*Poaceae*, *Monocotyledones*) which were adapted to either shady or sunny environments, a set of positively selected genes (besides *rbcL* and *matK*) was found: *accD*, *ndhD*, *ndhF*, *ndhH*, *psaA*, *psbB*, *psbD*, *psbH*, *rpl16*, *rpoA*, *rpoC2*, and *ycf68* (Gao et al., 2019). A completely different set of genes (*ycf1*, *ycf2*, *rps14*, *rps15*, and *rps16)* was found to be under selection in *Cardamine* and *Nasturtium* species (*Brassicaceae*) (Yan et al., 2019). Another analysis of adaptive evolution in *Brassicaceae* plastomes in general and in the *Cardamine* genus in particular resulted in detection of signatures of positive selection in the following genes (besides *rbcL* and *matK*): *ycf1*, *rpoC2*, *rpl14*, *petD*, *ndhF*, *ccsA*, *accD*, and *rpl20* at a significant level (Hu et al., 2015). Presumably, detected signals of positive selection in genes may reflect specific adaptations of species to a particular habitat. For example, in *Brassicaceae* it could possibly be a consequence of adaptation to high altitude environments (Hu et al., 2015).

Furthermore, extensive structural changes, such as large inversions, deletions, loss of functionality (pseudogenization or complete loss) of genes are regularly observed in plastomes of certain plant species or groups of species (various taxa of higher rank) (Haberle et al., 2008; Guisinger et al., 2011; Schwarz et al., 2015; Hsu et al., 2016; Fonseca and Lohmann, 2017).For example, the loss of all *ndh* (NADH dehydrogenase subunit complex) genes is apparently a result of convergent evolution during the land adaptation of photosynthesis in neighboring clades of different taxa. The *ndh* genes loss in plastomes is usually linked to heterotrophic type of nutrition, like in *Orchidaceae* (Lin et al., 2015; Lin et al., 2017) or *Lentibulariaceae* (Silva et al., 2016). However, the *NDH* complex may also be lost in photoautotrophic plant species, as it was discovered for a wide spectrum of

angiosperms and gymnosperms (de Pamphilis and Palmer, 1990; Haberhausen and Zetsche, 1994; Wakasugi et al., 1994; Martín and Sabater, 2010; Blazier et al., 2011; Omelchenko et al., 2020). Presumably, the loss of *ndh* genes is linked with inability to tolerate light intensity stress (Omelchenko et al., 2020). But for most plastome genes implications of their loss on the physiological and ecological functionality of plants remain inexplicable and are interpreted as random evolution events or just stated as fact. For example, *rpl23* has been found to be a pseudogene in monocots (Ogihara et al., 1991; Morton and Clegg, 1993; Omelchenko et al., 2020) and in all species of *Caryophyllales* (eudicots) (Raman and Park, 2015), while a wide range of species belonging to other clades of eudicots (e.g., rosids, asterids) have no traces of pseudogenization of *rpl23* (Raman and Park, 2015).

Amaryllidaceae is a large family with more than 1,600 species, divided into three subfamilies: *Agapanthoideae*, *Amaryllidoideae*, and *Allioideae*. The genus *Allium* (*Allioideae*) is vast, being one of the largest monocotyledonous genera and comprising more than 800 species (Fritsch et al., 2010; Fritsch, 2012; Shah, 2014; Maragheh et al., 2019) or about 973 species according to WCSP database (Govaerts and Fritsch, 2021). *Allium* species form 15 subgenera that are grouped into three intrageneric evolutionary lineages according to molecular data analysis of plastid and nuclear DNA barcode sequences (Friesen et al., 2006; Li et al., 2010; Wheeler et al., 2013). Besides the *Allium* genus (comprising monogeneric tribe *Allieae*), *Allioideae* has 17 more genera belonging to another three tribes (*Gilliesieae*, *Leucocoryneae*, and *Tulbaghieae*) (Chase et al., 2009; Sassone et al., 2014). *Allium* species (as well as other members of the *Amaryllidaceae*) are widespread and diversified, they are adapted to a wide range of habitats from shady forests to open habitats like meadows, steppes and deserts and highlands and are therefore well suited as study objects in studies of plants' adaptations to various sun irradiation levels and other geophysical factors (Vvedensky, 1935; Wendelbo, 1971). Complete plastome sequences were first sequenced and arranged for the most economically important species of *Allium*: onion, *A. cepa* (von Kohn et al., 2013), garlic, *A. sativum* (Filyushin et al., 2016), edible species *A. ursinum* and *A. paradoxum* (Omelchenko et al., 2020). Another wide spectrum of wild and cultivated *Allium* species plastomes were sequenced and partially analyzed (Filyushin et al., 2018, 2019; Xie et al., 2019, 2020; Liu et al., 2020; Namgung et al., 2021).

According to data available, plastomes of most plant species contain the same number of tRNA genes (30 in total, of them 9 are represented by two copies in IR) and rRNA genes (eight genes, all four are represented by two copies in IR), while the rest 79 genes encode proteins (Wicke et al., 2011; Wambugu et al., 2015). Investigated *Allium* species had similar number, arrangement and orientation of genes (Filyushin et al., 2016, 2018; Huo et al., 2019; Yusupov et al., 2020). Only *A. paradoxum* demonstrated notable alterations, such as large 4,825 bp long local inversion in the SSC region and elimination or pseudogenization of the whole *ndh* gene family, as well as large number of other genes: *rps16*, *rps2*, *rpl22*, *petD*, *infA*, *rpl23*, *rps3* (Omelchenko et al., 2020). Recently positive selection pressure was calculated in subgenus *Anguinum* of *Allium* and three genes were found with Kn/Ks >1 (*accD*, *rps14*, *rpl33*) (Jin et al., 2019). Soon after this an analysis of average Kn/Ks in a total of 39 complete chloroplast genomes of *Allium* was performed and quite unexpectedly it revealed an absolutely different set of positively selected genes (with Kn/Ks > 1): *psbC*, *rps11*, and *psaI* (Xie et al., 2020). These inconsistent data prompted us to choose a simple and easy to analyze ecological trait—the highest tolerable altitude of the species. High altitude habitats have high levels of solar radiation, often low humidity and other special meteorological conditions, influencing the whole plant physiology (Gale, 2004). We took the highest tolerable altitude of the *Allium* species as a selective factor and performed an evolutionary analysis of plastome genes in species with contrasting altitude limits (see **Supplementary Material 2** for species list). In this manuscript we only report results that relate to protein-coding genes of examined plastomes. We report *de novo* sequencing and assembling of complete cp-genomes of wild and cultured *Allium* species as well as outgroup species belonging to the same subfamily of *Allioideae* in *Amaryllidaceae—Nothoscordum bivalve* (L.) Britton. We have identified rapidly evolving plastome regions and genes under positive selection, and detected plastome traits that differentiate between species of three evolutionary lines of *Allium* genus adapted to certain contrasting habitat types: high-altitude vs. lowland habitats.

MATERIALS AND METHODS

Sampling, DNA Extraction, cpDNA Isolation, and Sequencing

Plastomes were newly sequenced and assembled for the following species obtained alive from the outdoor collection of the Moscow State University Botanical Garden: *A. fistulosum* L., *A. macleanii* Baker, *A. nutans* L., *A. obliquum* L., *A. pskemense* B. Fedtsch., *A. schoenoprasum* L., *A. victorialis* L., *A. zebdanense* Boiss. & Noë. The specimens of *A. semenovii* Regelas well as *A. platyspathum* Schrenk were collected in the wild in 2016. The specimens of the abovementioned species from MSU Botanical Garden living collections and from wild natural habitats were deposited in the Moscow State University Herbarium, extracted DNA from *Allium* plants were deposited in DNA collection in MSU Biology Department. The plastomes of *A. moly* L and *Nothoscordum bivalve* (L.) Britton were sequenced using dry plant material obtained from Osnabrück Botanical Garden (Germany), kindly provided by Nikolai Friesen. All additional details, including taxonomic position of species used in the study, the source of specimens, GenBank/ENA accession numbers of plastome sequences, information about ecology and geographical location of wild collected specimens, accession number of vouchers are provided in **Supplementary Table 2**.

For cpDNA extraction from samples obtained from living specimens (Lomonosov MSU botanical garden collections),fresh leaves were cut and stored in the dark in a refrigerator at +4°C for 10 days. Chloroplasts were isolated from about 2 g (fresh weight) of leaves using protocol based on Shi et al. (2012) and Vieira et al. (2014) and described in detail in Logacheva et al. (2017). Then cpDNA was extracted using standard CTAB protocol (Doyle and Doyle, 1987). Quality of cpDNA was evaluated visually by gel-electrophoresis.

Another approach was used for herbarium samples of *A. semenovii*, *A. platyspathum*, *A. moly*, and *N. bivalve*. Fragments of dry leaves were used as material for total DNA extraction according to Doyle and Doyle (1987) with an additional stage of purification according to Krinitsina et al. (2015) because of high amount of impurities (presumably phenolic compounds). Quality of total DNA was evaluated using spectrophotometry (Nanophotometr-N-60, Implen). Quantification of nucleic acids was done using fluorimetry (Qubit 3.0, Thermo Fisher Scientific).

Library Preparation, Sequencing, and Data Assembling

The libraries with insert sizes of 400–800 bp were constructed and then sequenced using high-throughput sequencing platform Illumina MiSeq (PE 2 × 250 bp or 2 × 300 bp). Paired-end libraries for each DNA sample were constructed at least twice using two different library preparation strategies to minimize protocol-associated biases and maximize assembly efficiency. First strategy included physical fragmentation by Covaris 220 followed by protocol with adapter ligation, using NEBNext® DNA Library Prep Master Mix Set for Illumina (E6040, NEB reagents) with single indexed primers from NEBNext® Multiplex Oligos for Illumina Kits (Index Primers Set 1–4), used according to the manufacturer's recommendation. Another approach was DNA library preparation using a transposase-based method (Nextera), developed by Illumina. After tagmentation the libraries were amplified using NEB Q5® High-Fidelity DNA Polymerase (up to 12 cycles of amplification) and Nextera-compatible dual indexed primers.

The chloroplast genome assembly protocol included (1) quality trimming with Trimmomatic (Bolger et al., 2014), (2) filtering of reads using known chloroplast genome sequences of *A. cepa* (NC024813) and *A. sativum* (NC031829) by Bowtie2 mapper (Langmead and Salzberg, 2012), (3) producing of two contig sets for both filtered and non-filtered reads using *de novo* assemblers Velvet (Zerbino and Birney, 2008) and Spades (Bankevich et al., 2012). Assembled contigs were selected for the next assembly if they showed similarity to published *Allium* plastomes. The final *de novo* assembly was checked and fixed where necessary by PE reads mapping to the assembly in order to check for potential assembly artifacts using Bowtie2, VarScan (v.2.3.7) and SAMtools/BCFtools software packages (Li et al., 2009; Langmead and Salzberg, 2012; Koboldt et al., 2012). The coverage was as follows: *A. semenovii* ∼30×, *A. macleanii* ∼35×, *A. fistulosum* > 200×, *A. platyspathum* ∼35×, *A. nutans* > 300×, *A. obliquum* > 200×, *A. pskemense* ∼39×, *A. schoenoprasum* L. > 250×, *A. victorialis* ∼80×, *A. zebdanense* > 150×.

Plastomes of *A. macleanii*, *A. nutans*, *A. obliquum*, *A. platyspathum*, *A. pskemense*, *A. schoenoprasum*, *A. victorialis*., *A. zebdanense* were completely assembled using only the high-throughput sequencing approach. Some unassembled regions of *A. semenovii* plastome were sequenced by Sanger method. Primers for amplification were: **Alsem 1 for:** GTC CTC GGT AAC GAG ACA TAA; **Alsem 1 rev:** ACG TAG TCA ACT CCA TTC GT; **Alsem 2 for:** GTG CCC AAA ATG GTG TCA AT; **Alsem 2 rev:** ATC CAT GGT TTA TTC CTT ATC TCT; **Alsem 3 for:** GTA TGC CGT CTT CTG CTT G; **Alsem 3 rev:** AAG GGT TCT TTT AAA CTC TTT TGT T; **Alsem 4 for:** TGT TGG ACA ATA CTC GAC AC; **Alsem 4 rev:** GAC CAT AGA GGA GCC GTA TG; **Alsem 5 for:** GAG TGG AGC TAT ACC CAA TAG ATA; **Alsem 5 rev:** TAA GGT TAT CTC CCG CCA AT.

Plastome Annotation

The Dual Organellar GenoMe Annotator (DOGMA) (Wyman et al., 2004) and GeSeq (Tillich et al., 2017) programs were used for preliminary gene annotation. From this initial annotation, putative start codons, stop codons, and intron positions were determined. Then putative start and stop codons together with intron positions were manually corrected based on comparisons with homologous genes of cp genomes of *Lycoris squamigera* (NC_040164.1), *Narcissus poeticus* voucher WSY: WSY0108940 (NC_039825.1), *Lycoris radiata* (NC_045077.1), *Allium cepa* (NC_024813.1), *Agapanthus coddii* voucher K:20081397 (NC_035971.1), *Allium paradoxum* (NC_039661), *Allium herderianum* (NC_042156.1), and *Allium victorialis* (NC_037240.1). All identified tRNAs were further verified by tRNAscan-SE 1.21 (Chan and Lowe, 2019). Complete nucleotide sequences of plastomes sequenced in this study were deposited in the GenBank database under accession numbers listed in **Supplementary Table 1**. After this results of annotations were manually checked to correct errors that appear as a result of the work of algorithms implemented in the annotation software.

Plastome Features Comparison

To investigate the sets of protein coding genes that reflect the evolution of plastomes in the three main clades of the genus *Allium* (Friesen et al., 2006), we used all complete and partially assembled sequences and complete sequences of several other *Allium* species plastomes obtained from GenBank. All annotated sequences previously published in GenBank were re-annotated using in-house scripts and *A. cepa* (NC_024813) as basic reference sequences. *A. praemixtum* (NC_044412) and *N. poeticus* (NC_039825) were used for annotation of genes pseudogenized or absent in *A. cepa*. Then we excluded most cultivated species, besides taken as reference, to avoid possible bias due to unknown processes accompanying artificial selection of cultivated forms. The sequences of the protein coding genes were extracted, than extraction and translation were manually verified to correct possible errors that appear as a result of the work of algorithms implemented in the annotation software. Although a fairly large number of complete *Allium* plastome sequences are available online, taxonomic attribution of many species looks uncertain. We used the species that we sequenced and assembled during this project and added more plastomes from GenBank, choosing species that were important for the aims of our research, namely representatives of the first and the second evolutionary line (Friesen et al., 2006) with contrasting highest tolerated altitudes: *A. paradoxum* (M. Bieb.) G. Don (MH053150), *A. ursinum* L. (MH157875.1), *A. prattii* L. (NC_037432.1), two important agricultural species *A. cepa* L. (NC_024813), *A. sativum* L. (NC_031829), and two species with known intraspecific variability [*A. obliquum* (MG670111) and *A. victorialis* (MF687749)]. For the complete list of analyzed species see **Supplementary Table 2**.

Detecting of Pseudogenes, GC-Content, Positive Selection, and Evolution Rate Analysis

Primary alignment was prepared with MAFFT v7.471 (Kazutaka and Standley, 2013). To identify pseudogenes, we aligned nucleotide sequences with MACSE and marked the first position of either frameshift, stop-codon or deletion spanning to the sequence end. Alignments were visualized with JalView v2.11.1.3 (Waterhouse et al., 2009). Indels-containing regions were inspected by eye for possible alignment errors. No additional alignments filtering was applied. Pseudogene heatmap was constructed with pandas (Reback et al., 2021), numpy (Harris et al., 2020), seaborn (Waskom et al., 2020), and matplotlib (Hunter, 2007) Python3 packages (Cock et al., 2009; Van Rossum and Drake, 2009).

We calculated GC content and proportion of gaps by window 100 bp in length using custom python script. For every window substitution count was estimated using a fixed tree topology (ParsimonyScorer class in Biopython v 1.78 (Talevich et al., 2012). Tree topology was the same as for HyPhy methods (explained below). Plots were constructed with ggplot2 R package (Wickham, 2016; R Core Team, 2018).

For selection and evolutionary rate analysis we used the concatenated sequences of 78 genes (see **Supplementary Table 3**). All individual gene sequences recognized as pseudogenes were replaced with gaps. Phylogenetic analysis was conducted with IQ-TREE v 2.0.6 (Chernomor et al., 2016; Minh et al., 2020). Evolutionary models and partitioning schemes were optimized using built-in methods (Kalyaanamoorthy et al., 2017) starting from individual partitions for every gene. The final scheme had seven partitions. Obtained tree was used for assessment of per-gene evolutionary rates and selective pressure detection. Tree was visualized with FigTree v1.4.4 (Rambaut and Drummond, 2012). We applied FUBAR v2.2 (Murrell et al., 2013), aBSREL 2.1 (Smith et al., 2015), and MEME 2.1.1 (Murrell et al., 2012) methods implemented in HyPhy framework v2.5.2 MP (Kosakovsky Pond et al., 2005). The latter was used to obtain Kn/Ks ratios for genes.

We have carried out analysis and identification of genes under positive or diversifying selection in *Allium* genus in its adaptive evolution to different solar irradiation and other abiotic conditions of its wide range in highlands and various lowlands in the temperate zone. A total of 49 species was analyzed, of which data for 13 species was obtained by our group. Two species sequenced by our group already had been stored in databases earlier (*A. obliquum* and *A. victorialis*), so we took 51 sequences in analysis total. We have chosen 8 species from the first evolutionary line, 7 from the second and 23 from the third. Cultivated species *A. cepa* was included in the analysis as a reference, *A. sativum*, *A. fistulosum*, *A. tuberosum*, *A. chinense* were excluded from the analysis. As far as species vertical limits are often species niche-limits (Hargreaves et al., 2014), we took the highest tolerable altitude of the species listed, using the data available from literature sources (see **Supplementary Table 2**) as well as the information presented on the herbarium labels for the corresponding species in the herbarium of the Moscow State University[1]. Our analysis consisted of a classical approach with Kn/Ks calculation, aBSREL (Smith et al., 2015), MEME (Murrell et al., 2012), and FUBAR (Murrell et al., 2013). We considered evidence of positive selection as sufficiently reliable only for those cases where it was confirmed by several methods.

To compare ratio of genes under selection among different gene ontologies we performed Fisher's exact test for multiple samples. All genes, used for selection seeking, were divided into seven ontologies—ATP synthesis, Cytochrome complex and electron-transport chain, *ndh* genes, photosynthesis, translation, transcription and other. For all the ontologies were calculated ratio of sites under selection to sites without selection and compared with Fisher's exact test for multiple samples with R package (Holm–Bonferroni method was used to counteract the problem of multiple comparison).

RESULTS

We have sequenced, assembled and annotated plastomes of *A. zebdanense* (MZ019480), *A. moly* (MZ019477), *A. victorialis* (NC_037240.1), *A. macleanii* (LT699703.1), *A. nutans* (LT799837.1), *A. obliquum* (MZ019478), *A. schoenoprasum* (LT699700.1), *A. platyspathum* (LT673892.1), *A. semenovii* (MZ019479), *A. pskemense* (MZ147623), *A. fistulosum* (LT674586.1). In addition, we have obtained a complete nucleotide sequence of the plastome of *Nothoscordum bivalve*. Together with genus *Allium* it belongs to the *Allioideae* subfamily, but to a different tribe *Leucocoryneae*. The plastome sequence of *Nothoscordum bivalve* was submitted as (MZ019481). The main parameters, i.e., total/LSC/SCC/IRs lengths and complete comparative annotations for most of assembled plastomes are represented in the (**Supplementary Material 1**). To expand the sampling, we took the plastome sequences of other wild *Allium* species published in GenBank. In total 41 plastome sequences of wild *Allioideae* species were analyzed. In addition our analysis also included representatives of two other *Amaryllidaceae* subfamilies, *Agapanthoideae* and *Amaryllideae*, data on which were found at Genbank. For the complete list of the studied species and descriptions of general ecology features of their natural habitats (see **Supplementary Material 2**).

To identify and correct possible annotation errors that arise as a result of the work of algorithms implemented in annotation software, we performed automatical reannotation of all previously published plastome sequences that were selected for comparative analysis in this work, using GeSeq. In addition to this we had determined start and stop codons of all protein coding genes using in-house scripts. The results of both procedures were then manually compared and finalized. This step allowed us to find some noteworthy inconsistencies with previous reports, in particular, regarding pseudogenization status of *rps2*, *rps16*, and *ndhD* reported by Xie et al. (2019). The authors claim that *rps2* was lost in all *Allium* species (*Allioideae*), but we have found that pseudogenization occurred in only about a third of studied *Allium* species, while in the rest two thirds these genes proved

[1]https://plant.depo.msu.ru/

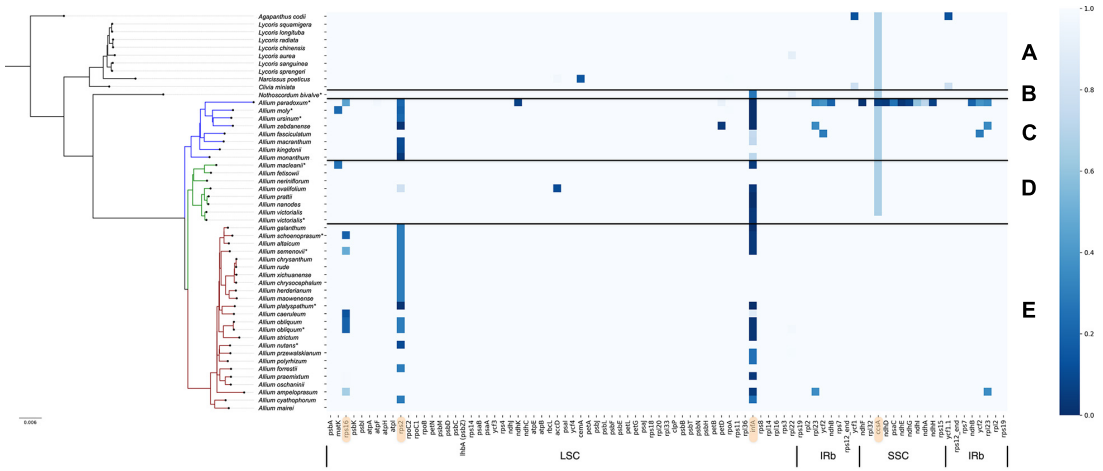


FIGURE 1 | Pseudogenization or loss of protein-coding genes heatmap in the plastome of *Allioideae* (including all three evolution lines of *Allium*). Taxonomic position of species is denoted at the left side. Species sequenced by our group are marked with (*). **(A)** Shows *Agapanthoideae* and *Amaryllidoideae* subfamily species; **(B)** shows *Nothoscordum bivalve* as an outgroup for *Allium* genus; **(C)** shows the first evolutionary line *Allium* species; **(D)**—the second and **(E)**—the third. Genes are sorted in natural order in plastid. *LSC*, large single-copied region; *SSC*, small single-copied region; *IRa* and *IRb*, inverted repeats A and B. The darkest blue marks completely deleted genes; skyblue of varying shades marks pseudogenized genes (depending on the length of the preserved gene fragment); light blue marks true genes (translated without errors). Genes described below in detail are marked with peach color (*rps16*, *rps2*, *infA*, *ccsA*).

to be in their true state, coding proteins without stop-codons (see **Figure 1**). Manual correction of automatic annotations has also influenced our conclusions regarding functionality of various genes like *rps16* in *A. platyspathum* and *A. nutans*: we found that these genes in these species are true. The *ndhD* gene also proved to be in its true state in all investigated *Allium* species (excluding *A. paradoxum*), in contradiction to conclusions made by Xie et al. (2019).

Interrelation of Plant Preferences to Habitats/Phylogenetics and Deletion/Pseudogenization of Genes in Plastomes

The set of plastomes obtained were used for analysis of deleted/pseudogenized genes patterns and their association with evolutionary lineages and with environmental adaptation of species. Plastomes of *Allium* species as well as other known *Amaryllidaceae* are generally similar in composition of their functional gene sets. The main differences were found in the functionality of the following genes: *ccsA* (responsible for heme attachment to cytochromes c), *rps16* and *rps2* (proteins of the small ribosomal subunit), *infA* (translation initiation factor I), see **Figure 1**. Pseudogenization of some other genes is presumably a sporadic event in the *Allium* genus, sometimes happening in all evolutionary lines: *petD* (encodes a subunit of the cytochrome b6/f complex), *ycf1* (translocon on the inner plastid membrane), *ycf2* (encodes a subunit of the 2-MD heteromeric AAA-ATPase complex which associated with the TIC complex (Kikuchi et al., 2018) and some other genes can undergo occasional defunctionalization.

Undoubtedly the sequence features of the *rps2*, *infA*, *rps16*, *ccsA* genes as well as their "normal" or "pseudo" state correlate with the genus evolutionary line to which species belong to.

The *rps16* Gene Encoding Plastid 30S Ribosomal Protein S16

We found that pseudogenization affects *rps16* genes only in species of the third evolutionary line, namely *A. schoenoprasum*, *A. obliquum*, *A. ampeloprasum*, *A. caeruleum*, and *A. platyspathum* (see **Figure 2**). In plastomes of first and second evolutionary lines of *Allium* (with only one exception, *A. paradoxum*) as well as in all non-*Allium Amaryllidaceae*, *rps16* gene is translated without errors.

The *rps2* Gene Encoding Plastid 30S Ribosomal Protein S2

Defunctionalization of the *rps2* gene (encoding plastid 30S ribosomal protein S2) was found to be characteristic of most *Allium* species (see **Figure 3**). There is a considerable amount of single nucleotide substitutions along with several indels in the sequence of the *rps2* gene which leads to many non-synonymous AA substitutions and premature stop codons. Interestingly, at first glance, the species of the second evolutionary line seem to be less affected by this process, *rps2* being pseudogenized only in one species, *A. ovalifolium*.

The *infA* Gene Encoding Translation Initiation Factor I

The differences in the nucleotide sequences of the *infA* genes (encoding translation initiation factor IF-1) are undoubtedly correlated with the phylogeny of the *Amaryllidaceae* family. While *Agapanthoideae* and *Amaryllidoideae* species contain genes translated without errors, most *Allioideae*, including *N. bivalve* and most *Allium* species of the first, second and third evolutionary lines, have pseudogenized *infA* (see **Figure 4**). In some *Allium* species the *infA* genes are completely deleted (for example, in *A. macleanii*). In many species they contain stop codons at different distances from the start of translation (**Figure 4**). Pseudogenized *infA* genes, containing a stop codon

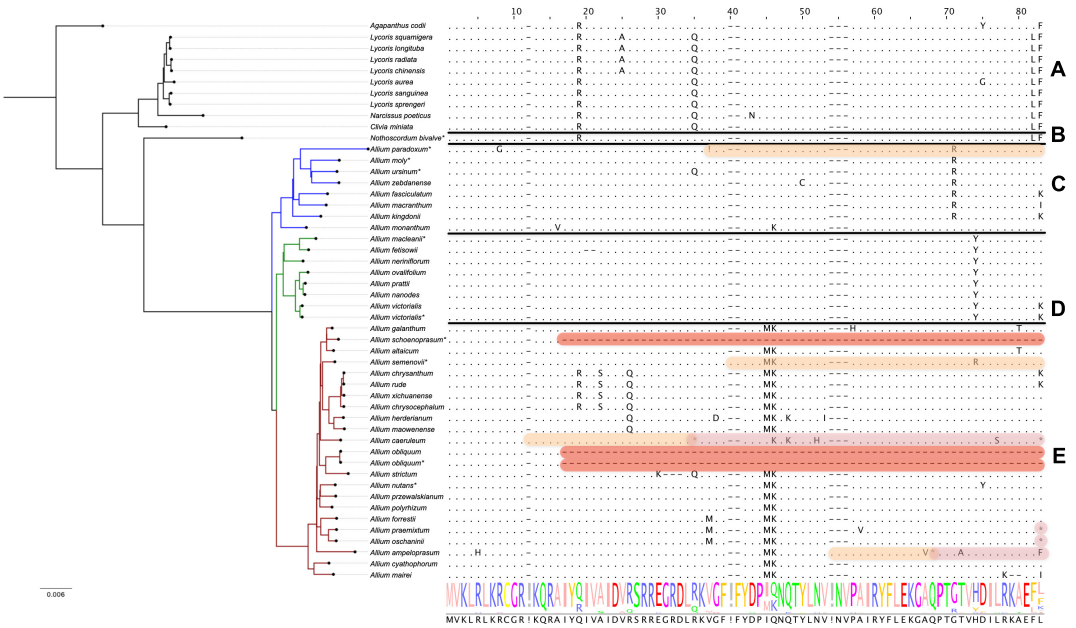


FIGURE 2 | Protein sequence alignment of the plastid 30S ribosomal protein S16 encoded by *rps16* genes. Àll species are sorted by phylogenetic tree: the first evolutionary line of *Allium* is blue, the second is green, the third is red and the remaining species are black. Species sequenced by our group are marked with *. **(A)** Shows *Agapanthoideae* and *Amaryllidoideae* subfamily species; **(B)** shows *Nothoscordum bivalve* as an outgroup for *Allium* genus; **(C)** shows the first evolutionary line *Allium* species; **(D)**—the second and **(E)**—the third. Non-conserved AA positions are highlighted as letters and conserved as dots. The whole AA sequence is shown as a normalized alignment logo below. Possible frameshifts are shown as! and stop codons as *. Damaged sequences are highlighted in peach color in case of frameshift, in purple in case of stop codon and in red in case of total AA loss. Colored sequences may be assumed to be a pseudogene. The gene is absent in *A. platyspathum*.

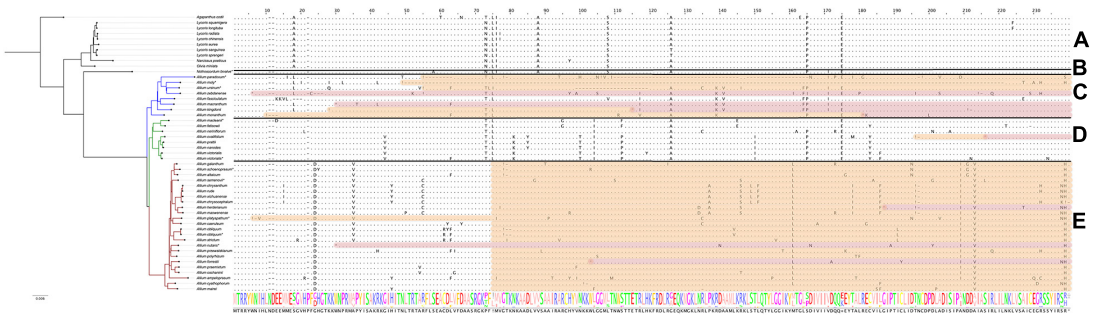


FIGURE 3 | Protein sequence alignment of the plastid 30S ribosomal protein S2 encoded by *rps2* genes. Àll species are sorted by phylogenetic tree: the first evolutionary line of *Allium* is blue, the second is green, the third is red and the remaining species are black. Species sequenced by our group are marked with *. **(A)** Shows *Agapanthoideae* and *Amaryllidoideae* subfamily species; **(B)** shows *Nothoscordum bivalve* as an outgroup for *Allium* genus; **(C)** shows the first evolutionary line *Allium* species; **(D)**—the second and **(E)**—the third. Non-conserved AA positions are highlighted as letters and conserved as dots. The whole AA sequence is shown as a normalized alignment logo below. Possible frameshifts are shown as ! and stop codons as *. Damaged sequences are highlighted in peach color in case of frameshift and in purple in case of stop codon. Colored sequences may be assumed to be a pseudogene.

at a short distance from translation start point, are found more often in plastomes of the species belonging to the first and second evolutionary lines. The observed patterns indicate that *infA* genes were pseudogenized multiple times after the separation of the *Allioideae* from other groups within *Amaryllidaceae*.

Analysis of amino acid sequences that lack stop codons and frameshifts in the middle of the sequence has shown that in species of the second evolutionary line there occurs a radical amino acid substitution in Y34S position, while all *Allium* species have such substitutions at I46T. Amino acids of these positions participate in forming rRNA binding sites (nucleotide binding) (Hiratsuka et al., 1989).

The *ccsA* Gene Encoding Cytochrome c Biogenesis Protein CcsA

The *ccsA* gene is undergoing pseudogenization within all non-*Allium* species as well as in all species of first and second evolution lines (see **Figure 5**). The plastome of *A. victorialis* (LT699702.1) sequenced in this work was the only exception to this rule, containing normally translated *ccsA*. Interestingly,

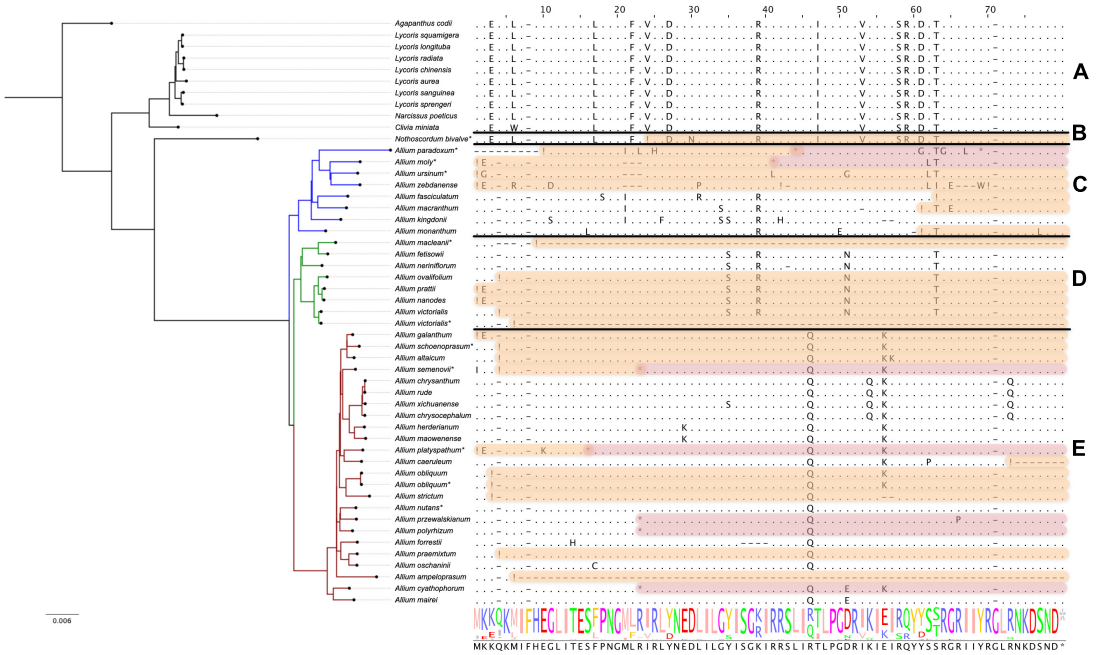


FIGURE 4 | Protein sequence alignment of the translation initiation factor IF-1 encoded by *infA* genes. Àll species are sorted by phylogenetic tree: the first evolutionary line of *Allium* is blue, the second is green, the third is red and the remaining species are black. Species sequenced by our group are marked with *. **(A)** Shows *Agapanthoideae* and *Amaryllidoideae* subfamily species; **(B)** shows *Nothoscordum bivalve* as an outgroup for *Allium* genus; **(C)** shows the first evolutionary line *Allium* species; **(D)**—the second and **(E)**—the third. Non-conserved AA positions are highlighted as letters and conserved as dots. The whole AA sequence is shown as a normalized alignment logo below. Possible frameshifts are shown as ! and stop codons as *. Damaged sequences are highlighted in peach color in case of frameshift and in purple in case of stop codon. Colored sequences may be assumed to be a pseudogene.

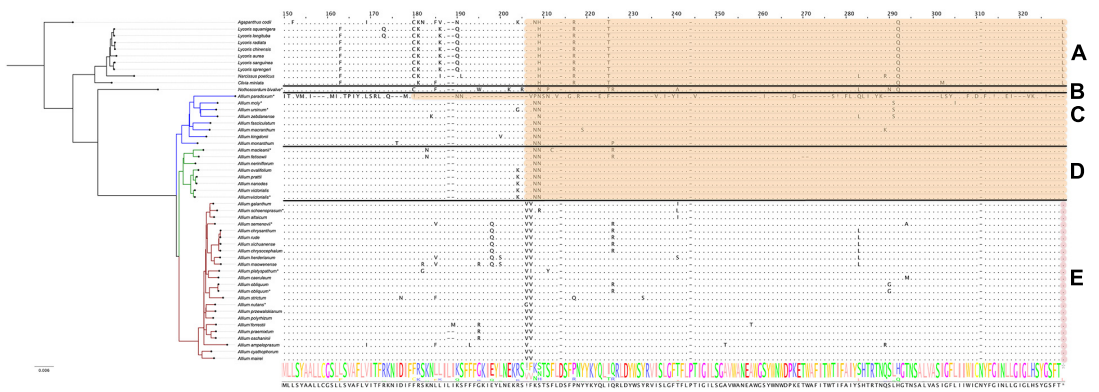


FIGURE 5 | Protein sequence alignment of the cytochrome c biogenesis protein encoded by *ccsA* genes. Àll species are sorted by phylogenetic tree: the first evolutionary line of *Allium* is blue, the second is green, the third is red and the remaining species are black. Species sequenced by our group are marked with *. **(A)** Shows *Agapanthoideae* and *Amaryllidoideae* subfamily species; **(B)** shows *Nothoscordum bivalve* as an outgroup for *Allium* genus; **(C)** shows the first evolutionary line *Allium* species; **(D)**—the second and **(E)**—the third. Alignment here starts from the 150th AA position for better visualization. Non-conserved AA positions are highlighted as letters and conserved as dots. The whole AA sequence is shown as a normalized alignment logo below. Possible frameshifts are shown as ! and stop codons as *. Damaged sequences are highlighted in peach color in case of frameshift and in purple in case of stop codon. Colored sequences may be assumed to be a pseudogene.

the other specimen of *A. victorialis* (NC_037240.1) sequenced by Lee et al. (2017) has *ccsA* pseudogenized like other species of the second evolutionary line. We suppose that the plastome of *A. victorialis* (LT699702.1) could contain some inexplicable mistake in the *ccsA* gene and could be ignored when patterns of pseudogenization are discussed. Thus, the *ccsA* gene status ("normal" or "pseudo") reflects the phylogenetic evolution of the *Amaryllidaceae*: *Agapanthoideae* and *Amaryllidoideae* species as well as basic clades of *Allioideae* form *ccsA* pseudogenes.

Despite the fact that their products are obviously necessary for the normal plant system functioning, the reasons why the *infA*, *rps16*, *rps2*, and *ccsA* genes had lost their functionality remains unclear. It is possible that copies of these genes can be found in nuclear genomes. We checked the sequences of the assembled whole nuclear genome of *A. sativum* which is the only one available in NCBI presently (assembling accession number: GCA_014155895.1) for the presence of the sequences *infA*, *rps16*, *rps2* but gained no positive results. Perhaps these genes can be found in nuclear genomes of other *Allium* species.

Evolution Rate and Plastome Gene Selective Pressure Analysis

The evolution rate estimation is widely used in phylogenetic and evolutionary studies. Usually pseudogenes and genes with loose selective constraints have the highest rates, followed by positively selected genes. Methods for detecting positive or diversifying selection events are sensitive to evolutionary models, so we treated them as putatively selected only sites, discovered by both MEME and FUBAR methods of analysis (see **Supplementary Table 3**). The highest rate of evolution, 3,16, was detected in gene *rpl32*. No sites under positive selection are found by MEME and FUBAR in sequence of 59 a.a. length in protein 60S ribosome 32, coded by *rpl32*. We can therefore suppose a neutral model of evolution in most lineages, with selective constraints loosen in some species. Amino acid sequence of *Allium paradoxum rpl32* protein has relatively low homology with sequences of the other *Allium* species and high number of gaps in alignment. In *A. macranthum* and *A. fetisovii* amino acid sequence of 60S ribosome 32 also differs from consensus alignment, so we can either suppose that this protein plays a role in adaptation or rather has lost its selective value in these species.

The next gene in the high evolutionary rate list is *ycf1*, encoding the long protein of 1,761 a.a. with partly known functions and a story of pseudogenization in various plant families. MEME gives 51 sites under positive selection, while only 4 of them coincide with FUBAR list. Selection was detected by AbsRel method in two lineages—*A. mairei* and *A. oschaninii*. We can suggest predominantly neutral evolution of *ycf1* in most lineages, but in some cases we can suppose selective evolution as well. In *A. macranthum*we can see long compensated frameshift, it is an evident trace of stabilizing selection.

The third gene in this list is *psbZ*, encoding a small protein of Photosystem II reaction center. MEME revealed only one site under selection and it does not match with FUBAR data. But the Kn/Ks ratio in *psbZ* is 1,26, one of the highest in all the examined genes. Natural selection obviously affects the sequence, but we cannot detect the particular site of its action. So, we would rather suggest neutral evolution after relaxation of selective constraints (Akashi et al., 2012). The last gene with evo rate >2 is the ndhF gene, coding NADH-plastoquinone_oxidoreductase_subunit_5. We detected it by MEME selection in 15 sites of 734 examined, 4 of them coinciding with FUBAR data. The aBSREL detected positive selection in the ndhF gene in three lineages—*A. macranthum*, *A. forrestii*, and *A. neriniflorum*. Kn/Ks ratio in the *ndhF* gene is 0,28, so we should stay away from the conclusion that selective mode of evolution is significant in all examined lineages, but we may say that in some of them natural selection can play a certain role.

matK has an evolutionary rate of 1,98, which is insignificantly lower than 2. MEME has found 13 sites from 521 under selection, and only 3 of them are the same as those found by FUBAR. aBSREL found only one lineage under selection—*A. paradoxum*, and in two species *(A. moly* and *A. macleanii) matK* was pseudogenized. *matK* is widely used as a phylogenetic marker and we agree here that it is reasonable in most cases, but some *matK* sites are still under positive or diversifying selection.

Two genes of small ribosomal proteins (*rps15* and *rps16)* have close evo rates, *rps15* evo rate is 1,68 and rps16 evo rate is 1,52. No selection was detected in rps15 by any method, so we would assume that its evolution is neutral and only negative selection affects it. In *rps16* both MEME and FUBAR detect the same site under selection, so we are not able to argue there is no evidence of selection at all. In some species of the third evolutionary line *rps16* is pseudogenized, but in other lineages it is functional and thus it can be presumably under selection in these lineages.

High percentage of gaps in the small single copy region is very interesting at first glance (**Figure 6A**), but can be reasonably explained by lineage-specific insertion in *Nothoscordum* and *Narcissus*. Species of genus *Allium* do not have this insertion, so they have many common gaps in alignment.

GC content in different genes varies from 0.29 to 0.45 (**Supplementary Table 3** and **Figure 6B**). Genes with high evolutionary rates usually have lower GC content, e.g., *ycf1 (*0,29) or *rpl32* (0,3), whereas genes with low evolutionary rates usually have higher GC content, like *psbN* (0,46) or *atpH* (0,44), but this is not a tight dependence. Genes with low evolutionary rate can also have low GC content, like *psbL* (0,30) (see **Supplementary Table 3** and **Figures 6B,C**). The highest GC content is reported in repeats containing no protein-coding genes (see **Figures 6B,D**).

We found evidence of positive selection in the *matK* genes (its product takes part in the type II introns splicing), *accD* (the beta subunit of carboxyltransferase, which is part of the plastid acetyl-coA carboxylase), as well as in some *ndh* genes and *petL* (see **Table 1**). All these genes are involved in electron transport in photosystem II, in particular, in the cyclic transport of electrons. *petL* gene product is cytochrome b6-f complex subunit VI, a component of the cytochrome b6-f complex (Shi et al., 2016). Although only FUBAR results suggest some selection in *petL*, we found it worth mentioning due to its high relevance for cyclic electron transport. The *ndh* genes products are subunits of NADH dehydrogenase and are found to be under positive selection in many angiosperms (for example, Wang et al., 2018). Plants have mechanisms for enduring excess light irradiation using cyclic electron transfer (for example, it is known that mutant tobacco plants with a lost function of the *ndhB* gene demonstrate a greater sensitivity of photosynthesis to the width of stomata opening (Horváth et al., 2000). It is also known that the synthesis of the

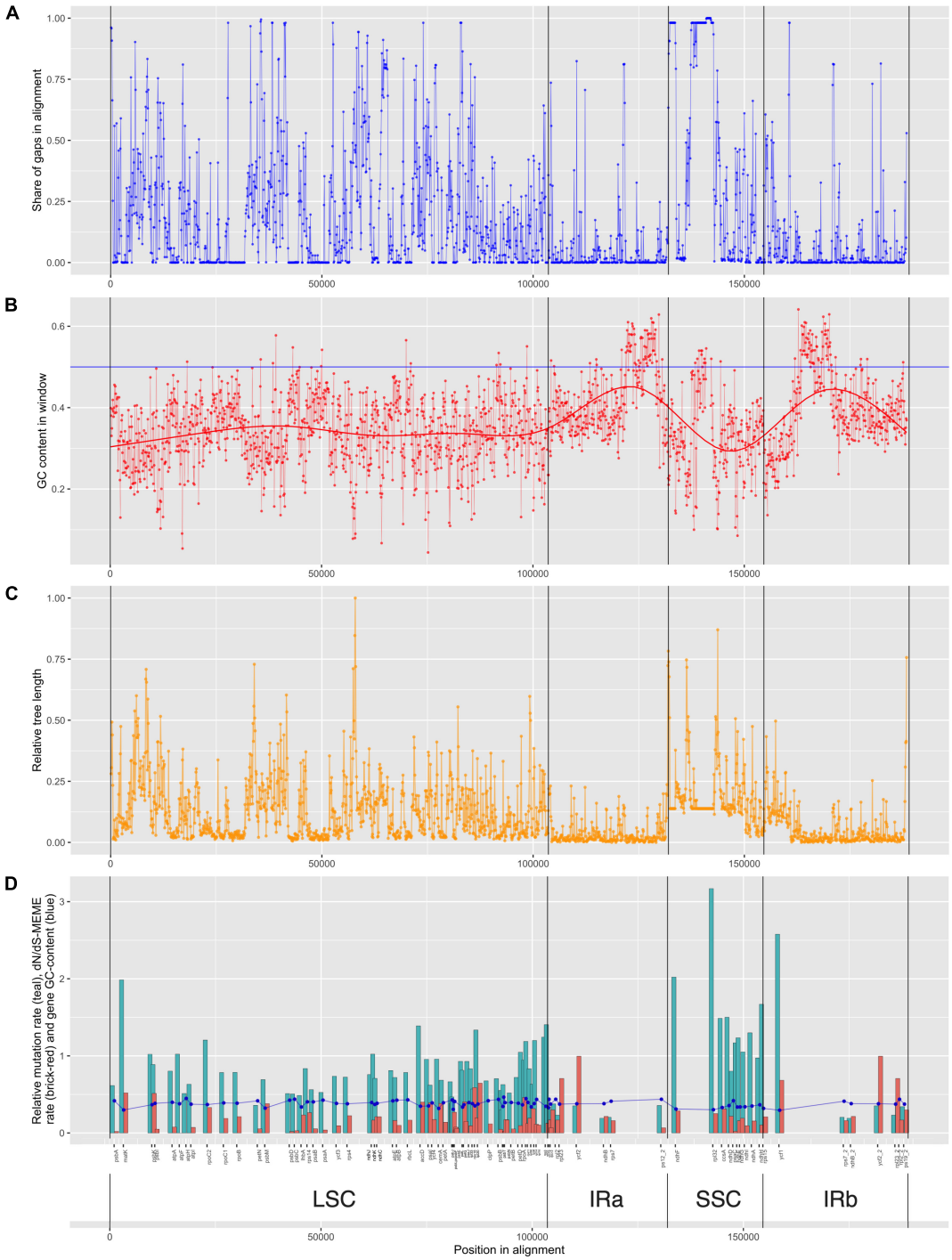


FIGURE 6 | Information on alignment and gene features. Structural components of plastome genomes and genes' positions are shown [*LSC*, large single-copied region; *SSC*, small single-copied region; *IRa* and *IRb*, inverted repeats **(A,B)**]. **(A)** Share of gaps in alignment in the window of 100 bp. Dots are connected with a line for visibility. **(B)** GC-content in the window of 100 bp. Dots are connected with a line for visibility. A bold line shows regression estimating of GC-content. **(C)** Relative tree length. Dots are connected with a line for visibility. **(D)** Relative mutation rate (teal), Kn/Ks-MEME rate (brick-red) bar-plot and gene GC-content (blue). Dots are connected with a line for visibility.

protein encoded by the *ndhA* gene occurs as a response to photooxidative stress (Martín et al., 1996, etc.). Photosystem II protein M gene *psbM* also has one confirmed site under positive selection. Taking all the above into account, we can assume that adaptive evolution of genes affecting cyclic electron transport around photosystem II, especially encoding the subunits of

TABLE 1 | Sites under positive of diversifying selection in protein coding genes.

Protein	Gene	Number of sites in alignment	Number of MEME sites under + selection	Number of FUBAR sites under + and − selection	Positions and number of MEME sites, confirmed by FUBAR
Acetyl-CoA carboxylase carboxyltransferase beta subunit	*accD*	480	6	+5 −23	96,156, 176, 477 Total 4
Cytochrome c biogenesis protein CcsA	*ccsA*	330	2	+3 −8	241
Protein TIC 214	*ycf1*	1,761	51	+9 −44	352, 705, 810, 851 Total 4
Hypothetical chloroplast RF21	*ycf2*	2,294	15	+35 −14	474, 475, 595, 691, 1,786 Total 5
NADH-plastoquinone_oxidoreductase_subunit_1	*ndhA*	366	3	+1 −76	206
NADH-plastoquinone_oxidoreductase_subunit_2	*ndhB*	511	1	+2 −12	380
NADH-plastoquinone_oxidoreductase_subunit_4L	*ndhE*	102	1	+1 −13	47
NADH-plastoquinone_oxidoreductase_subunit_4	*ndhD*	507	5	+3 −88	404, 454 Total 2
NADH-plastoquinone oxidoreductase subunit 5	*ndhF*	734	15	+6 −146	299, 510, 514,676 Total 4
NADH-plastoquinone_oxidoreductase_subunit_6	*ndhG*	183	1	+1 −18	34
NADH-plastoquinone oxidoreductase subunit K	*ndhK*	249	3	+2 −20	235, 240 Total 2
Photosystem II protein M	*psbM*	35	1	+1 −1	33
Ribosomal protein L16	*rpl16*	137	1	+1 −22	120
Ribosomal protein L20	*rpl20*	118	2	+3, −3	73
Maturase K	*matK*	521	13	+7 −40	92,324, 345 Total 3
Ribosomal protein S16	*rps16*	83	1	+2 −6	83
Ribulose-1,5-bisphosphate carboxylase oxygenase large subunit	*rbcL*	480	9	+10 −70	91,97, 225, 265 Total 4
RNA polymerase alpha subunit	*rpoA*	341	2	+10 −23	225
RNA polymerase beta subunit	*rpoB*	1,071	9	+5 −158	7, 160, 1,061 Total 3

NADH-plastoquinone oxidoreductase, is supposedly caused by abiotic factors of high altitude regions, mostly by intense light and UV radiation.

More support of this assumption came from aBSrel analysis of selected branches. We analyzed the list of aBSrel findings of selection events and the maximal altitudes of habitats and found most species to be referred to as alpine species (occurring at altitudes of 3,000 m and higher), except for two rather lowland species—*A. neriniflorum and A. strictum* (**Table 2**). We did not find any links between selection events and taxonomic position of the species in the genus. In all branches besides *A. paradoxumndh* genes were discovered to be under selection, only in *A. paradoxum* signatures of selection were found in maturase K gene.

Fisher's exact text for multiple samples provided some more support for hypothesis of selection pressure on *ndh* genes. *P*-value 8.822e-06 allows us to reject H0, some annotations has more selected sites. The only differing annotation is "*ndh* genes," it differs significantly from "photosynthesis" and marginally from "translation." Other ontologies do not give significant signal (**Supplementary Material 4**).

DISCUSSION

According to our analysis, plastomes in the genus *Allium* evolve predominantly in neutral mode. It is important to remark that only *A. paradoxum* significantly differed from other *Allium* species in having large genomic rearrangements. First of all, major rearrangements occurred in the 4,825 bp inversion (in the region between the *ndhE* and *rpl32* genes in SSC). *A. paradoxum* also demonstrates complete loss or pseudogenization of all its *ndh* (NADH dehydrogenase-like) genes. It was shown that *A. paradoxum* has the shortest known plastome of *Allium* species due to a large number of small deletions (145,819 bp). All the NADH-dehydrogenase complex genes (*ndh* genes) were found defunctionalized in the *A. paradoxum* plastome. This event may be supposedly associated with its adaptation to specific environmental conditions, rather uncommon among the species of *Allium* (shady humid forests). Some of *A. paradoxum* plastome genes have also lost their functionality in contrast to the rest of the studied species: *rpl22* was deleted and *rpl23* underwent pseudogenization (Omelchenko et al., 2020). We found that all other *Allium* species of the first evolutionary line had neither major rearrangements nor significant differences in the number of genes compared to *Allium* species of two other lines. The

TABLE 2 | Genes and branches with selection events.

Evolutionary line	Species	Habitat altitude range	Genes shown as being under selection with aBSREL
I	*Allium paradoxum*	No data (lives in shady forests)	*matK*
	Allium macranthum	2,700–4,200	*ndhF*, *rpl16*, *rpoC*
	Allium kingdonii	4,500–5,000	*ndhK*
II	*Allium neriniflorum*	500–2,000	*ndhF*, *rpoB*
III	*Allium strictum*	No data (lives on open rocks)	*ndhK*
	Allium przewalskianum	2,000–4,800	*ndhJ*
	Allium forrestii	2,700–4,200	*ndhF*
	Allium oschaninii	3,000	*ndhJ*, *rpoB*
	Allium cyathophorum	2,700–4,600	*ndhD*, *ndhJ*

results comparing *A. paradoxum* and *A. ursinum* were published (Omelchenko et al., 2020).

Plants have evolved a suite of adaptive responses to cope with variable environmental conditions. Ranges of *Allium* species are wide not only regarding latitude and longitude, but also altitude. Numerous species occupy mountain territories, budding high altitude species from each evolutionary line (Friesen et al., 2006). That is why putative adaptations to specific high altitude environments were of our special interest. Natural areas of species habitats are different in climatic parameters in local areas. Considering temperature or humidity we can only speak of parameters that are more or less typical for the habitat of a particular species. But maximal altitudes as well as the total amount and spectrum of sunlight are detected more precisely and we can consider it as a limiting factor that could direct natural selection.

Some genes with discovered sites under positive selection—*rpoA*,*rpoB* and*rbcL*—code essential enzymes for chloroplast protein synthesis (*rpoA* and *rpoB)* and sugar synthesis (*rbcL)*. *RbcL* gene evolves rapidly in many clades (see, for example, Sen et al., 2011; Galmés et al., 2014; Gao et al., 2019), and there is some data that its adaptive evolution may promote successful colonization of high altitude habitats (Zhao et al., 2019). The *rbcL* and *rpoC2* genes were involved in adaptation of *Cardamine* species to high or low altitudes. Authors of the study speculated that amino acid residues found to be under positive selection in RUBISCO could possibly be involved in the modulation of RUBISCO aggregation/activation and enzymatic specificity (Hu et al., 2015). Also the *rbcL* is under selection pressure in shade-tolerant *Oryza* species (Gao et al., 2019).

Other genes of those that were found under positive diversifying selection in *Allium* genus in this work were also found under selection pressure in shade-tolerant *Oryza* species, namely *accD* (Gao et al., 2019) and in *Cardamine* species adapted to different altitudes, namely *accD*, *ccsA*, *ycf1,rpl20* and *matK* (Hu et al., 2015). The *matK* gene also deserves a separate discussion. This gene is found under positive selection in our analysis of maturase K protein sequence in 3 sites from 521. The *matK* gene is under positive selection in many clades and its role in splicing of type II introns may also play a role in adaptation. Nevertheless, Kn/Ks ratio of *matK* is rather low (0,76) and we cannot be sure it is under selective pressure in the course of *Allium* adaptation to a wide range of environments, characteristic for this widespread genus.

Our analysis revealed that many genes from the *ndh* family are under positive selection—*ndhA*, *ndhB*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, and *ndhK*. *PetL* is found only by FUBAR, but we are considering it here as well because its product is involved in the same biological process. All these genes play certain roles in electron transport across the PSII and especially in cyclic electron transport. PETL codes a component of the cytochrome b6-f complex (Shi et al., 2016) and genes from *ndh* family code subunits of NADH-dehydrogenase. We know that loss-of-function mutants of *ndhB* gene in tobacco demonstrate sensitivity of photosynthesis to moderate stomatal closure (Horváth et al., 2000) and *ndhA* protein is synthesized in response to photooxidative treatment (Martín et al., 1996). *NdhK* is a part of NADH-oxidizing subcomplex (Elortza et al., 1999) and is positively selected in some angiosperm clades (Wang et al., 2018), in particular in species adapted to contrasting high/low altitude habitats (Hu et al., 2015) and shade-tolerant and sun-loving species (Gao et al., 2019).

Cyclic electron transport across PSII is activated in response to intense light or when stomata are closed to prevent production of reactive oxygen compounds and photodamage of PSII and PSI. Under excess light and/or closed stomata (CO_2 deficiency) plastoquinone pool regenerates and thereby a regeneration of NADP + occurs. NADPH cannot be expended in Calvin cycle in case of CO_2 deficiency. Substrate for ferredoxin-NADP-reductase is deficient and excess of excited electrons can cause membrane oxidation, generation of active oxygen species and damage of photosystems. To prevent oxidative stress, plants start cyclic electron transport.

Genes from the *ndh* family are strongly overrepresented in the selected gene list in our study—we found sites under selection in 7 genes of 11 of this family. The most plausible explanation of this fact is their relevance in adaptation to conditions of high light intensity. Taking into account pseudogenization of *ndh* genes in *A. paradoxum*, we can accept this hypothesis as the first choice. Data from Fisher's exact test provide moderate support to this hypothesis. *Ndh* is the only ontology that gives significant difference in ratio of selected to unselected sites and not in all comparisons. If we can speak of signatures of selection in this study, we can mention only genes from *ndh* family.

High altitudes have specific climate parameters, namely temperature extremes, strong winds, high solar radiation and lower atmospheric pressure (Rangwala and Miller, 2012). All these conditions affect leaf temperature, gas exchange constants, vapor pressure deficit, Michaelis-Menten constants for carboxylation and oxygenation and some other parameters affecting photosynthesis (Wang et al., 2017). Each extra 1,000 m in altitude makes photosynthesis more effective because of enhanced photosynthetic photon flux density and at the same time more problematic because of decreasing χ, which

is a decreasing function of the leaf-to-air vapor pressure deficit (VPD) (Wang et al., 2017).

In mountain habitats with their fast circadian changes of light intensity and humidity and, consequently, with long periods of light phase of photosynthesis going on with closed stomata, cyclic electron transport is essential for preventing photodamage of PSII and PSI. In some plants high contents of phenolic chemicals play a role in preventing photodamage without switching to cyclic electron transport, and in *A. ursinum* high concentration of polyphenolic compounds was indeed reported (Lachowicz et al., 2017; Tóth et al., 2018).

Oxidative stress may affect not only electron transport, but RNA synthesis as well. In *Mycobacterium tuberculosum* mutations in *rpoB* gene can modulate bacterial survival in high oxygenic environments inside macrophages (O'Sullivan et al., 2005). In *Escherichia coli* cells, growing in aerobic conditions, the guanine base is oxidized to 8-oxo-7,8-dihydroguanine, which can pair with adenine as well as cytosine. So fidelity of RNA synthesis depends upon proper work of RNA polymerases and substitutions in its β and β' subunit, coded by *rpoB* and *rpoC* genes are under selection in aerobic conditions (Inokuchi et al., 2013). Based on these facts we can hypothesize that highly oxygenic conditions in mountain habitats can cause selective pressure upon RNA polymerase function in chloroplasts, changing its protein sequence to establish stable function in conditions characterized by an excess of active oxygen, produced by overloaded Photosystem II.

So we can conclude that species of genus *Allium* appear to use various possible ways to prevent photodamage, but at high altitudes maintaining cyclic electron transport is most essential.

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**. The custom scripts used in this study for figures preparation have been deposited in Github (https://github.com/nikitin-p/Allium_analysis).

AUTHOR CONTRIBUTIONS

VS, AK, and AS: conceptualization, ideas formulation, or evolution of overarching research goals and aims. IA, PN, MB, EK, and AES: software and bioinformatics analysis. VS, IA, AK, DO, ML, and AS: investigation, conducting a research and investigation process, specifically performing the experiments, or data/evidence collection. MA and SK: resources and provision of plants. VS, PN, and AS: data curation, management activities to annotate (produce metadata), scrub data, and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse. VS, IA, PN, AK, and AS: writing—original draft preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation). ML: writing—review and editing. PN: visualization and preparation of the figures. AS: project administration, management and coordination responsibility for the research activity planning and execution, funding acquisition, and acquisition of the financial support for the project. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by RFBR Grant 18-04-01203.

ACKNOWLEDGMENTS

Authors thank Dr. Sergei Lysenkov for useful discussion of statisitcal methods.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.674783/full#supplementary-material

REFERENCES

Akashi, H., Osada, N., and Ohta, T. (2012). Weak selection and protein evolution. *Genetics* 192, 15–31. doi: 10.1534/genetics.112.140178

Bankevich, A., Nurk, S., Antipov, D. A., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing». *J. Comput. Biol.* 19, 455–477. doi: 10.1089/cmb.2012.0021

Blazier, J. C., Guisinger, M. M., and Jansen, R. K. (2011). Recent loss of plastid-encoded ndh genes within *Erodium* (Geraniaceae). *Plant Mol. Biol.* 76, 263–272. doi: 10.1007/s11103-011-9753-5

Bolger, A. M., Lohse, M., and Usade, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962, 1–14. doi: 10.1007/978-1-4939-9173-0_1

Chase, M. W., Reveal, J. L., and Fay, M. F. (2009). A subfamilial classification for the expanded asparagalean families Amaryllidaceae, Asparagaceae and Xanthorrhoeaceae. *Bot. J. Lin. Soc.* 161, 132–136. doi: 10.1111/j.1095-8339.2009.00999.x

Chernomor, O., von Haeseler, A., and Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Syst. Biol.* 65, 997–1008. doi: 10.1093/sysbio/syw037

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Daniell, H., Lin, C. S., Yu, M., and Chang, W.-J. (2016). Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol.* 17:134. doi: 10.1186/s13059-016-1004-2

de Pamphilis, C., and Palmer, J. (1990). Loss of photosynthetic and chlororespiratory genes from the plastid genome of a parasitic flowering plant. *Nature* 348, 337–339. doi: 10.1038/348337a0

Dong, W. L., Wang, R. N., Zhang, N. Y., Fan, W. B., Fang, M. F., and Li, Z. H. (2018). Molecular evolution of chloroplast genomes of *Orchid* species: insights into phylogenetic relationship and adaptive evolution. *Int. J. Mol. Sci.* 19:716. doi: 10.3390/ijms19030716

Doyle, J. J., and Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phyt. Bull.* 19, 11–15.

Elortza, F., Asturias, J. A., and Arizmendi, J. M. (1999). Chloroplast NADH dehydrogenase from *Pisum sativum*: characterization of its activity and cloning of ndhK gene. *Plant Cell Physiol.* 40, 149–154. doi: 10.1093/oxfordjournals.pcp.a029522

Filyushin, M. A., Beletsky, A. V., and Kochieva, E. Z. (2019). Characterization of the complete chloroplast genome of leek *Allium porrum* L. (Amaryllidaceae). *Mitochondrial. DNA Part B.* 4, 2602–2603. doi: 10.1080/23802359.2019.1640090

Filyushin, M. A., Beletsky, A. V., Mazur, A. M., and Kochieva, E. Z. (2016). The complete plastid genome sequence of garlic *Allium sativum* L. *Mitochondrial. DNA Part B.* 1, 831–832. doi: 10.1080/23802359.2016.1247669

Filyushin, M. A., Beletsky, A. V., Mazur, A. M., and Kochieva, E. Z. (2018). Characterization of the complete plastid genome of lop-sided onion *Allium obliquum* L. (Amaryllidaceae). *Mitochondrial. DNA part B.* 3, 393–394. doi: 10.1080/23802359.2018.1456369

Fonseca, L. H. M., and Lohmann, L. G. (2017). Plastome rearrangements in the Adenocalymma-Neojobertia“ clade (Bignonieae, Bignoniaceae) and its phylogenetic implications. *Front. Plant Sci.* 8:1875. doi: 10.3389/fpls.2017.01875

Friesen, N., Fritsch, R. M., and Blattner, F. R. (2006). Phylogeny and new intrageneric classification of *Allium* (Alliaceae) based on nuclear ribosomal DNA ITS sequences. *Aliso Syst. Evol. Bot.* 22, 372–395. doi: 10.5642/aliso.20062201.31

Fritsch, R. M. (2012). Geographic relations and morphological variation inside molecular clades of Central Asian *Allium* species of subg. Melanocrommyum (Amaryllidaceae)*. *Verh. Zool. Bot. Ges. Österreich* 148/149, 245–263.

Fritsch, R. M., Blattner, F. R., and Gurushidze, M. (2010). New classification of *Allium* L. subg. *Melanocrommyum* (Webb & Berthel) *Rouy* (Alliaceae) based on molecular and morphological characters. *Phyton* 49, 145–220.

Gale, J. (2004). Plants and altitude–revisited. *Ann. Bot.* 94:199. doi: 10.1093/aob/mch143

Galmés, J., Andralojc, P. J., Kapralov, M. V., Flexas, J., Keys, A. J., Molins, A., et al. (2014). Environmentally driven evolution of rubisco and improved photosynthesis and growth within the genus *Limonium* (Plumbaginaceae)». *N. Phytol.* 203, 989–999. doi: 10.1111/nph.12858

Gao, L. Z., Liu, Y. L., Zhang, D., Li, W., Gao, J., Liu, Y., et al. (2019). Evolution of Oryza chloroplast genomes promoted adaptation to diverse ecological habitats. *Commun. Biol.* 2:278. doi: 10.1038/s42003-019-0531-2

Govaerts, R., and Fritsch, R. (2021). *World Checklist of [Allium]*. Richmond: Facilitated by the Royal Botanic Gardens, Kew.

Green, B. R. (2011). Chloroplast genomes of photosynthetic eukaryotes: chloroplast genomes of photosynthetic eukaryotes. *Plant J.* 66, 34–44. doi: 10.1111/j.1365-313X.2011.04541.x

Guisinger, M. M., Kuehl, J. V., Boore, J. L., and Jansen, R. K. (2011). Extreme reconfiguration of plastid genomes in the angiosperm family Geraniaceae: rearrangements, repeats, and codon usage. *Mol. Biol. Evol.* 28, 583–600. doi: 10.1093/molbev/msq229

Haberhausen, G., and Zetsche, K. (1994). Functional loss of all ndh genes in an otherwise relatively unaltered plastid genome of the holoparasitic flowering plant *Cuscuta reflexa*. *Plant Mol. Biol.* 24, 217–222. doi: 10.1007/BF00040588

Haberle, R. C., Fourcade, H. M., Boore, J. L., and Jansen, R. K. (2008). Extensive rearrangements in the chloroplast genome of *Trachelium caeruleum* are associated with repeats and tRNA genes. *J. Mol. Evol.* 66, 350–361. doi: 10.1007/s00239-008-9086-4

Hanaoka, M., Kato, M., Anma, M., and Tanaka, K. (2012). SIG1, a Sigma Factor for the chloroplast RNA polymerase, differently associates with multiple DNA regions in the chloroplast chromosomes in vivo. *Int. J. Mol. Sci.* 13, 12182–12194. doi: 10.3390/ijms131012182

Hargreaves, A. L., Samis, K. E., and Eckert, C. G. (2014). Are species' range limits simply niche limits writ large? a review of transplant experiments beyond the range. *Am. Nat.* 183, 157–173. doi: 10.1086/674525

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with numPy. *Nature* 585, 357–362. doi: 10.1038/s41586-020-2649-2

Hilu, K. W., Borsch, T., Müller, K., Soltis, D. E., Soltis, P. E., Savolainen, V., et al. (2003). Angiosperm phylogeny based on matK sequence information. *Am. J. Bot.* 90, 1758–1776. doi: 10.3732/ajb.90.12.1758

Hiratsuka, J., Shimada, H., Whittier, R., Ishibashi, T., Sakamoto, M., Mori, M., et al. (1989). The complete sequence of the rice (*Oryza sativa*) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol. Gen. Genet.* 217, 185–194. doi: 10.1007/BF02464880

Horváth, E. M., Peter, S. O., Joët, T., Rumeau, D., Cournac, L., Horváth, G. V., et al. (2000). Targeted inactivation of the plastid ndhB gene in tobacco results in an enhanced sensitivity of photosynthesis to moderate stomatal closure. *Plant physiol.* 123, 1337–1350. doi: 10.1104/pp.123.4.1337

Hsu, C. Y., Wu, C. S., and Chaw, S. M. (2016). Birth of four chimeric plastid gene clusters in Japanese umbrella pine. *Genome Biol. Evol.* 8, 1776–1784. doi: 10.1093/gbe/evw109

Hu, S., Sablok, G., Wang, B., Qu, D., Barbaro, E., Viola, R., et al. (2015). Plastome organization and evolution of chloroplast genes in *Cardamine* species adapted to contrasting habitats. *BMC Genomics* 16:306. doi: 10.1186/s12864-015-1498-0

Hunter, J. D. (2007). Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi: 10.1109/mcse.2007.55

Huo, Y., Gao, L., Liu, B., Yang, Y., Kong, S., Sun, Y., et al. (2019). Complete chloroplast genome sequences of four *Allium* species: comparative and phylogenetic analyses. *Sci. Rep.* 9:12250. doi: 10.1038/s41598-019-48708-x

Inokuchi, H., Ito, R., Sekiguchi, T., and Sekiguchi, M. (2013). Search for proteins required for accurate gene expression under oxidative stress. *J. Biol. Chem.* 288, 32952–32962. doi: 10.1074/jbc.M113.507772

Jin, F. Y., Y, X., Xie, D. F., Li, H., Yu, Y., Zhou, S. D., et al. (2019). Comparative complete chloroplast genome analyses and contribution to the understanding of chloroplast phylogeny and adaptive evolution in subgenus *Anguinum*. *Russ. J. Genet.* 55, 872–884. doi: 10.1134/S1022795419070081

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kapralov, M. V., and Filatov, D. A. (2006). Molecular adaptation during adaptive radiation in the Hawaiian endemic genus *Schiedea*. *PLoS One* 1:e8. doi: 10.1371/journal.pone.0000008

Kazutaka, K., and Standley, D. M. (2013). MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kikuchi, S., Asakura, Y., Imai, M., Nakahira, Y., Kotani, Y., Hashiguchi, Y., et al. (2018). A Ycf2-FtsHi heteromeric AAA-ATPase complex is required for chloroplast protein import. *Plant Cell* 30, 2677–2703. doi: 10.1105/tpc.18.00357

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., et al. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. doi: 10.1101/gr.129684.111

Kosakovsky Pond, S. L., Frost, S. D. W., and Muse, S. V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676–679. doi: 10.1093/bioinformatics/bti079

Krinitsina, A. A., Sizova, T. V., Zaika, M. A., Speranskaya, A. S., and Sukhorukov, A. P. (2015). A rapid and cost-effective method for DNA extraction from archival herbarium specimens. *Biochemistry* 80, 1478–1484. doi: 10.1134/s0006297915110097

Lachowicz, S., Kolniak-Ostek, J., Oszmiański, J., and Wiśniewski, R. (2017). Comparison of phenolic content and antioxidant capacity of bear garlic (*Allium ursinum* L.) in different maturity stages: phenolics and antioxidants in bear garlic. *J. Food Process. Preserv.* 41, 1–10. doi: 10.1111/jfpp.12921

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Lee, J., Chon, J. K., Lim, J. S., Kim, E.-K., and Nah, G. (2017). Characterization of complete chloroplast genome of *Allium* victorialis and its application for barcode markers. *Plant Breed. Biotechnol.* 5, 221–227. doi: 10.9787/PBB.2017.5.3.221

Lee-Yaw, J. A., Kharouba, H. M., Bontrager, M., Mahony, C., Csergö, A. M., Noreen, A. M., et al. (2016). A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. *Ecol. Lett.* 19, 710–722. doi: 10.1111/ele.12604

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

Li, Q.-Q., Zhou, S.-D., He, X.-J., Yu, Y., Zhang, Y.-C., and Wei, X.-Q. (2010). Phylogeny and biogeography of *Allium* (Amaryllidaceae: Allieae) based on nuclear ribosomal internal transcribed spacer and chloroplast rps16 sequences, focusing on the inclusion of species endemic to China. *Ann. Bot.* 106, 709–733. doi: 10.1093/aob/mcq177

Lin, C. S., Chen, J., Huang, Y. T., Chan, M.-T., Daniell, H., Chang, W.-J., et al. (2015). The location and translocation of Ndh genes of chloroplast origin in the Orchidaceae family. *Sci. Rep.* 5:9040. doi: 10.1038/srep09040

Lin, C. S., Chen, J. J. W., Chiu, C. C., Hsiao, H. C. W., Yang, C. J., Jin, X. H., et al. (2017). Concomitant loss of NDH complex-related genes within chloroplast and nuclear genomes in some Orchids. *Plant J. Cell Mol. Biol.* 90, 994–1006. doi: 10.1111/tpj.13525

Liu, L., Yusupov, Z., Suyunkulov, H., and Jiang, Z. (2020). The complete chloroplast genome of *Allium ferganicum*. *Mitoch. DNA B Resour.* 5, 2772–2773. doi: 10.1080/23802359.2020.1788454

Logacheva, M. D., Krinitsina, A. A., Belenikin, M. S., Khafizov, K., Konorov, E. A., Kuptsov, S. V., et al. (2017). Comparative analysis of inverted repeats of polypod fern (Polypodiales) plastomes reveals two hypervariable regions. *BMC Plant Biol.* 17:255. doi: 10.1186/s12870-017-1195-z

Macadlo, L. A., Ibrahim, I. M., and Puthiyaveetil, S. (2020). Sigma Factor 1 in chloroplast gene transcription and photosynthetic light acclimation. *J. Exp. Bot.* 71, 1029–1038. doi: 10.1093/jxb/erz464

Maragheh, F. P., Janus, D., Senderowicz, M., Haliloglu, K., and Kolano, B. (2019). Karyotype analysis of eight cultivated *Allium* species». *J. Appl. Genet.* 60, 1–11. doi: 10.1007/s13353-018-0474-1

Martín, M., Casano, L. M., and Sabater, B. (1996). Identification of the product of ndhA gene as a thylakoid protein synthesized in response to photooxidative treatment. *Plant Cell Physiol.* 37, 293–298. doi: 10.1093/oxfordjournals.pcp.a028945

Martín, M., and Sabater, B. (2010). Plastid Ndh genes in plant evolution. *Plant Physiol. Biochem.* 48, 636–645. doi: 10.1016/j.plaphy.2010.04.009

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., et al. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534. doi: 10.1093/molbev/msaa015

Morton, B. R., and Clegg, M. T. (1993). A chloroplast DNA mutational hotspot and gene conversion in a noncoding region near rbcL in the Grass family (Poaceae). *Curr. Genet.* 24, 357–365. doi: 10.1007/BF00336789

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., et al. (2013). FUBAR: a fast, unconstrained bayesian approximation for inferring selection. *Mol. Biol. Evol.* 30, 1196–1205. doi: 10.1093/molbev/mst030

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., and Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* 8:e1002764. doi: 10.1371/journal.pgen.1002764

Namgung, J., Do, H. D. K., Kim, C., Choi, H. J., and Kim, J. H. (2021). Complete chloroplast genomes shed light on phylogenetic relationships, divergence time, and biogeography of Allioideae (Amaryllidaceae). *Sci. Rep.* 11:3262. doi: 10.1038/s41598-021-82692-5

Ogihara, Y., Terachi, T., and Sasakuma, T. (1991). Molecular analysis of the hot spot region related to length mutations in wheat chloroplast DNAs. I. nucleotide divergence of genes and intergenic spacer regions located in the hot spot region. *Genetics* 129, 873–884. doi: 10.1093/genetics/129.3.873

Omelchenko, D. O., Krinitsina, A. A., Belenikin, M. S., Konorov, E. A., Kuptsov, S. V., Logacheva, M. D., et al. (2020). Complete plastome sequencing of *Allium paradoxum* reveals unusual rearrangements and the loss of the Ndh genes as compared to *Allium ursinum* and other onions. *Gene* 726:144154. doi: 10.1016/j.gene.2019.144154

O'Sullivan, D. M., McHugh, T. D., and Gillespie, S. H. (2005). Analysis of rpoB and pncA mutations in the published literature: an insight into the role of oxidative stress in Mycobacterium tuberculosis evolution? *J. Antimicrob. Chemother.* 55, 674–679. doi: 10.1093/jac/dki069

R Core Team, R. (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raman, G., and Park, S. (2015). Analysis of the complete chloroplast genome of a medicinal plant, Dianthus superbus var. longicalycinus, from a comparative genomics perspective. *PLoS One* 10:e0141329. doi: 10.1371/journal.pone.0141329

Rambaut, A., and Drummond, A. J. (2012). *FigTree (version 1.4. 0)*. Edinburgh: Andrew Rambaut. Available online at: http://tree.bio.ed.ac.uk/software/figtree/

Rangwala, I., and Miller, J. R. (2012). Climate change in mountains: a review of elevation-dependent warming and its possible causes. *Clim. Change* 114, 527–547. doi: 10.1007/s10584-012-0419-3

Reback, J., McKinney, W., Jbrockmendel, Van Den Bossche, J., Augspurger, T., Cloud, P., et al. (2021). pandas-dev/pandas: pandas 1.2.3 (v1.2.3). *Zenodo*. doi: 10.5281/ZENODO.3509134

Sanchez-Puerta, M., Bachvaroff, T., and Delwiche, C. (2005). The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. *DNA Res.* 12, 151–156. doi: 10.1093/dnares/12.2.151

Sassone, A. B., Arroyo-Leuenberger, S. C., and Giussani, L. M. (2014). Nueva circunscripción de la tribu Leucocoryneae (Amaryllidaceae. Allioideae) *Darwinia. Nueva Serie* 2, 197–206. doi: 10.14522/darwiniana/2014.22.584

Schwarz, E. N., Ruhlman, T. A., Sabir, J. S. M., Hajrah, N. H., Alharbi, N. S., Al−Malki, A. L., et al. (2015). Plastid genome sequences of Legumes reveal parallel inversions and multiple losses of Rps16 in Papilionoids: parallel inversions and Rps16 losses in Legumes. *J. Syst. Evol.* 53, 458–468. doi: 10.1111/jse.12179

Sen, L., Fares, M. A., Liang, B. Gao L., Wang B., Wang T., et al. (2011). Molecular evolution of *rbcL* in three gymnosperm families: identifying adaptive and coevolutionary patterns. *Biol. Direct* 6:29. doi: 10.1186/1745-6150-6-29

Shah, N. C. (2014). Status of cultivated & wild *Allium* species in India, A review. *Scit. J.* 1, 28–36.

Shi, C., Hu, N., Huang, H., Gao, J., Zhao, Y. J., and Gao, L. Z. (2012). An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. *PLoS One* 7:e31468. doi: 10.1371/journal.pone.0031468

Shi, C., Wang, S., Xia, E. H., Jiang, J. J., Zeng, F. C., and Gao, L. Z. (2016). Full transcription of the chloroplast genome in photosynthetic eukaryotes. *Sci. Rep.* 6:30135. doi: 10.1038/srep30135

Silva, S. R., Diaz, Y. C., Penha, H. A., Pinheiro, D. G., Fernandes, C. C., Miranda, V. F., et al. (2016). The chloroplast genome of *Utricularia* reniformisSheds light on the evolution of the Ndh gene complex of terrestrial carnivorous plants from the Lentibulariaceae family. *PloS One* 11:e0165176. doi: 10.1371/journal.pone.0165176

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., and Kosakovsky Pond, S. L. (2015). Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* 32, 1342–1353. doi: 10.1093/molbev/msv022

Talevich, E., Invergo, B. M., Cock, P. J., and Chapman, B. A. (2012). Bio.Phylo: a unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinform.* 13:209. doi: 10.1186/1471-2105-13-209

Tamura, M. N., Yamashita, J., Fuse, S., and Haraguchi, M. (2004). Molecular phylogeny of monocotyledons inferred from combined analysis of plastid matK and rbcL gene sequences. *J. Plant Res.* 117, 109–120. doi: 10.1007/s10265-003-0133-3

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq – versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi: 10.1093/nar/gkx391

Tóth, T., Kovarovič, J., Bystrická, J., Vollmannova, A., Musilová, J., and Lenková, M. (2018). The content of polyphenols and antioxidant activity in leaves and flowers of wild garlic (*Allium ursinum* L.). *Acta Aliment.* 47, 252–258. doi: 10.1556/066.2018.47.2.15

Van Rossum, G., and Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vieira, L. N., Faoro, H., Fraga, H. P., Rogalski, M., de Souza, E. M., de Oliveira Pedrosa, F., et al. (2014). An improved protocol for intact chloroplasts and cpDNA isolation in conifers. *PLoS One* 9:e84792. doi: 10.1371/journal.pone.0084792

von Kohn, C., Kiełkowska, A., and Havey, M. J. (2013). Sequencing and annotation of the chloroplast DNAs and identification of polymorphisms distinguishing

normal male-fertile and male-sterile cytoplasms of onion. *Genome* 56, 737–742. doi: 10.1139/gen-2013-0182

Vvedensky, A. I. (1935). *Flora of USSR. V. 4. Genus 267*, ed. L. Allium (Moscow; Leningrad: Ed.of USSR Academy of Science).

Wakasugi, T., Tsudzuki, J., Ito, S., Nakashima, K., Tsudzuki, T., and Sugiura, M. (1994). Loss of all Ndh genes as determined by sequencing the entire chloroplast genome of the black pine *Pinus thunbergii*. *Proc. Natl. Acad. Sci.* 91, 9794–9798. doi: 10.1073/pnas.91.21.9794

Wambugu, P. W., Brozynska, M., Furtado, A., Waters, D. L., and Henry, R. J. (2015). Relationships of wild and domesticated rices (Oryza AA genome species) based upon whole chloroplast genome sequences. *Sci. Rep.* 5:13957. doi: 10.1038/srep13957

Wang, H., Prentice, I. C., Davis, T. W., Keenan, T. F., Wright, I. J., and Peng, C. (2017). Photosynthetic responses to altitude: an explanation based on optimality principles. *New Phytol.* 213, 976–982. doi: 10.1111/nph.14332

Wang, X., Zhou, T., Bai, G., and Zhao, Y. (2018). Complete chloroplast genome sequence of Fagopyrum dibotrys: genome features, comparative analysis and phylogenetic relationships. *Sci. Rep*. 8:12379. doi: 10.1038/s41598-018-30398-6

Waskom, M., Botvinnik, O., Gelbart, M., Ostblom, J., Hobson, P., Lukauskas, S., et al. (2020). mwaskom/seaborn: v0.11.1. *Zenodo*. doi: 10.5281/ZENODO.592845

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Wendelbo, P. (1971). *Flora Iranica. V. 76. Alliaceae*. Graz: Akademische Druck- und Verlagsanstalt.

Wheeler, E. J., Mashayekhi, S., McNeal, D. W., Columbus, J. T., and Pires, J. C. (2013). Molecular systematics of *Allium* subgenus *Amerallium* (Amaryllidaceae) in North America. *Am. J. Bot.* 100, 701–711. doi: 10.3732/ajb.1200641

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Müller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol. Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Xie, D. F., Tan, J. B., Yu, Y., Gui, L. J., Su, D. M., Zhou, S. D., et al. (2020). Insights into phylogeny, age and evolution of *Allium* (Amaryllidaceae) based on the whole plastome sequences. *Ann. Bot.* 125, 1039–1055. doi: 10.1093/aob/mcaa024

Xie, D. F., Yu, H. X., Price, M., Xie, C., Deng, Y. Q., Chen, J. P., et al. (2019). Phylogeny of Chinese *Allium* species in section daghestanica and adaptive evolution of *Allium* (Amaryllidaceae, Allioideae) species revealed by the chloroplast complete genome. *Front. Plant Sci.* 10:460. doi: 10.3389/fpls.2019.00460

Yan, C., Du, J., Gao, L., Li, Y., and Hou, X. (2019). The complete chloroplast genome sequence of watercress (*Nasturtium officinale* R. Br.): genome organization, adaptive evolution and phylogenetic relationships in Cardamineae. *Gene* 699, 24–36. doi: 10.1016/j.gene.2019.02.075

Yao, X., Tan, Y. H., Yang, J. B., Wang, Y., Corlett, R. T., and Manen, J. F. (2019). Exceptionally high rates of positive selection on the rbcL gene in the genus Ilex (Aquifoliaceae). *BMC Evol. Biol.* 19:192. doi: 10.1186/s12862-019-1521-1

Yusupov, Z., Deng, T., Volis, S., Khassanov, F., Makhmudjanov, D., and Tojibaev, K. (2020). Phylogenomics of *Allium* section Cepa (Amaryllidaceae) provides new insights on domestication of onion. *Plant Divers.* 43, 102–110. doi: 10.1016/j.pld.2020.07.008

Zeng, S., Zhou, T., Han, K., Yang, Y., Zhao, J., and Liu, Z. L. (2017). The complete chloroplast genome sequences of six Rehmannia species. *Genes* 8:103. doi: 10.3390/genes8030103

Zerbino, D. R., and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn Graphs. *Genome Res.* 18, 821–829. doi: 10.1101/gr.074492.107

Zhang, J., Yuan, H., Yang, Y., Fish, T., Lyi, S. M., Thannhauser, T. W., et al. (2016). Plastid ribosomal protein S5 is involved in photosynthesis, plant development, and cold stress tolerance in *Arabidopsis*. *J. Exp. Bot.* 67, 2731–2744. doi: 10.1093/jxb/erw106

Zhao, Y., Xu, F., Liu, J., Guan, F., Quan H., and Meng, F. (2019). The adaptation strategies of *Herpetospermum pedunculosum* (Ser.) Baill at altitude gradient of the Tibetan plateau by physiological and metabolomic methods. *BMC Genom.* 20:451. doi: 10.1186/s12864-019-5778-y

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.

frontiers in Genetics

BRIEF RESEARCH REPORT
published: 11 August 2021
doi: 10.3389/fgene.2021.676935

Genome and Transcriptome Sequencing of *Populus* × *sibirica* Identified Sex-Associated Allele-Specific Expression of the *CLC* Gene

*Elena N. Pushkova[1], George S. Krasnov[1], Valentina A. Lakunina[1], Roman O. Novakovskiy[1], Liubov V. Povkhova[1,2], Ekaterina M. Dvorianinova[1,2], Artemy D. Beniaminov[1], Maria S. Fedorova[1], Anastasiya V. Snezhkina[1], Anna V. Kudryavtseva[1], Alexey A. Dmitriev[1] and Nataliya V. Melnikova[1]**

[1] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia, [2] Moscow Institute of Physics and Technology, Moscow, Russia

OPEN ACCESS

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

Reviewed by:
Yuepeng Song,
Beijing Forestry University, China
Andrey Knyazev,
Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Russia
Konstantin V. Krutovsky,
University of Göttingen, Germany

***Correspondence:**
Nataliya V. Melnikova
mnv-4529264@yandex.ru

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

Received: 06 March 2021
Accepted: 28 May 2021
Published: 11 August 2021

Citation:
Pushkova EN, Krasnov GS, Lakunina VA, Novakovskiy RO, Povkhova LV, Dvorianinova EM, Beniaminov AD, Fedorova MS, Snezhkina AV, Kudryavtseva AV, Dmitriev AA and Melnikova NV (2021) Genome and Transcriptome Sequencing of Populus × sibirica Identified Sex-Associated Allele-Specific Expression of the CLC Gene. Front. Genet. 12:676935. doi: 10.3389/fgene.2021.676935

Transcriptome sequencing of leaves, catkin axes, and flowers from male and female trees of *Populus* × *sibirica* and genome sequencing of the same plants were performed for the first time. The availability of both genome and transcriptome sequencing data enabled the identification of allele-specific expression. Such an analysis was performed for genes from the sex-determining region (SDR). *P.* × *sibirica* is an intersectional hybrid between species from sections *Aigeiros* (*Populus nigra*) and *Tacamahaca* (*Populus laurifolia*, *Populus suaveolens*, or *Populus* × *moskoviensis*); therefore, a significant number of heterozygous polymorphisms were identified in the SDR that allowed us to distinguish between alleles. In the SDR, both allelic variants of the *TCP* (T-complex protein 1 subunit gamma), *CLC* (Chloride channel protein CLC-c), and *MET1* (DNA-methyltransferase 1) genes were expressed in females, while in males, two allelic variants were expressed for *TCP* and *MET1* but only one allelic variant prevailed for *CLC*. Targeted sequencing of *TCP*, *CLC*, and *MET1* regions on a representative set of trees confirmed the sex-associated allele-specific expression of the *CLC* gene in generative and vegetative tissues of *P.* × *sibirica*. Our study brings new knowledge on sex-associated differences in *Populus* species.

Keywords: *Populus*, poplar, sex, transcriptome sequencing, gene expression, DNA polymorphism, *CLC* gene

INTRODUCTION

In contrast to animals, hermaphroditism is common in angiosperm plants—about 90% of these species produce bisexual flowers, about 5% are monoecious and have both male and female flowers on the same individual, and only about 5% are dioecious and have male and female flowers on separate individuals (Yampolsky and Yampolsky, 1922; Renner, 2014). Besides, there are gynodioecious plant species (with female and hermaphroditic individuals) and androdioecious species (with male and hermaphroditic individuals) that are less common (Bawa, 1980). It is known that dioecy can emerge from hermaphroditism directly or form *via* gynodioecy, androdioecy, monoecy, and, in some cases, heterostyly (Bawa, 1980). Evolution to dioecy involves a positive selection of mutations that lead to male and female sterility (Spigler and Ashman, 2012). In

plants, divergence of sexes took place relatively recently and occurred fast, independently, and repeatedly, which led to very polymorphic mechanisms of sex determination (Charlesworth, 2002; Tanurdzic and Banks, 2004; Diggle et al., 2011; Bachtrog et al., 2014; Kafer et al., 2017). The genus *Populus* is presented by dioecious species and, due to extensive genetic studies (Tuskan et al., 2006; Jansson and Douglas, 2007; Jansson et al., 2010), is a promising object for research of differences between sexes. Poplars are wind-pollinated trees, and different species are easily crossed, resulting in the emergence of natural interspecific hybrids and a high level of genetic diversity (Rae et al., 2007; Roe et al., 2014; Jiang et al., 2016).

Most *Populus* species have an XY system of sex determination, while *Populus alba* has a ZW system. Sex-specific DNA polymorphisms were identified in *Populus* species, and the sex-determining region (SDR) was mapped to the pericentromeric region of chromosome 19 in aspens and the peritelomeric region of chromosome 19 in most studied poplars, except for *Populus euphratica* (the SDR is located on chromosome 14) (Gaudet et al., 2008; Yin et al., 2008; Pakull et al., 2009, 2011, 2015; Paolucci et al., 2010; Kersten et al., 2014; Geraldes et al., 2015; McKown et al., 2017; Melnikova et al., 2019; Müller et al., 2020; Xue et al., 2020; Yang et al., 2020; Zhou et al., 2020).

It was shown that *ARABIDOPSIS RESPONSE REGULATOR 17* (*ARR17*) ortholog plays a key role in sex determination. Involved in the cytokinin pathway, this gene is presented in genomes of males and females of *Populus* species with the XY system of sex determination, but its partial repeats were identified only in SDRs of the males. It was suggested that the locus of these repeats produces small RNAs that silence *ARR17 via* DNA methylation. *ARR17* works as a sex switch: male flowers are formed when *ARR17* is off and female ones when *ARR17* is on (Müller et al., 2020).

In the SDR of male *Populus deltoides*, two long Y-specific hemizygous sequences (YHSs) were identified: YHS1 was about 35 kb, and YHS2 was about 4.3 kb. YHS1 included two male-specific genes: one, named *FERR-R*, contained partial duplications of the *FERR* gene (named *ARR17* in the study of Müller et al., 2020) and repressed the formation of female generative organs, likely through the methylation of the *FERR* gene and cleavage of its transcript, and the other one, named *MSL*, belonged to the LTR/Gypsy transposon family and transcribed long non-coding RNA, which probably promoted the development of male reproductive organs of *P. deltoides*. Complete *MSL* sequence was also identified in males of *Populus simonii* but not *Populus davidiana* and *Populus tremula*. The SDR also contained genes encoding T-complex protein 1 subunit gamma (TCP), Chloride channel protein CLC-c (CLC), and DNA-methyltransferase 1 (MET1), which were present in both males and females of *P. deltoides* (Xue et al., 2020).

In *Populus trichocarpa*, about 50 kb of its SDR were identified as male-specific (these sequences were absent in female plants). The SDR also contained five genes, which were present in both males and females of *P. trichocarpa*—genes encoding TCP, CLC, MET1, and leucine-rich repeat-containing protein (NB-ARC), and also an unknown gene. Genomic sites that were homozygous in females and heterozygous in males were revealed, and alternative Y and X haplotypes were identified (Zhou et al., 2020).

Differences in gene expression between male and female poplars and aspens were also studied. Overexpression of a gene that is necessary for the formation of female reproductive organs (named *ARR17*, *FERR*, or *RR* in different studies) was revealed in female genotypes at early stages of flower development (Cronk et al., 2020; Müller et al., 2020; Xue et al., 2020; Yang et al., 2020), while the expression of *MSL* was male-specific and continuous (Xue et al., 2020). Sex-specific gene expression was observed in flowers but not in leaves of *Populus balsamifera*: female-biased genes were related to photosynthesis, while male-biased genes were related to mitochondria (Sanderson et al., 2019). The analysis of transcriptomic data for male and female reproductive organs of *P. balsamifera* revealed gene expression trajectories during flower development and male-biased expression of two MADS-box genes, *APETALA3* and *PISTILLATA* (Cronk et al., 2020). Besides, in several studies of *Populus* species, sex-associated differences in gene expression were observed under particular, predominantly unfavorable, conditions (Melnikova et al., 2017; Han et al., 2018; Song et al., 2019).

Despite the significant improvement in our understanding of sex determination in *Populus* species, our knowledge of sex-specific differences in this genus is still incomplete. Further studies concerning male and female distinctions in poplars and aspens are necessary, and genomic and transcriptomic data for different species of the genus *Populus* should be obtained and analyzed. In the present study, to identify sex differences, we performed genome and transcriptome sequencing for male and female plants of one of the most common poplars in the cities of central Russia—*Populus* × *sibirica*, which is an intersectional hybrid likely between species from section *Aigeiros* (*Populus nigra*) and section *Tacamahaca* (the exact progenitor is still unknown, it could be *Populus laurifolia*, *Populus suaveolens*, or *Populus* × *moskoviensis*) (Mayorov et al., 2012; Kostina and Nasimovich, 2014; Kostina et al., 2017).

MATERIALS AND METHODS

Plant Material

We used plant material of *P.* × *sibirica* trees growing in Moscow within the territory from 55°41′29″N to 55°42′35″N and from 37°33′33″E to 37°35′26″E. Leaves, catkin axes, and flowers were collected from male and female plants during the beginning of flowering, immediately frozen in liquid nitrogen, and stored at −70°C until further use. Samples from one male and one female trees were used for whole-genome and transcriptome sequencing, and samples from 10 male and 10 female trees for targeted sequencing of DNA and cDNA.

RNA Extraction and Transcriptome Sequencing

For RNA extraction, the Quick-RNA Miniprep Kit (Zymo Research, Irvine, CA, United States) was used. Plant samples (three for each tissue) were homogenized in a lysis buffer with solid glass beads (Sigma-Aldrich, St. Louis, MO, United States)

using MagNA Lyser (Roche, Basel, Switzerland), and further manipulations were performed following the manufacturer's protocol with DNase I treatment step. RNA quality and concentration were evaluated on 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, United States) with Agilent RNA 6000 Nano Kit (Agilent Technologies) and Qubit 2.0 (Life Technologies, Carlsbad, CA, United States) with Qubit RNA BR Assay Kit (Life Technologies) respectively.

For cDNA library preparation from 1 μg of total RNA, the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, Hitchin, United Kingdom) and NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (New England Biolabs) were used. cDNA library quality and concentration were evaluated on 2100 Bioanalyzer (Agilent Technologies) with Agilent DNA 1000 Kit (Agilent Technologies) and Qubit 2.0 (Life Technologies) with Qubit dsDNA HS Assay Kit (Life Technologies) respectively. Transcriptome sequencing of leaves, catkin axes, and flowers of male and female *P.* × *sibirica* plants (in total, six libraries) was performed on NextSeq 500 (Illumina, San Diego, CA, United States) with a read length of 86 bp.

DNA Extraction and Whole-Genome Sequencing

Populus × *sibirica* leaves were homogenized in the lysis buffer from the DNeasy Plant Mini Kit (Qiagen, Germantown, MD, United States) with solid glass beads (Sigma-Aldrich) using MagNA Lyser (Roche), then, DNA extraction was performed following the manufacturer's protocol. The extracted DNA was fragmented on an S220 ultrasonic homogenizer (Covaris, Woburn, MA, United States), and 1 μg of fragmented DNA was used to prepare the library using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer's protocol with size selection of adapter-ligated DNA of about 600 bp. The quality and concentration of DNA libraries were evaluated using 2100 Bioanalyzer (Agilent Technologies) with Agilent DNA 1000 Kit (Agilent Technologies) and Qubit 2.0 (Life Technologies) with Qubit dsDNA HS Assay Kit (Life Technologies) respectively. DNA libraries were sequenced on HiSeq 2500 (Illumina) with a read length of 125 + 125 bp.

Targeted Genome and Transcriptome Sequencing

DNA was extracted from leaves, and RNA was extracted from leaves, catkin axes, and flowers of 10 male and 10 female plants of *P.* × *sibirica* as described above. One microgram of RNA for each sample was treated with DNase I (Thermo Fisher Scientific, Waltham, MA, United States) and used for reverse transcription with random hexamer primers (Evrogen, Moscow, Russia) and Mint reverse transcriptase (Evrogen) following the manufacturer's protocol. Primers were designed for amplification of parts of the *TCP*, *CLC*, and *MET1* genes and further sequencing on the Illumina platform as described in our previous works (Melnikova et al., 2019; Dmitriev et al., 2020; Novakovskiy et al., 2020). In brief, two PCRs were used for the library preparation. The first PCR enabled amplification of target gene regions and the addition of universal adapters to amplicons according to the Illumina protocol[1]. In the second PCR, Nextera XT v2 index primers containing dual-index barcodes and sequencing adapters were used. Primer sequences are listed in **Supplementary Data 1**. For each of the 20 genotypes, the amplicons were obtained independently for DNA from leaves and cDNA from leaves, catkin axes, and flowers at the first PCR and then labeled with unique sample-specific indexes at the second PCR.

Data Analysis

For the expression analysis, RNA-Seq reads were trimmed using Trimmomatic 0.32 (Bolger et al., 2014), mapped to the male *P. trichocarpa* "Stettler 14" genome[2] (Hofmeister et al., 2020) using STAR 2.8 (Dobin et al., 2013), and then quantified (a) for the annotated genes using the featureCounts tool from the Subread package 1.6.0 (Liao et al., 2014) and (b) for the region Chr18:16,200,000–16,320,000 (200-bp intervals) corresponding to the SDR of the male *P. trichocarpa* "Stettler 14" (Zhou et al., 2020) with BEDTools 2.26.0 (Quinlan and Hall, 2010). It should be noted that in the "Stettler 14" genome assembly, the SDR is located on chromosome 18, but the genetic map showed that its correct place is on chromosome 19 (Zhou et al., 2020). The derived read count values were analyzed using edgeR 3.28.1 (for R 3.6.3) (Robinson et al., 2010). TMM normalization and quasi-likelihood *F*-test were applied. Pairwise distance (dissimilarity) was calculated as "1 – *r*," where *r* is the Spearman's rank correlation coefficient between gene expression profiles. For distance calculation, we used genes with an average CPM (counts per million) >16.

For the identification of polymorphisms in the SDR of *P.* × *sibirica* based on the obtained sequencing data, reads were trimmed using Trimmomatic and mapped to the male *P. trichocarpa* "Stettler 14" genome with BWA-MEM 0.7.17 (Li, 2013). The derived BAM files were preprocessed with Picard tools 2.21.3, including verification of mate information (the FixMateInformation tool), marking duplicated reads [the MarkDuplicatesWithMateCigar tool; only for whole-genome sequencing (WGS) paired-end reads], or splitting reads spanning introns (SplitNCigarReads from GATK 4.1.9.0; only for RNA-Seq reads). Variant calling was performed using FreeBayes 1.3.2 (Garrison and Marth, 2012) for the SDR of the male *P. trichocarpa* "Stettler 14." We filtered out variant candidates with Phred quality <15.

RESULTS

Transcriptome sequencing of leaves, catkin axes, and flowers of male and female *P.* × *sibirica* plants produced from 10.7 to 12.8 million 86-bp single-end reads for each sample. WGS gave 36 million paired-end 2 × 125-bp reads for the male poplar and 52 million reads for the female one that corresponded to about 18× and 26× genome coverage respectively. The raw data were

[1]https://support.illumina.com/documents/documentation/chemistry_documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf
[2]https://phytozome-next.jgi.doe.gov/info/PtrichocarpaStettler14_v1_1

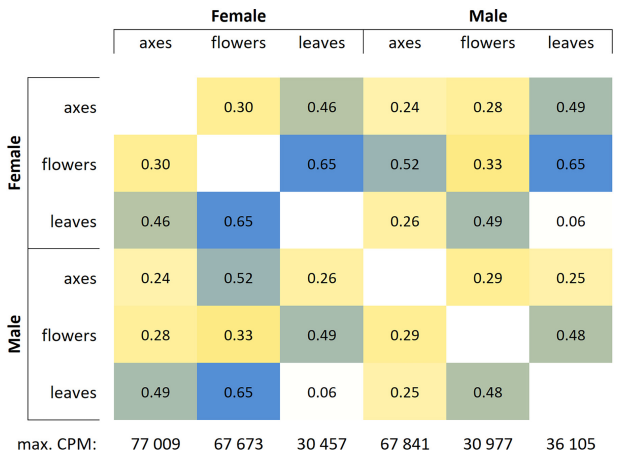


FIGURE 1 | Distance matrix for gene expression profiles in leaves, catkin axes, and flowers of male and female *Populus* × *sibirica* plants. Transcriptome sequencing data; genes with CPM (counts per million) >16 were taken into account; max. CPM, maximum CPM among all genes.

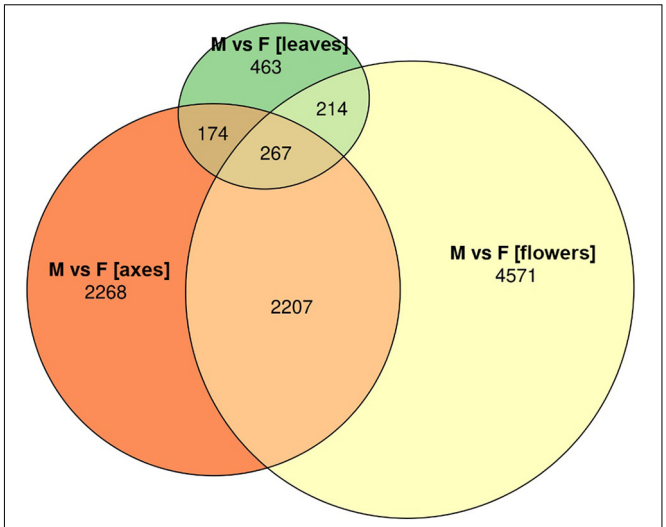


FIGURE 2 | Euler diagram for genes differentially expressed between male and female *Populus* × *sibirica* plants in leaves, catkin axes, and flowers. Transcriptome sequencing data; genes with CPM (counts per million) >8 and at least twofold expression level differences were taken into account; M, male; F, female.

deposited in the NCBI Sequence Read Archive (SRA) under the BioProject accession number PRJNA644206.

Based on RNA-Seq data, we assessed distinctions in gene expression profiles between different male and female *P.* × *sibirica* tissues. Multidimensional scaling plots (**Supplementary Data 2, 3**) and a distance matrix (**Figure 1**) showed that the expression profiles of male and female leaves differed much less than those of flowers and catkin axes. So, when comparing male and female leaves, we revealed 1,118 differentially expressed genes (DEGs), having at least twofold expression level differences and a sufficient expression level (CPM > 8), while for catkin axes and flowers, 4,916 and 7,259 DEGs were identified (**Figure 2**). Thus, sex did not result in such significant differences in the transcriptomes of vegetative organs, as it happened for generative ones. These results are in concordance with those from previous studies in *P. balsamifera* that revealed more significant expression differences between male and female flowers compared to the leaves (Sanderson et al., 2019).

Based on whole-genome and transcriptome sequencing data, coverage profiles and polymorphisms were evaluated for the SDR of *P.* × *sibirica* using *P. trichocarpa* "Stettler 14" genome assembly as a reference [as mentioned above, the SDR in this assembly is located on chromosome 18, but its correct place is on chromosome 19 (Zhou et al., 2020); however, the SDR sequence itself is complete that allows us to use it for the analysis]. The results of the analysis are presented in **Supplementary Data 4**. As in previous studies on SDRs of male poplars (Müller et al., 2020; Xue et al., 2020; Zhou et al., 2020), we revealed male-specific genomic regions, the coverage of which was identified only within genomic data of the male; however, expression from them was not observed (**Supplementary Data 4**). This can be associated with the type of plant material used in the present study (particular tissue and stage of development).

A more detailed analysis was performed for five genes, which were identified in the SDR of *P. trichocarpa* "Stettler 14" in the study of Zhou et al. (2020): *TCP* (PtStettler14.18G127900, Chr18:16,275,593-16,279,859, reverse), *CLC* (PtStettler14 .18G127800, Chr18:16,268,169-16,272,894, reverse), *MET1* (PtStettler14.18G127700, Chr18:16,249,745-16,259,190, reverse), *NB-ARC* (PtStettler14.18G127600, Chr18:16,226,313-16,236,215, reverse), and an unknown gene (PtStettler14.18G127500, Chr18:16,214,199-16,215,600, forward) (Zhou et al., 2020). Three of these genes (*TCP*, *CLC*, and *MET1*) were expressed in *P.* × *sibirica* tissues under study (leaves, catkin axes, and flowers) that enabled the identification of polymorphisms in their transcripts and further comparison with polymorphisms from our genome sequencing data. *P.* × *sibirica* is an intersectional hybrid (Mayorov et al., 2012; Kostina and Nasimovich, 2014; Kostina et al., 2017), so a significant number of heterozygous polymorphisms were identified in the SDR. In the tissues of the female, two variants of nucleotides were revealed in transcriptomic data in sites where two variants of nucleotides were found in genomic data; thus, both allelic variants of *TCP*, *CLC*, and *MET1* genes were expressed in the female (**Supplementary Data 5**). In the male, we observed the same for *TCP* and *MET1* but not *CLC*. For this gene, only one nucleotide variant prevailed in male transcriptomic data for leaves, catkin axes, and flowers in sites that had two variants of nucleotides in male genomic data (**Supplementary Data 6**).

To confirm our results, we performed targeted sequencing of *TCP*, *CLC*, and *MET1* gene regions for DNA and cDNA (RNA extracted from leaves, catkin axes, and flowers) samples of 10 male and 10 female plants of *P.* × *sibirica* (in total, 80 samples). It should be noted that two primer pairs for the *CLC* gene amplified regions with introns. Such an approach

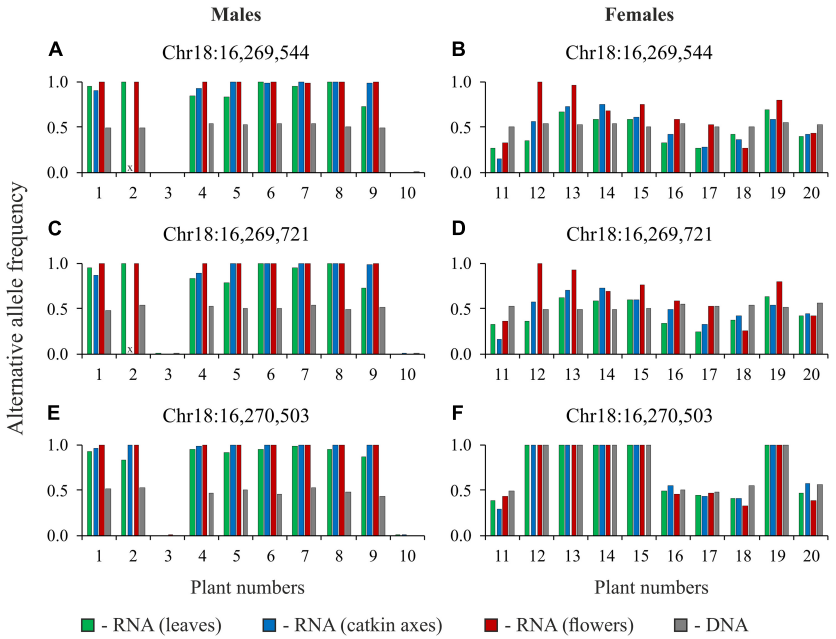


FIGURE 3 | Alternative allele frequencies for heterozygous sites of the *CLC* gene for DNA and cDNA (RNA from leaves, catkin axes, and flowers) samples of 10 male **(A,C,E)** and 10 female **(B,D,F)** plants of *Populus* × *sibirica*. Targeted sequencing data; chromosome number and nucleotide position are given above each pair of charts according to the *P. trichocarpa* "Stettler 14" genome assembly; plant numbers (1–20) correspond to plant numbers in **Supplementary Data 7**; x, no data.

enabled us to control the absence of DNA in RNA samples through the analysis of sequencing data. The results of variant calling in amplicons of the *TCP*, *CLC*, and *MET1* genes are presented in **Supplementary Data 7**. When two variants of nucleotides (frequencies of alternative and reference alleles about 0.5) were revealed in a particular site of the *TCP* or *MET1* amplicons obtained from DNA, two variants of nucleotides (with significant frequencies of both alternative and reference alleles) were also revealed in the amplicons obtained from cDNA (RNA from leaves, catkin axes, and flowers) of the same genotype for most males and females. As expected, the same was observed for the *CLC* gene in females. However, for *CLC* in males, in sites with frequencies of alternative and reference alleles about 0.5 for the amplicons from DNA, only one nucleotide variant dominated (frequency of an alternative allele or a reference one was close to 1) in the amplicons from cDNA (RNA from leaves, catkin axes, and flowers) of the same genotype. The sets of heterozygous sites in the *TCP*, *CLC*, and *MET1* genes were different between males and females of *P.* × *sibirica* that is probably associated with suppression of recombination in the SDR. However, some common heterozygous sites were also revealed, and data for three of such sites of the *CLC* gene are presented in **Figure 3** as an illustration of the sex-associated allele-specific expression of this gene.

DISCUSSION

Due to recent studies on the genetics of sex in *Populus* species, especially obtaining high-quality genome assemblies of male individuals (Geraldes et al., 2015; McKown et al., 2017; Xue et al., 2020; Yang et al., 2020; Zhou et al., 2020), a detailed investigation of SDRs became possible. In the present work, the analysis of the SDR of *P.* × *sibirica* was performed for the identification of sex-specific differences. Based on the comparison of whole-genome and transcriptome sequencing data for male and female plants of *P.* × *sibirica*, we evaluated allele frequencies in heterozygous sites of genes from the SDR. Since *P.* × *sibirica* is an intersectional hybrid, a significant number of heterozygous sites were revealed that enabled us to identify the predominant expression of the *CLC* gene only from one allele in generative and vegetative tissues of the male but not female. Then, we used targeted sequencing of DNA and cDNA samples of 20 *P.* × *sibirica* plants (10 males and 10 females) and confirmed the sex-associated allele-specific expression of the *CLC* gene in this poplar. Moreover, based on obtained by us high-quality genome assembly of male *P.* × *sibirica* with separated Y and X haplotypes of the SDR (unpublished data), it can be concluded that it is the Y SDR haplotype with the suppressed *CLC* expression.

The expression of genes from the SDR, including *CLC*, was analyzed in previous studies of poplars. In *P. balsamifera*, *CLC* upregulation was revealed in flowers of females compared to those of males (Sanderson et al., 2019). Opposite expression polarity was revealed during the development of reproductive organs in male and female *P. balsamifera* for two variants of *TCP*, *CLC*, and *MET1* (located on both chromosome 18 and scaffold 42) when the genome of female *P. trichocarpa* was used as a reference (Cronk et al., 2020). However, no consistent differences in the expression of the *TCP*, *CLC*, and *MET1*

genes were revealed between male and female *P. deltoides* in developing flowers (Xue et al., 2020). Thus, for the first time, we revealed sex-associated allele-specific expression for one of the three genes from the SDR (for *CLC* but not for *TCP* and *MET1*) in generative and also vegetative tissues of poplar, namely *P.* × *sibirica*.

CLC genes participate in Cl^- and NO_3^- transport, are involved in response to salt stress, and are related to the efficiency of nitrogen use in plants (Liao et al., 2018; Liu et al., 2020; Subba et al., 2020). Therefore, differences in *CLC* expression in vegetative tissues of *P.* × *sibirica* could be of interest in terms of determination of the possible role of sex in adaptation to particular environments. A significant number of studies concerning differences in stress response of male and female plants of *Populus* species were performed; however, there is no consensus on the association of stress resistance with sex (Melnikova et al., 2017). The increased transcription level of *CLC* was revealed under high salinity conditions in female poplars but not in male ones, and the authors suggested that males had more efficiency in Cl^- homeostasis than females (Jiang et al., 2012). Besides, under salt stress, *CLC* was upregulated in *Populus pruinosa*, which is distributed in deserts with underground water close to the surface, but not in *P. euphratica*, which occurs in deserts with deep underground water (Zhang et al., 2014). It was also shown that overexpression of the *CLC* gene from soybean in transgenic poplars improved salt tolerance (Sun et al., 2013).

Thus, the *CLC* gene is involved in salt stress response in poplars, and our data on its sex-associated allele-specific expression in vegetative tissues of *P.* × *sibirica* suggest that differences between male and female poplars could take place not only in generative organs but also in whole plants and are implicated in stress resistance. However, further research is needed to clarify this issue.

CONCLUSION

Comprehensive genetic and epigenetic studies of phylogenetically distant *Populus* species are essential for the identification of male- and female-specific differences and sex-associated pathways. In our study, transcriptome sequencing of leaves, catkin axes, and flowers from male and female trees of *P.* × *sibirica* and genome sequencing of the same plants were performed for the first time. In the SDR, both allelic variants of the *TCP*, *CLC*, and *MET1* genes were expressed in females, while, in males, both allelic variants were expressed for *TCP* and *MET1*, but only one variant prevailed for the *CLC* gene. Targeted sequencing of *TCP*, *CLC*, and *MET1* gene regions obtained from DNA and cDNA (RNA extracted from leaves, catkin axes, and flowers) samples of 10 male and 10 female plants of *P.* × *sibirica* confirmed the predominant expression of only one allelic variant of the *CLC* gene in males. Generalization of the currently available data on *CLC* expression allows us to suggest that the identified sex-associated allele-specific expression of this gene in both generative and vegetative organs can be involved in different salt resistance of males and females of *Populus* species.

DATA AVAILABILITY STATEMENT

The obtained sequencing data can be found in NCBI under the BioProject accession number PRJNA644206.

AUTHOR CONTRIBUTIONS

AD and NM conceived and designed the work. EP, RN, LP, AB, MF, and AS performed the experiments. EP, GK, VL, ED, AK, AD, and NM analyzed the data. EP, GK, ED, AD, and NM wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was funded by RFBR according to the research projects 18-34-20113 mol_a_ved (genome and transcriptome sequencing of *P.* × *sibirica*) and 20-34-90159 (targeted sequencing of the *TCP*, *CLC*, and *MET1* genes for DNA and cDNA samples extracted from 20 plants of *P.* × *sibirica*).

ACKNOWLEDGMENTS

We thank the Center for Precision Genome Editing and Genetic Technologies for Biomedicine, EIMB RAS for providing computing power and techniques for data analysis. This work was performed using the equipment of EIMB RAS "Genome" center (http://www.eimb.ru/ru1/ckp/ccu_genome_ce.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.676935/full#supplementary-material

Supplementary Data 1 | Primers for the first stage of DNA library preparation.

Supplementary Data 2 | Multidimensional scaling plot (dimensions 1 and 2) for gene expression profiles in leaves, catkin axes, and flowers of male and female *Populus* × *sibirica* plants.

Supplementary Data 3 | Multidimensional scaling plot (dimensions 1 and 3) for gene expression profiles in leaves, catkin axes, and flowers of male and female *Populus* × *sibirica* plants.

Supplementary Data 4 | Polymorphisms in the SDR (including *TCP*, *CLC*, and *MET1* genes) for male and female *Populus* × *sibirica* plants based on genome and transcriptome sequencing data.

Supplementary Data 5 | The selected polymorphisms in the *TCP*, *CLC*, and *MET1* genes for the female *Populus* × *sibirica* plant based on genome and transcriptome sequencing data.

Supplementary Data 6 | The selected polymorphisms in the *TCP*, *CLC*, and *MET1* genes for the male *Populus* × *sibirica* plant based on genome and transcriptome sequencing data.

Supplementary Data 7 | Polymorphisms in the *TCP*, *CLC*, and *MET1* genes for male and female *Populus* × *sibirica* plants based on targeted sequencing data.

REFERENCES

Bachtrog, D., Mank, J. E., Peichel, C. L., Kirkpatrick, M., Otto, S. P., Ashman, T. L., et al. (2014). Sex determination: why so many ways of doing it? *PLoS Biol.* 12:e1001899. doi: 10.1371/journal.pbio.1001899

Bawa, K. S. (1980). Evolution of dioecy in flowering plants. *Annu. Rev. Ecol. Syst.* 11, 15–39.

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

Charlesworth, D. (2002). Plant sex determination and sex chromosomes. *Heredity* 88, 94–101. doi: 10.1038/sj.hdy.6800016

Cronk, Q., Soolanayakanahally, R., and Bräutigam, K. (2020). Gene expression trajectories during male and female reproductive development in balsam poplar (*Populus balsamifera* L.). *Sci. Rep.* 10:8413. doi: 10.1038/s41598-020-64938-w

Diggle, P. K., Di Stilio, V. S., Gschwend, A. R., Golenberg, E. M., Moore, R. C., Russell, J. R., et al. (2011). Multiple developmental processes underlie sex differentiation in angiosperms. *Trends Genet.* 27, 368–376. doi: 10.1016/j.tig.2011.05.003

Dmitriev, A. A., Kezimana, P., Rozhmina, T. A., Zhuchenko, A. A., Povkhova, L. V., Pushkova, E. N., et al. (2020). Genetic diversity of *SAD* and *FAD* genes responsible for the fatty acid composition in flax cultivars and lines. *BMC Plant Biol.* 20(Suppl. 1):301. doi: 10.1186/s12870-020-02499-w

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Garrison, E., and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv* [Preprint] 1207.3907v2[q–bio.GN]. arXiv:1207.3907v2 [q-bio.GN]

Gaudet, M., Jorge, V., Paolucci, I., Beritognolo, I., Mugnozza, G. S., and Sabatti, M. (2008). Genetic linkage maps of *Populus nigra* L. including AFLPs, SSRs, SNPs, and sex trait. *Tree Genet. Genomes* 4, 25–36. doi: 10.1007/s11295-007-0085-1

Geraldes, A., Hefer, C. A., Capron, A., Kolosova, N., Martinez-Nunez, F., Soolanayakanahally, R. Y., et al. (2015). Recent Y chromosome divergence despite ancient origin of dioecy in poplars (*Populus*). *Mol. Ecol.* 24, 3243–3256. doi: 10.1111/mec.13126

Han, Q., Song, H., Yang, Y., Jiang, H., and Zhang, S. (2018). Transcriptional profiling reveals mechanisms of sexually dimorphic responses of *Populus cathayana* to potassium deficiency. *Physiol. Plant.* 162, 301–315. doi: 10.1111/ppl.12636

Hofmeister, B. T., Denkena, J., Colome-Tatche, M., Shahryary, Y., Hazarika, R., Grimwood, J., et al. (2020). A genome assembly and the somatic genetic and epigenetic mutation rate in a wild long-lived perennial *Populus trichocarpa*. *Genome Biol.* 21:259. doi: 10.1186/s13059-020-02162-5

Jansson, S., Bhalerao, R. P., and Groover, A. T. (2010). *Genetics and Genomics of Populus*. New York, NY: Springer-Verlag.

Jansson, S., and Douglas, C. J. (2007). *Populus*: a model system for plant biology. *Annu. Rev. Plant Biol.* 58, 435–458. doi: 10.1146/annurev.arplant.58.032806.103956

Jiang, D., Feng, J., Dong, M., Wu, G., Mao, K., and Liu, J. (2016). Genetic origin and composition of a natural hybrid poplar Populus x jrtyschensis from two distantly related species. *BMC Plant Biol.* 16:89. doi: 10.1186/s12870-016-0776-6

Jiang, H., Peng, S., Zhang, S., Li, X., Korpelainen, H., and Li, C. (2012). Transcriptional profiling analysis in *Populus yunnanensis* provides insights into molecular mechanisms of sexual differences in salinity tolerance. *J. Exp. Bot.* 63, 3709–3726. doi: 10.1093/jxb/ers064

Kafer, J., Marais, G. A., and Pannell, J. R. (2017). On the rarity of dioecy in flowering plants. *Mol. Ecol.* 26, 1225–1241. doi: 10.1111/mec.14020

Kersten, B., Pakull, B., Groppe, K., Lueneburg, J., and Fladung, M. (2014). The sex-linked region in *Populus tremuloides* Turesson 141 corresponds to a pericentromeric region of about two million base pairs on *P. trichocarpa* chromosome 19. *Plant Biol.* 16, 411–418. doi: 10.1111/plb.12048

Kostina, M. V., and Nasimovich, J. A. (2014). On the systematics of *Populus* L. II. Importance of fruit characters for identification of cultivated and adventive species in Moscow region. *Bull. Mosc. Soc. Nat. Biol. Ser.* 119, 74–79.

Kostina, M. V., Puzyryov, A. N., Nasimovich, J. A., and Parshevnikova, M. S. (2017). Representatives of the sections *Aigeiros* Duby and *Tacamahaca* Spach (genus *Populus* L., *Salicaceae*) and their hybrids in cities of central and eastern European Russia. *Skvortsovia* 3, 97–119.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [Preprint] 1303.3997v2[q–bio.GN]. arXiv:1303.3997v2 [q-bio.GN]

Liao, Q., Zhou, T., Yao, J. Y., Han, Q. F., Song, H. X., Guan, C. Y., et al. (2018). Genome-scale characterization of the vacuole nitrate transporter Chloride Channel (CLC) genes and their transcriptional responses to diverse nutrient stresses in allotetraploid rapeseed. *PLoS One* 13:e0208648. doi: 10.1371/journal.pone.0208648

Liao, Y., Smyth, G. K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi: 10.1093/bioinformatics/btt656

Liu, C., Zhao, Y., Zhao, X., Dong, J., and Yuan, Z. (2020). Genome-wide identification and expression analysis of the *CLC* gene family in pomegranate (*Punica granatum*) reveals its roles in salt resistance. *BMC Plant Biol.* 20:560. doi: 10.1186/s12870-020-02771-z

Mayorov, S. R., Bochkin, V. D., Nasimovich, J. A., and Shcherbakov, A. V. (2012). "Family *Salicaceae*," in *Adventive Flora of Moscow and Moscow Region*, ed. M. S. Ignatov (Moscow: KMK Scientific Press Ltd.), 78–108.

McKown, A. D., Klapste, J., Guy, R. D., Soolanayakanahally, R. Y., La Mantia, J., Porth, I., et al. (2017). Sexual homomorphism in dioecious trees: extensive tests fail to detect sexual dimorphism in *Populus* (dagger). *Sci. Rep.* 7:1831. doi: 10.1038/s41598-017-01893-z

Melnikova, N. V., Borkhert, E. V., Snezhkina, A. V., Kudryavtseva, A. V., and Dmitriev, A. A. (2017). Sex-specific response to stress in *Populus*. *Front. Plant Sci.* 8:1827. doi: 10.3389/fpls.2017.01827

Melnikova, N. V., Kudryavtseva, A. V., Borkhert, E. V., Pushkova, E. N., Fedorova, M. S., Snezhkina, A. V., et al. (2019). Sex-specific polymorphism of *MET1* and *ARR17* genes in *Populus* x *sibirica*. *Biochimie* 162, 26–32. doi: 10.1016/j.biochi.2019.03.018

Müller, N. A., Kersten, B., Leite Montalvão, A. P., Mähler, N., Bernhardsson, C., Bräutigam, K., et al. (2020). A single gene underlies the dynamic evolution of poplar sex determination. *Nat. Plants* 6, 630–637. doi: 10.1038/s41477-020-0672-9

Novakovskiy, R. O., Dvorianinova, E. M., Rozhmina, T. A., Kudryavtseva, L. P., Gryzunov, A. A., Pushkova, E. N., et al. (2020). Data on genetic polymorphism of flax (*Linum usitatissimum* L.) pathogenic fungi of *Fusarium*, *Colletotrichum*, *Aureobasidium*, *Septoria*, and *Melampsora* genera. *Data Brief* 31:105710. doi: 10.1016/j.dib.2020.105710

Pakull, B., Groppe, K., Mecucci, F., Gaudet, M., Sabatti, M., and Fladung, M. (2011). Genetic mapping of linkage group XIX and identification of sex-linked SSR markers in a *Populus tremula* × *Populus tremuloides* cross. *Can. J. For. Res.* 41, 245–253. doi: 10.1139/X10-206

Pakull, B., Groppe, K., Meyer, M., Markussen, T., and Fladung, M. (2009). Genetic linkage mapping in aspen (*Populus tremula* L. and *Populus tremuloides* Michx.). *Tree Genet. Genomes* 5, 505–515. doi: 10.1007/s11295-009-0204-2

Pakull, B., Kersten, B., Luneburg, J., and Fladung, M. (2015). A simple PCR-based marker to determine sex in aspen. *Plant Biol.* 17, 256–261. doi: 10.1111/plb.12217

Paolucci, I., Gaudet, M., Jorge, V., Beritognolo, I., Terzoli, S., Kuzminsky, E., et al. (2010). Genetic linkage maps of *Populus alba* L. and comparative mapping analysis of sex determination across *Populus* species. *Tree Genet. Genomes* 6, 863–875. doi: 10.1007/s11295-010-0297-7

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

Rae, A. M., Street, N. R., and Rodríguez-Acosta, M. (2007). "*Populus* trees," in *Forest Trees*, ed. C. Kole (Berlin: Springer Berlin Heidelberg), 1–28.

Renner, S. S. (2014). The relative and absolute frequencies of angiosperm sexual systems: dioecy, monoecy, gynodioecy, and an updated online database. *Am. J. Bot.* 101, 1588–1596. doi: 10.3732/ajb.1400196

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

Roe, A. D., MacQuarrie, C. J., Gros-Louis, M. C., Simpson, J. D., Lamarche, J., Beardmore, T., et al. (2014). Fitness dynamics within a poplar hybrid zone:

II. Impact of exotic sex on native poplars in an urban jungle. *Ecol. Evol.* 4, 1876–1889. doi: 10.1002/ece3.1028

Sanderson, B. J., Wang, L., Tiffin, P., Wu, Z., and Olson, M. S. (2019). Sex-biased gene expression in flowers, but not leaves, reveals secondary sexual dimorphism in *Populus balsamifera*. *New Phytol.* 221, 527–539. doi: 10.1111/nph.15421

Song, H., Cai, Z., Liao, J., Tang, D., and Zhang, S. (2019). Sexually differential gene expressions in poplar roots in response to nitrogen deficiency. *Tree Physiol.* 39, 1614–1629. doi: 10.1093/treephys/tpz057

Spigler, R. B., and Ashman, T. L. (2012). Gynodioecy to dioecy: are we there yet? *Ann. Bot.* 109, 531–543. doi: 10.1093/aob/mcr170

Subba, A., Tomar, S., Pareek, A., and Singla-Pareek, S. L. (2020). The chloride channels: silently serving the plants. *Physiol. Plant*. 171, 688–702. doi: 10.1111/ppl.13240

Sun, W., Deng, D., Yang, L., Zheng, X., Yu, J., Pan, H., et al. (2013). Overexpression of the chloride channel gene (GmCLC1) from soybean increases salt tolerance in transgenic *Populus deltoides* x *P. euramericana* 'Nanlin895'. *Plant OMICS* 6, 347–354.

Tanurdzic, M., and Banks, J. A. (2004). Sex-determining mechanisms in land plants. *Plant Cell* 16(Suppl. 1), S61–S71. doi: 10.1105/tpc.016667

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Xue, L., Wu, H., Chen, Y., Li, X., Hou, J., Lu, J., et al. (2020). Evidences for a role of two Y-specific genes in sex determination in *Populus deltoides*. *Nat. Commun.* 11:5893. doi: 10.1038/s41467-020-19559-2

Yampolsky, C., and Yampolsky, H. (1922). *Distribution of Sex Forms in the Phanerogamic Flora*. Leipzig: Gebrüder Borntraeger.

Yang, W., Wang, D., Li, Y., Zhang, Z., Tong, S., Li, M., et al. (2020). A general model to explain repeated turnovers of sex determination in the *Salicaceae*. *Mol. Biol. Evol.* 38, 968–980. doi: 10.1093/molbev/msaa261

Yin, T., Difazio, S. P., Gunter, L. E., Zhang, X., Sewell, M. M., Woolbright, S. A., et al. (2008). Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Res.* 18, 422–430. doi: 10.1101/gr.7076308

Zhang, J., Feng, J., Lu, J., Yang, Y., Zhang, X., Wan, D., et al. (2014). Transcriptome differences between two sister desert poplar species under salt stress. *BMC Genomics* 15:337. doi: 10.1186/1471-2164-15-337

Zhou, R., Macaya-Sanz, D., Schmutz, J., Jenkins, J. W., Tuskan, G. A., and DiFazio, S. P. (2020). Sequencing and analysis of the sex determination region of *Populus trichocarpa*. *Genes* 11:843. doi: 10.3390/genes11080843

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.

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