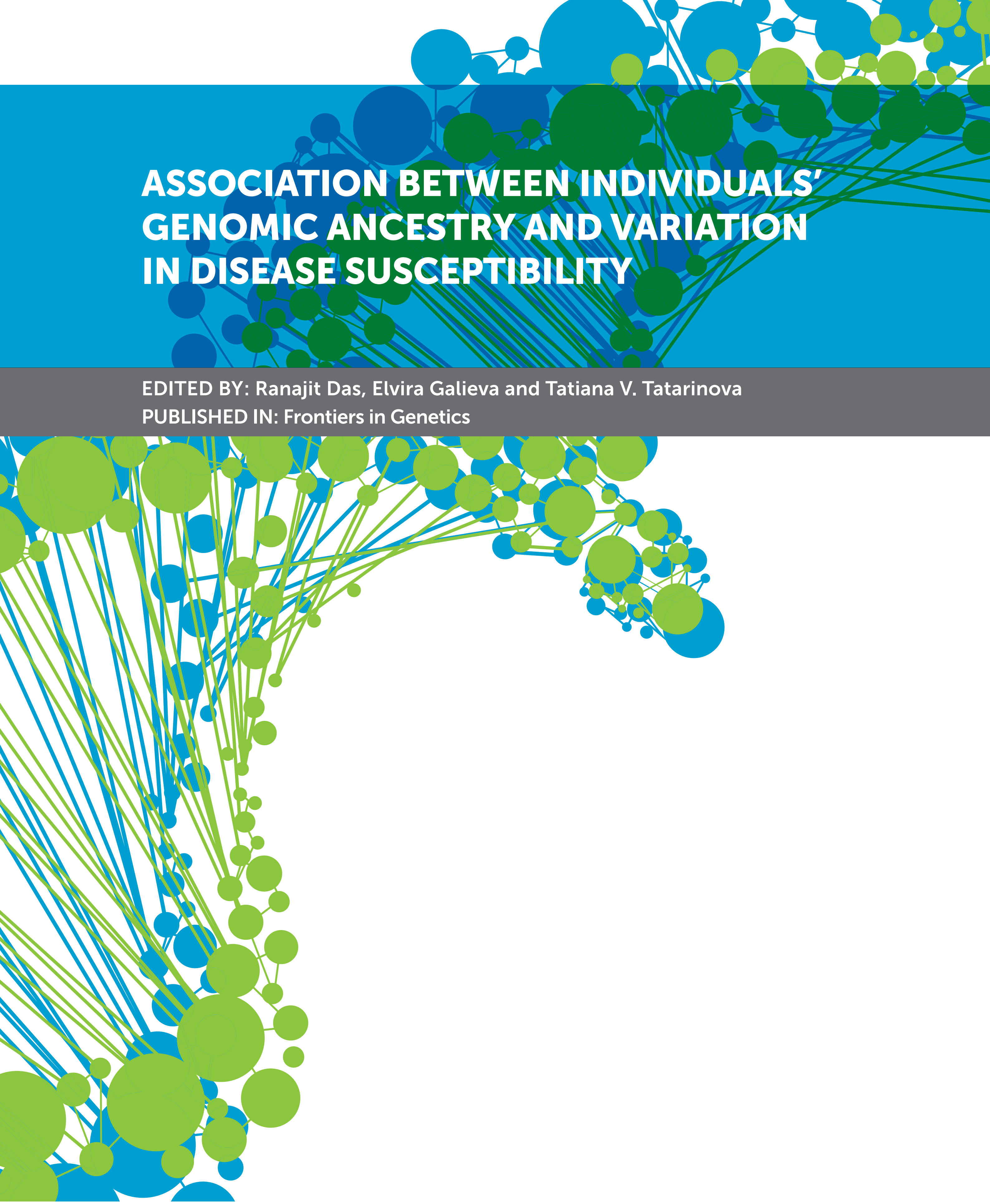


ASSOCIATION BETWEEN INDIVIDUALS' GENOMIC ANCESTRY AND VARIATION IN DISEASE SUSCEPTIBILITY

EDITED BY: Ranajit Das, Elvira Galieva and Tatiana V. Tatarinova
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-571-5
DOI 10.3389/978-2-88974-571-5

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

ASSOCIATION BETWEEN INDIVIDUALS' GENOMIC ANCESTRY AND VARIATION IN DISEASE SUSCEPTIBILITY

Topic Editors:
Ranajit Das, Yenepoya University, India
Elvira Galieva, Novosibirsk State University, Russia
Tatiana V. Tatarinova, University of La Verne, United States

Topic Editor Ranajit Das is the Founder Director of Genome Mapster and Infygene Genomic Healthcare. Topic Editor Tatiana Tatarinova holds patents related to the Research Topic subject. All other Topic Editors declare no competing interests with regard to the Research Topic subject.

Citation: Das, R., Galieva, E., Tatarinova, T. V., eds. (2022). Association Between Individuals' Genomic Ancestry and Variation in Disease Susceptibility. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-571-5

Table of Contents

05 ***Editorial: Association Between Individuals' Genomic Ancestry and Variation in Disease Susceptibility***
Ranajit Das, Tatiana V. Tatarinova, Elvira R. Galieva and Yuriy L. Orlov

08 ***Delineation of Mitochondrial DNA Variants From Exome Sequencing Data and Association of Haplogroups With Obesity in Kuwait***
Mohammed Dashti, Hussain Alsaleh, Muthukrishnan Eaaswarkhanth, Sumi Elsa John, Rasheeba Nizam, Motasem Melhem, Prashantha Hebbar, Prem Sharma, Fahd Al-Mulla and Thangavel Alphonse Thanaraj

20 ***Evolutionary Changes in the Interaction of miRNA With mRNA of Candidate Genes for Parkinson's Disease***
Saltanat Kamenova, Assel Aralbayeva, Aida Kondybayeva, Aigul Akimniyazova, Anna Pyrkova and Anatoliy Ivashchenko

38 ***RHIVDB: A Freely Accessible Database of HIV Amino Acid Sequences and Clinical Data of Infected Patients***
Olga Tarasova, Anastasia Rudik, Dmitry Kireev and Vladimir Poroikov

45 ***Ethnic Differences in the Frequency of* CFTR *Gene Mutations in Populations of the European and North Caucasian Part of the Russian Federation***
Nika Petrova, Natalia Balinova, Andrey Marakhonov, Tatyana Vasilyeva, Nataliya Kashirskaya, Varvara Galkina, Evgeniy Ginter, Sergey Kutsev and Rena Zinchenko

54 ***Heterogeneity of Genetic Admixture Determines SLE Susceptibility in Mexican***
Susana Hernández-Doño, Juan Jakez-Ocampo, José Eduardo Márquez-García, Daniela Ruiz, Víctor Acuña-Alonzo, Guadalupe Lima, Luis Llorente, Víctor Hugo Tovar-Méndez, Rafael García-Silva, Julio Granados, Joaquín Zúñiga and Gilberto Vargas-Alarcón

74 ***Genetic Susceptibility to Fungal Infections and Links to Human Ancestry***
Bharati Naik, Sumayyah M. Q. Ahmed, Suparna Laha and Shankar Prasad Das

86 ***Epidemiology of Rare Hereditary Diseases in the European Part of Russia: Point and Cumulative Prevalence***
Rena A. Zinchenko, Eugeny K. Ginter, Andrey V. Marakhonov, Nika V. Petrova, Vitaly V. Kadyshev, Tatyana P. Vasilyeva, Oksana U. Alexandrova, Alexander V. Polyakov and Sergey I. Kutsev

98 ***A Role of Variance in Interferon Genes to Disease Severity in COVID-19 Patients***
Leonid Gozman, Kellie Perry, Dimitri Nikogosov, Ilya Klabukov, Artem Shevlyakov and Ancha Baranova

108 Targeted Sequencing of 242 Clinically Important Genes in the Russian Population From the Ivanovo Region

Vasily E. Ramensky, Alexandra I. Ershova, Marija Zaicenoka, Anna V. Kiseleva, Anastasia A. Zharikova, Yuri V. Vyatkin, Evgeniia A. Sotnikova, Irina A. Efimova, Mikhail G. Divashuk, Olga V. Kurilova, Olga P. Skirko, Galina A. Muromtseva, Olga A. Belova, Svetlana A. Rachkova, Maria S. Pokrovskaya, Svetlana A. Shalnova, Alexey N. Meshkov and Oxana M. Drapkina

121 Local Ancestry Adjusted Allelic Association Analysis Robustly Captures Tuberculosis Susceptibility Loci

Yolandi Swart, Caitlin Uren, Paul D. van Helden, Eileen G. Hoal and Marlo Möller

133 Identification of Bovine miRNAs with the Potential to Affect Human Gene Expression

Moldir Myrzabekova, Siegfried Labeit, Raigul Niyazova, Aigul Akimniyazova and Anatoliy Ivashchenko

frontiers in Genetics

EDITORIAL
published: 02 February 2022
doi: 10.3389/fgene.2022.831320

Editorial: Association Between Individuals' Genomic Ancestry and Variation in Disease Susceptibility

Ranajit Das[1]*, Tatiana V. Tatarinova[2,3], Elvira R. Galieva[4,5] and Yuriy L. Orlov[4,5,6,7]

[1]Yenepoya Research Centre, Yenepoya University, Mangalore, India, [2]Natural Science Division, La Verne University, La Verne, CA, United States, [3]Department of Fundamental Biology and Biotechnology, Siberian Federal University, Krasnoyarsk, Russia, [4]Life Sciences Department, Novosibirsk State University, Novosibirsk, Russia, [5]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, [6]Agrarian and Technological Institute, Peoples' Friendship University of Russia (RUDN University), Moscow, Russia, [7]Institute of Digital Medicine, I.M.Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia

Keywords: genetic ancestry and disease susceptibility, ancestry and genetic tapestry, ancestry-specific medicine, genomic profiling, genomic healthcare, personalized therapeutics, genome wide association study (GWAS)

Editorial for the Research Topic

Association between Individuals' Genomic Ancestry and Variation in Disease Susceptibility

This *Association between Individuals' Genomic Ancestry and Variation in Disease Susceptibility* issue presents the studies related to the fields of human genetics, population genetics, and genetic and ancestral affiliations of various diseases. Our DNA determines not only who we are but also holds the key to uncovering our true ancestral past. A plethora of information stored inside the genome reflects our uniqueness and proximity to different ancestral and modern-day populations. Given the high correspondence between our genetic make-up and the geographical origin of our forefathers, it is possible to glean into precise ancestral origin using the genetic information. Understanding one's ancestry is not only a 'homing' tool bringing someone closer to human evolutionary past but also holds the key in determining population-specific variability in disease susceptibility, drug responsiveness, and other health and fitness related traits.

Accordingly, we organized this Research Topic to collect the papers focused on studying ancestry specific variation in various human diseases. This Topic complements recent Research Topic *Bioinformatics of Genome Regulation* in *Frontiers in Genetics* considered more genetic background rather than molecular mechanisms of the human diseases (Orlov et al., 2021a).

Understanding one's ancestry can play a monumental role in understanding variation in disease susceptibility across various populations and glean into the complex gene and environment interplay in ancestry-specific disorders. For instance, cardiovascular diseases tend to manifest in distinct ways unique to the ancestry of the patient such that people with high African ancestry proportion tend to have strokes as a result of cardiovascular disease, while people of South Asian ancestry tend to have heart attacks (Harshfield et al., 2021). Traditional high fat and protein diets in cold regions of Siberia and North Asia have consequences on obesity and diabetes-related diseases (Bai et al., 2015; Tiis et al., 2020). Individuals with larger fractions of Western Hunter Gatherer (WHG) related ancestry has been shown to develop more severe COVID-19 symptoms (Upadhyai et al., 2021).

In the realms of human population genetics, we strived to address similar questions in a global context, in an attempt to expand our existing knowledge to better understand the association between individuals' genetic ancestry and disease susceptibility. This understanding can facilitate the development of novel therapeutics or repurposing of existing treatment strategies, particularly aiding in identifying population-specific therapies. Subsequently, our knowledge of ancestry-specific

OPEN ACCESS

Edited and reviewed by:
Marc Via,
University of Barcelona, Spain

****Correspondence:***
Ranajit Das
neel_291@yahoo.co.in

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *08 December 2021*
Accepted: *14 January 2022*
Published: *02 February 2022*

Citation:
Das R, Tatarinova TV, Galieva ER and Orlov YL (2022) Editorial: Association Between Individuals' Genomic Ancestry and Variation in Disease Susceptibility. Front. Genet. 13:831320. doi: 10.3389/fgene.2022.831320

variation in complex disorders can be used towards developing personalized and ancestry-specific precision medicine approaches to ameliorate several complex disorders.

We considered the following themes for this special issue:

- Unravel predisposition of various modern-day populations towards common disorders and conditions, including but not limited to cancers, heart diseases, and infectious diseases.
- The association of alleles with complex disorders, evaluated at a population level; discovery of novel disease marker panels.
- Identification of novel medically relevant genetic variants that can be used as diagnostic markers in genetic diagnostics and healthcare.
- Selection dynamics of various genes. Investigation of the spatial and temporal distributions of positively selected alleles in response to population specific disease susceptibility.

The papers published in this Research Topic correspond to the themes stated above and extend the studies presented in *Frontiers in Genetics* Topics (see https://www.frontiersin.org/research-topics/17947/bioinformatics-of-genome-regulation-volume-ii). Recently we had organized series of conferences on human population genetics and computational genomics in Russia that allowed formalize the idea of special journal issues. The conference "Century of Human Population Genetics" was held in Moscow in 2019 (Tatarinova et al., 2020a). The conference was focused on the discussion of the research on gene pools of the world's nations, ancient DNA analysis, possibilities of judicial genetics, population-genetic database development, biobanks, and new genomics technologies (Tatarinova et al., 2020b). We had series of publications on human ancestry based on new genomics data initially discussed at this meeting (Das et al., 2020; Orlov et al., 2021b). The BGRS\SB (Bioinformatics of Genome Regulation and Structure\Systems Biology) multi-conference (https://bgrssb.icgbio.ru/2020) has been organized in Novosibirsk, Russia in 2020 and was associated with the Research Topic on gene expression regulation in *Frontiers in Genetics* (see also https://www.frontiersin.org/research-topics/8383/bioinformatics-of-genome-regulation-and-systems-biology) (Orlov et al., 2016a, Orlov et al., 2016b, Tatarinova et al., 2019; Orlov et al., 2021a).

In this Research Topic a total of 11 papers could be arranged by two main areas - the human population genetics and ancestry studies, and the works on molecular mechanisms of the diseases.

Among the ancestry-based studies, Dashti et al. studied association of mtDNA haplogroups with obesity using high throughput sequencing technologies. Previous studies indicated that certain mtDNA variants and haplogroup lineages were associated with obesity. Dashti et al. presents the first study that used whole-exome data to extract entire mitochondrial haplogroups and consecutively study its association with obesity in an Arab population.

Susana Hernández-Doño et al. studied genetic determinants of systemic lupus erythematosus in Mexican population based on Human Leukocyte Antigen (HLA) haplotypes. Consistent with the admixture estimations, the origin of all risk alleles and haplotypes found in this study were found to be European, while the protection alleles were found to be Mexican Native American. Petrova et al. studied human genome variation in CFTR gene related to cystic fibrosis in the European and North Caucasian Part of Russia. The widespread introduction of technologies of whole genome and whole-exome analysis into practical healthcare has significantly increased the number of genetically determined diseases in the structure of human morbidity. Zinchenko et al. (2021) discussed the point (PP) and cumulative prevalence (PP), as well as the burden of the most common rare hereditary diseases (RHDs) (autosomal dominant, autosomal recessive, and X-linked) among 14 remote populations of the European part of Russia. Ramensky et al. studied genetic predisposition to cardiovascular diseases. The authors performed a targeted sequencing of 242 clinically important genes mostly associated with cardiovascular diseases.

The second set of papers contains the research articles, database application and the reviews highlighting genetic background and molecular mechanisms of the diseases. Similar studies have been discussed in previous issues on the topic (Snezhkina et al., 2020).

Kamenova et al. studied interactions of miRNA with mRNA in human genes associated with neurodegenerative diseases. Parkinson's disease has complex genetic background that challenges bioinformatics methods for analysis of genes' network and search for target genes (Orlov et al., 2021c). Recent studies have established a correlation between the disease and miRNA expression (Mukushkina et al., 2020). The authors described quantitative characteristics of the interactions between miRNAs and the mRNAs of candidate Parkinson's disease genes. Myrzabekova et al. have studied potential effects of exogenous miRNA to human gene expression. The authors described potential human gene targets for such miRNA.

Tarasova et al. presented bioinformatics database on Human immunodeficiency virus (HIV) to study associations of clinical data of infected patients and viral sequences. Resistance of HIV to current drugs raises problems of treatment of this complex disease. The authors developed the RHIVDB database with free access that could help in drug target search.

Swart et al. studied genetic factors for developing active form of pulmonary *tuberculosis*, caused by *Mycobacterium tuberculosis*. The authors found *tuberculosis* susceptibility loci using local ancestry adjusted allelic association analysis. Naik et al. have reviewed genetic susceptibility to fungal infections in humans. Here, the authors discussed the polymorphisms in the genes of the immune system, the way it contributes toward some common fungal infections.

Gozman et al. highlighted the putative role of patients' genetic background in response to the exposure to coronavirus. Hosts' resistance to COVID-19 is thought to be related to the immune system. However, the genes are yet to be revealed. The variation in interferon genes to disease severity was discussed in this article. Gozman et al. raised important actual problem of the role of genetic variance in disease severity in COVID-19. Further, Prof Balanovsky et al. (2021) recently studied the variation of individual genomes associated with the severe COVID-19. This is a burning topic and the problem is continuing to be actively discussed (Upadhyai et al., 2021; Gerasimov et al., 2021; Lu et al., 2021) in relation to new data.

Overall, we are proud of the Research Topic at *Frontiers in Genetics* we collated. We hope that you will find this paper collection a stimulating reading and consider coming to the next conferences in this area in life format next year (https://bgrssb.icgbio.ru/2022/). The complementary Research Topics in *Frontiers* (https://www.frontiersin.org/research-topics/21036/high-throughput-sequencing-based-investigation-of-chronic-disease-markers-and-mechanisms) continues collection of papers on human diseases' markers.

AUTHOR CONTRIBUTIONS

RD, TT, and EG organized the Research Topic as guest editors, supervised the reviewing of the manuscripts, YO critically contributed both to the extension of the Topic and the reviewing process. All the authors wrote this Editorial paper. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The guest editors are grateful to all the authors contributing to this special issue papers collection and thank all the reviewers who helped improve the manuscripts. The population genetics studies were supported by the Russian Science Foundation (grant 19-15-00219, for EG). The publication has been prepared with the support of the RUDN University Strategic Academic Leadership Program (recipient YO).

REFERENCES

Bai, H., Liu, H., Suyalatu, S., Guo, X., Chu, S., Chen, Y., et al. (20152015). Association Analysis of Genetic Variants with Type 2 Diabetes in a Mongolian Population in China. *J. Diabetes Res.* 2015, 1–7. doi:10.1155/2015/613236

Balanovsky, O., Petrushenko, V., Mirzaev, K., Abdullaev, S., Gorin, I., Chernevskiy, D., et al. (2021). Variation of Genomic Sites Associated with Severe Covid-19 across Populations: Global and National Patterns. *Pgpm* 14, 1391–1402. doi:10.2147/PGPM.S320609

Das, R., Ivanisenko, V. A., Anashkina, A. A., and Upadhyai, P. (2020). The story of the Lost Twins: Decoding the Genetic Identities of the Kumhar and Kurcha Populations from the Indian Subcontinent. *BMC Genet.* 21 (Suppl. 1), 117. doi:10.1186/s12863-020-00919-2

Gerasimov, A., Galkina, E., Danilova, E., Ikonnikova, I., Novoselova, T., Orlov, Y. L., et al. (2021). Estimation of the Probability of Daily Fluctuations of Incidence of COVID-19 According to Official Data. *PeerJ* 9, e11049. doi:10.7717/peerj.11049

Harshfield, E. L., Fauman, E. B., Stacey, D., Paul, D. S., Ziemek, D., Ong, R. M. Y., et al. (2021). Genome-wide Analysis of Blood Lipid Metabolites in over 5000 South Asians Reveals Biological Insights at Cardiometabolic Disease Loci. *BMC Med.* 19 (1), 232. doi:10.1186/s12916-021-02087-1

Lu, H., Chen, M., Tang, S., and Yu, W. (2021). Association of Coagulation Disturbances with Severity of COVID-19: a Longitudinal Study. *Hematology* 26 (1), 656–662. doi:10.1080/16078454.2021.1968648

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

Myrzabekova, M., Labeit, S., Niyazova, R., Ivashchenko, A. A., and Akimniyazova, A. (2021). Identification of Bovine miRNAs with the Potential to Affect Human Gene Expression. *Front. Genet.* 12. doi:10.3389/fgene.2021.705350

Orlov, Y. L., Anashkina, A. A., Klimontov, V. V., and Baranova, A. V. (2021b). Medical Genetics, Genomics and Bioinformatics Aid in Understanding Molecular Mechanisms of Human Diseases. *Ijms* 22, 9962. doi:10.3390/ijms22189962

Orlov, Y. L., Anashkina, A. A., Tatarinova, T. V., and Baranova, A. V. (2021a). Editorial: Bioinformatics of Genome Regulation, Volume II. *Front. Genet.* 12, 795257. doi:10.3389/fgene.2021.795257

Orlov, Y. L., Baranova, A. V., Hofestädt, R., and Kolchanov, N. A. (2016b). Computational Genomics at BGRS\SB-2016: Introductory Note. *BMC genomics* 17 (Suppl. 14), 996. doi:10.1186/s12864-016-3350-6

Orlov, Y. L., Baranova, A. V., and Markel, A. L. (2016a). Computational Models in Genetics at BGRS\SB-2016: Introductory Note. *BMC Genet.* 17 (Suppl. 3), 155. doi:10.1186/s12863-016-0465-3

Orlov, Y. L., Galieva, A. G., Orlova, N. G., Ivanova, E. N., Mozyleva, Y. A., and Anashkina, A. A. (2021c). Reconstruction of Gene Network Associated with Parkinson Disease for Gene Targets Search. *Biomed. Khim* 67 (3), 222–230. Russian. doi:10.18097/PBMC20216703222

Snezhkina, A. V., Fedorova, M. S., Pavlov, V. S., Kalinin, D. V., Golovyuk, A. L., Pudova, E. A., et al. (2020). Mutation Frequency in Main Susceptibility Genes Among Patients with Head and Neck Paragangliomas. *Front. Genet.* 11, 614908. doi:10.3389/fgene.2020.614908

Tatarinova, T. V., Baranova, A. V., Anashkina, A. A., and Orlov, Y. L. (2020a). 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

Tatarinova, T. V., Tabikhanova, L. E., Eslami, G., Bai, H., and Orlov, Y. L. (2020b). Genetics Research at the "Centenary of Human Population Genetics" Conference and SBB-2019. *BMC Genet.* 21 (Suppl. 1), 109. doi:10.1186/s12863-020-00906-7

Tiis, R. P., Osipova, L. P., Lichman, D. V., Voronina, E. N., and Filipenko, M. L. (2020). Studying Polymorphic Variants of the NAT2 Gene (NAT2*5 and NAT2*7) in Nenets Populations of Northern Siberia. *BMC Genet.* 21 (Suppl. 1), 115. doi:10.1186/s12863-020-00909-4

Upadhyai, P., Suresh, G., Parit, R., and Das, R. (2021). Genomic and Ancestral Variation Underlies the Severity of COVID-19 Clinical Manifestation in Individuals of European Descent. *Life* 11, 921. doi:10.3390/life11090921

Zinchenko, R. A., Ginter, E. K., Marakhonov, A. V., PetrovaKadyshev, N. V. V. V., Kadyshev, V. V., Vasilyeva, T. P., et al. (2021). Epidemiology of Rare Hereditary Diseases in the European Part of Russia: Point and Cumulative Prevalence. *Front. Genet.* 12. doi:10.3389/fgene.2021.678957

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

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 11 February 2021
doi: 10.3389/fgene.2021.626260

Delineation of Mitochondrial DNA Variants From Exome Sequencing Data and Association of Haplogroups With Obesity in Kuwait

Mohammed Dashti[1]*, *Hussain Alsaleh*[2], *Muthukrishnan Eaaswarkhanth*[1], *Sumi Elsa John*[1], *Rasheeba Nizam*[1], *Motasem Melhem*[1], *Prashantha Hebbar*[1], *Prem Sharma*[3], *Fahd Al-Mulla*[1]* and *Thangavel Alphonse Thanaraj*[1]*

[1] Genetics and Bioinformatics Department, Dasman Diabetes Institute, Kuwait City, Kuwait, [2] Kuwait Identification DNA Laboratory, General Department of Criminal Evidence, Ministry of Interior, Kuwait City, Kuwait, [3] Department Special Services Facilities, Dasman Diabetes Institute, Kuwait City, Kuwait

OPEN ACCESS

Edited by:
Tatiana V. Tatarinova,
University of La Verne, United States

Reviewed by:
Veronica Fernandes,
University of Porto, Portugal
Shaoqing Wen,
Fudan University, China

****Correspondence:***
Fahd Al-Mulla
fahd.almulla@dasmaninstitute.org
Thangavel Alphonse Thanaraj
alphonse.thangavel@dasmaninstitute.org
Mohammed Dashti
mdashti@outlook.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *05 November 2020*
Accepted: *13 January 2021*
Published: *11 February 2021*

Citation:
Dashti M, Alsaleh H, Eaaswarkhanth M, John SE, Nizam R, Melhem M, Hebbar P, Sharma P, Al-Mulla F and Thanaraj TA (2021) Delineation of Mitochondrial DNA Variants From Exome Sequencing Data and Association of Haplogroups With Obesity in Kuwait. Front. Genet. 12:626260. doi: 10.3389/fgene.2021.626260

Background/Objectives: Whole-exome sequencing is a valuable tool to determine genetic variations that are associated with rare and common health conditions. A limited number of studies demonstrated that mitochondrial DNA can be captured using whole-exome sequencing. Previous studies have suggested that mitochondrial DNA variants and haplogroup lineages are associated with obesity. Therefore, we investigated the role of mitochondrial variants and haplogroups contributing to the risk of obesity in Arabs in Kuwait using exome sequencing data.

Subjects/Methods: Indirect mitochondrial genomes were extracted from exome sequencing data from 288 unrelated native Arab individuals from Kuwait. The cohort was divided into obese [body mass index (BMI) $\geq$ 30 kg/m^2] and non-obese (BMI $<$ 30 kg/m^2) groups. Mitochondrial variants were identified, and haplogroups were classified and compared with other sequencing technologies. Statistical analysis was performed to determine associations and identify mitochondrial variants and haplogroups affecting obesity.

Results: Haplogroup R showed a protective effect on obesity [odds ratio (OR) = 0.311; *P* = 0.006], whereas haplogroup L individuals were at high risk of obesity (OR = 2.285; *P* = 0.046). Significant differences in mitochondrial variants between the obese and non-obese groups were mainly haplogroup-defining mutations and were involved in processes in energy generation. The majority of mitochondrial variants and haplogroups extracted from exome were in agreement with technical replica from Sanger and whole-genome sequencing.

Conclusions: This is the first to utilize whole-exome data to extract entire mitochondrial haplogroups to study its association with obesity in an Arab population.

Keywords: mitochondrial, DNA, haplogroup, exome, obesity

INTRODUCTION

Mitochondria play a role in generating cellular energy via oxidative phosphorylation (OXPHOS), reactive oxygen species production, and apoptosis. Human mitochondrial DNA (mtDNA) is circular, double-stranded, and 16,569 base pairs (bp) in size and contains 37 genes that code for 22 transfer RNAs, two ribosomal RNAs that are necessary for protein synthesis, and 13 messenger RNAs that are required for OXPHOS (Anderson et al., 1981; Andrews et al., 1999). Each mitochondrion contains several copies of mtDNA, and each cell contains many mitochondria (Hosgood et al., 2010). mtDNA contains a major non-coding region called the control/D-loop region, which regulates mitochondrial transcription and replication. The mitochondrial control region is located at mitochondrial nucleotide positions 16,024–576 and is susceptible to a high rate of mtDNA alterations, particularly at the hypervariable regions (Greenberg et al., 1983) as well as under conditions of increased oxidative stress (Clayton, 2000). mtDNA variants are maternally inherited without recombination and may accumulate over time. A mitochondrial haplogroup is a group of individuals who share the same accumulated mtDNA variants and can be geographically restricted, making then traceable via maternal linage. Different haplogroups form diverse branches of a mitochondrial phylogenetic tree. The sub-Saharan Africans are characterized by L0–L6; the South Asians by haplogroups R5–R8, M2–M6, and M4–67; the Europeans, Southwest Asians, and North Africans by haplogroups U, HV, JT, N1, N2, and X; and the East Asians by haplogroups A–G, Z, and M7–M9 (Loogvali et al., 2004; Chaubey et al., 2007; Soares et al., 2010; Kivisild, 2015).

Sanger sequencing is considered the gold standard for detecting mtDNA variants. This approach has progressed to next-generation sequencing (NGS) platform, as it provides high-throughput sequence data for large cohort studies and is less labor-intensive and time-consuming than Sanger sequencing (Calvo et al., 2012; Chinnery et al., 2012; Tang et al., 2013; Wong, 2013). Recently, a number of studies demonstrated that whole-genome sequencing and off-target exome sequencing are able to target both nuclear DNA and mtDNA for the diagnosis of monogenic cases and association studies for multifactorial disorders (Picardi and Pesole, 2012; Delmiro et al., 2013; Samuels et al., 2013; Griffin et al., 2014; Li et al., 2014). Particularly interesting studies include the following: Wagner et al. (2019) evaluated if mtDNA analysis can be performed using exome data; Diroma et al. (2014) extracted mtDNA sequences from exome data to reconstruct human population history using mtDNA variant as marker and to illustrate the involvement of mtDNA in pathology; and Patowary et al. (2017) analyzed the mtDNA sequence derived from whole-exome sequencing and studied haplogroup and variant association in autism spectrum disorder. Nevertheless, to the best of our knowledge, there is no demonstration in the literature on the efficiency of mitochondrial variant calling from whole genome and exome data when compared with the calling using Sanger sequencing data. In addition, there are no studies on mitochondrial haplogroup and variant association with obesity using exome data with potential significant results.

Obesity has become a worldwide epidemic, particularly among Arab populations. In Kuwait, the prevalence of obesity ranges from 37 to 50% (Ng et al., 2014; World Health Organization [WHO], 2018). While obesity has a large heritable component, elucidating these determinants is complicated by the complex interplay between environmental, behavioral, and genetic factors. Genetic studies into obesity have identified monogenic genes using linkage analysis and common variants using genome-wide association studies (GWASs) (Ramachandrappa and Farooqi, 2011). However, obesity-associated genetic loci are often identified in nuclear DNA and have a modest effect that cannot explain the high heritability estimates, and well-defined genetic loci are often from rare familial syndromes (Stunkard et al., 1990; Bouchard and Perusse, 1993; Sorensen et al., 1998).

Mitochondrial variants, haplogroups, and copy number variations have been proposed as potential causative or protective factors for complex and multifactorial disorders and can explain the missing heritability for obesity.

Several studies have shown correlations between mitochondrial function and obesity (Wortmann et al., 2009; Fernandez-Sanchez et al., 2011; Naukkarinen et al., 2014). These findings have led to studies that have evaluated whether mitochondrial dysfunction in obesity is due to inherited sequence variations. Several mitochondrial variants and haplogroups have been associated with obesity in different ethnicities (Yang et al., 2011; Grant et al., 2012; Nardelli et al., 2013; Flaquer et al., 2014; Knoll et al., 2014; Ebner et al., 2015; Veronese et al., 2018; Eaaswarkhanth et al., 2019). The genotyping data used in these studies were from the mitochondrial control region (Nardelli et al., 2013; Ebner et al., 2015; Veronese et al., 2018; Eaaswarkhanth et al., 2019) and/or extracted from GWASs (Yang et al., 2011; Grant et al., 2012; Flaquer et al., 2014; Knoll et al., 2014).

The present study investigates the role of mitochondrial variants and haplogroups contributing to the risk of obesity in Arabs in Kuwait. Arabian Peninsula populations present unique features in the context of the worldwide genetic diversity (Alsmadi et al., 2013, 2014; Thareja et al., 2015): (1) they resulted from an old and continuous admixture between African, European, and Asian ancestries; (2) the high level of consanguineous mating increases frequencies of rare variants and extends stretches of homozygous chromosomal fragments. Further, the Arabian Peninsula, by virtue of being the exit point for the Southern Route of Africa, was indeed the first staging post in the spread of modern humans around the world (Fernandes et al., 2012). Hence, the characterization of Arabian exome variant data potentiates the easy detection of functional variants, contributing information to discover new disease mechanisms.

The present study examined the association of mitochondrial haplogroups and variants with obesity using off-target whole-exome data from a Kuwaiti population. The study used whole-genome data and Sanger sequencing data as quality control samples for mitochondrial variants called from exome reads.

MATERIALS AND METHODS

Exome Data

We analyzed 288 exomes from Kuwaiti individuals who were included in a previously published study (John et al., 2018). Samples from the individuals were divided into two groups according to body mass index (BMI): obese (BMI $\geq$ 30 kg/m^2; n = 152) and non-obese (<30 kg/m^2; n = 136). Samples were sequenced using two different exome kits: samples from 160 individuals were sequenced using the TruSeq Exome Enrichment kit, and samples from the remaining 128 individuals were sequenced using the Nextera Rapid Capture Exome kit, both using the Illumina HiSeq platform (Illumina Inc. United States) (John et al., 2018). Target files of both exome kits show that both contain the same 11 mtDNA regions where each target region covers an average of 1,000 bp. Whole genomes from three of the 288 individuals were sequenced in our previous studies (Alsmadi et al., 2014; Thareja et al., 2015), and mtDNA sequences extracted from these individuals were used as quality control samples for mitochondrial variants called from exome reads. Furthermore, we previously sequenced mtDNA D-loops from 173 individuals using conventional DNA Sanger sequencing (Eaaswarkhanth et al., 2019), and variants called using the control regions were used as quality controls for mitochondrial variant calling using exome sequences.

Mitochondrial DNA Sequences, Variant Calling, and Annotation

Raw paired-end reads (100 bp) from exome sequencing were mapped to human genome assembly GRCh37 using Burrows–Wheeler Aligner (BWA-MEM version v07-17) with default mapping options (Li, 2013). Duplicate reads were removed using Picard version 2.20.2[1]. The revised Cambridge Reference Sequence (rCRS) (Andrews et al., 1999) for human mtDNA as deposited in the GenBank NCBI database under accession number NC_012920.1 was extracted using SAMtools version 0.1.19 (Li et al., 2009), and the average mtDNA coverage was calculated using Genome Analysis Toolkit (GATK) version v3.8-1-0 (McKenna et al., 2010). mtDNA BAM files were generated for each sample. We subsequently used the GATK haplocaller with default parameters on the extracted mtDNA BAM files to generate variants for each sample in Genomic Variant Call Format (GVCF). All the GVCF files were combined into a single GVCF that was subsequently used to genotype the mtDNA variants. Annotation of the variants was performed using Ensembl Variant Effect Predictor (McLaren et al., 2016) and mitomap/mitomaster[2] (Lott et al., 2013).

Haplogroup Assignment

We used raw variant calling format files for whole mtDNA from 288 samples from Kuwaiti individuals and three whole-genome (technical replica) samples to determine their maternal haplogroups. Haplogroup calling was performed using HaploGrep 2.0 (Weissensteiner et al., 2016) based on PhyloTree Build 17 (accessed on 19 December 2019). To determine the accuracy of mitochondrial haplogroup prediction from exome data, we compared the results with a matched mitochondrial haplogroup profiling of 173 samples from Kuwaiti individuals whose mitochondrial control region variants were called using Sanger sequencing in our previously study (Eaaswarkhanth et al., 2019). We also assessed the agreement in assignment of major mitochondrial haplogroups between whole-exome and whole-genome samples. Further, the graphical phylogenetic trees for the haplogroups R and L were generated from HaploGrep 2.0. The Median-Joining networks of R and L haplogroups were constructed using PopART version 1.7 (Leigh and Bryant, 2015).

Statistical Analyses

Statistical analysis of clinical characteristics was performed using R Project for Statistical Computing software (version 3.6.2)[3]. Quantitative clinical variables (assuming continuous values) were ascertained for normality assumption using Shapiro–Wilk test and presented as either mean ± standard deviation or median and interquartile range. Non-parametric Mann–Whitney U test was used to compare age and BMI scores (which may have skewed distribution) between obese and non-obese groups. In the cases of categorical variables, descriptive statistics were presented as number and percentage, and chi-square test was applied to find associations or significant differences between them.

The differences in the counts of non-synonymous over synonymous mutations between the R (protective) and L (risk) haplogroups were examined using Fisher exact test. Further, the significance of the differences in dN/dS ratio between the two haplogroups was calculated using unpaired Wilcoxon rank sum test available in R software.

Principal component analysis (PCA) was conducted to determine whether the mtDNA profiles could cluster the samples based on obese/non-obese categorizations and assigned haplogroups. We used the PCA tools package of the R software to perform the PCA. Fisher exact test was used to investigate the differences in the distribution of mitochondrial haplogroups and variants between the obese and non-obese groups. Additionally, logistic regression analysis was performed to determine haplogroup association (adjusted for age and sex) with traits using IBM SPSS statistical software (version 25). PLINK software (version 1.9) (Chang et al., 2015) was used to test mtDNA variant association (adjusted for age, sex, and maternal haplogroups) with traits. A two-tailed P-value < 0.05 was considered statistically significant.

RESULTS

Study Population

Table 1 shows the descriptive statistics for the clinical characteristics of the study cohort and subcohorts of 152 (52.8%) obese and 136 (47.2%) non-obese Kuwaiti individuals. There were no significant differences between the obese and non-obese

[1] http://picard.sourceforge.net

[2] http://mitomap.org

[3] https://www.R-project.org/

groups in terms of sex; however, the mean age was significantly higher in the obese group compared with the non-obese group ($P > 0.001$). This difference was in agreement with the results reported in our recent study on Arab population from Kuwait (Eaaswarkhanth et al., 2019). As expected, BMI was significantly different between obese and non-obese groups ($P > 0.001$).

Mitochondrial DNA Coverage and Variants

The average coverage of extracted mtDNA sequences from off-target whole-exome samples in our study cohort was 50 × using Nextera Rapid Capture Exome kit and 8 × using TruSeq Exome Enrichment kit. The average coverage of mtDNA sequences from whole-genome (technical replica) samples was 2,491×. The coverage of mtDNA sequences was expected to be high using whole-genome sequencing as the large number of mitochondria present in the cytoplasm contributed to a greater number of reads.

A total of 1,241 mtDNA single-nucleotide polymorphisms (SNPs) and insertion/deletion (INDELs) variants were identified among the 288 whole-exome samples. A comparison of the detected mtDNA variants (SNPs only) from whole-exome data with variants called from the Sanger sequenced reads for the corresponding samples revealed that 77% of the variants were common. Nevertheless, a higher detection rate of variants was observed in Nextera Rapid Capture Exome kit samples (87%) compared with the TruSeq Exome Enrichment kit samples (70%).

Some variants were detected by only Sanger sequencing—such variants were MT:71, 209, 235, 311, 315, 398, 411, 523, 524, 571, 573, 582, 16086, 16186, 16188, 16207, 16217, 16249, 16256, 16293, 16351, and 16399. On the other hand, some mtDNA variants (such as MT:513 and MT:16183) were detected only in whole-exome data. Inconsistencies were seen in the called genotypes at MT:302, MT:309, and MT:310 between Sanger and exome sequencing. Furthermore, all the mtDNA variants identified from whole exomes were detected in whole genomes of the same samples. Only four variants (including a mtDNA variant at position MT:3492) were detected in the whole genomes, but not in whole exomes.

Mitochondrial Haplogroups Associated With Obesity

A total of 12 maternal haplogroups (H, HV, J, K, L, M, N, R, T, U, W, and X) were identified from the mitochondrial variants extracted from the whole-exome samples. The most common maternal haplogroups among the 288 Kuwaiti individuals were J (19%), H (16%) L (13%), R (11%), and U (11%) (**Figure 1**). Good concordance was observed in haplogroup calling using whole exomes versus whole genomes versus Sanger sequenced reads. Among the 173 samples used for both exome sequencing and Sanger sequencing, 123 had the same major maternal haplogroups detected in both exomes and Sanger sequences (**Supplementary Table 1**). Further, same haplogroups (even at the resolution of subclade) were detected in both exomes and whole genomes in three samples that were analyzed using both sequencing techniques. One sample that had the same mitochondrial haplogroup detected using whole-exome and whole-genome sequencing differed in Sanger sequencing reads.

To assess the amount of variation observed in mtDNA that could be attributed to BMI classification, PCA was performed for the 288 samples, in which 152 (52.8%) were classified as obese and 136 (47.2%) were non-obese (**Figure 2**). The samples did not cluster based on BMI classification or sex (data not shown), which may indicate that the genetic heritability of obesity in mtDNA is overestimated and/or our data are too small to demonstrate (**Figure 2A**). However, the samples clustered well based on haplogroup origin, emphasizing the importance of mtDNA when studying the maternal relatedness between individuals and populations (**Figure 2B**).

Supplementary Table 2 lists the variants used to assign haplogroup for each sample and the haplogroup assigned to the sample along with the HaploGrep 2 score (Weissensteiner et al., 2016). **Table 2** shows the frequencies of each haplogroup in the obese and non-obese groups. The results indicated that individuals with the R haplogroup are at low risk of obesity [odds ratio (OR) = 0.4; $P = 0.017$)] and remained significant after adjusting the model for age and sex [OR = 0.311; 95% confidence interval (CI) = 0.135–0.717; and $P = 0.006$]. In addition, males with haplogroup R had a greater likelihood of being non-obese (OR = 4.84; $P = 0.035$) than obese (data not shown). On the other hand, haplogroup L individuals had a twofold increased risk of obesity (OR = 1.94), which was not significant ($P = 0.074$) but became significant after adjusting for age and sex using multivariate logistic regression (OR = 2.285; 95% CI = 1.02–5.14; and $P = 0.046$) (**Table 2**). The frequencies of haplogroups H and L differed significantly between obese and non-obese groups, where haplotype R was more frequent in the non-obese group and L was more frequent in the obese group (**Figure 3**). The complete phylogeny and Median-Joining networks of these obesity risk-associated haplogroups R and L along with their subclades in Kuwaiti individuals are presented in **Supplementary Figures 1,2** and **Figures 4A,B**, respectively.

dN/dS Ratio Between the R and L Haplogroups

Upon performing Fisher exact test on the counts of non-synonymous and synonymous substitutions between R and L haplogroups, we did not find any significant difference (OR = 0.636; CI = 0.14–2.77; $P = 0.547$) in the distribution of non-synonymous and synonymous substitutions between them. However, upon computing the dN/dS ratios, we observed the median (IQR) of dN/dS ratio as 0.6 (0.425) and 0.364 (0.196) for R and L, respectively. A statistical test using unpaired Wilcoxon rank sum test between dN/dS ratio of R and L suggested significant differences ($P = 0.0024$) among them (**Figure 5**).

Mitochondrial DNA Variants Associated With Obesity

Significant associations ($P > 0.05$) with BMI classifications were found for 14 mtDNA variants (**Table 3**); however, three of these associations were no longer significant after adjusting for age, sex, and maternal haplogroups using multivariate logistic

TABLE 1 | Clinical characteristics of the Kuwaiti study samples.

	Obese	Non-obese	Total	*P*-value for obese versus non-obese groups
	(*N* = 152) *n* (%)	**(*N* = 136) *n* (%)**	**(*N* = 288) *n* (%)**	
Gender				
Male	54 (35.5%)	63 (46.3%)	117 (40.6%)	0.08
Female	98 (64.5%)	73 (53.7%)	171 (59.4%)	
Age (years)				
≤25	4 (63.4%)	7 (63.6%)	6 (2.1%)	2.90E-11
26–34	6 (10.5%)	51 (89.5%)	57 (19.8%)	
35–44	21 (45.6%)	25 (54.3%)	46 (15.9%)	
≥45	121 (69.5%)	53 (30.5%)	174 (60.4%)	
Mean ± SD	57.1 ± 14.4	43.8 ± 16.7	50.8 ± 16.8	
Median (IQ)	58.5 (48–66.25)	36.5 (31–57.5)	52.5 (35–64)	
BMI score				
Mean ± SD	39.5 ± 6.8	24.8 ± 2.9	32.5 ± 9.1	<2.2E-16
Median (IQ)	38.5 (33.8–43.7)	24.8 (22.6–27.1)	30.9 (25.2–38.8)	

P-value for age categories for obese versus non-obese groups were calculated using Mann–Whitney U test. P-value for sex counts in obese versus non-obese groups were calculated using Chi-sq test. Abbreviations: BMI, body mass index; N, number of individuals; SD, standard deviation; IQ, inter-quartile.

regression (**Table 3**). In addition, nine mtDNA variants were found when the model was corrected for age, sex, and maternal haplogroups. Thus, a total of 20 SNPs were correlated with obesity, among which 11 were positively (OR > 1) correlated with obesity. The missense variant MT:5460G > A (Ala331Thr) in the *MT-ND2* gene showed the most significant correlation (*P* = 0.009) and was associated with a threefold increased risk of obesity. Among the nine negatively (OR > 1) correlated SNPs, the upstream variant MT:16362T > C in the *MT-TP* gene (encoding microsomal triglyceride transfer protein) showed the most significant (*P* = 0.007; OR = 0.38) negative association with obesity.

Functional analysis of the consequences of these 20 variants revealed that 12 were located in coding exonic regions, four were in non-coding regions, and four were in gene upstream regions. The SIFT and PolyPhen-2 tools that assess the impact of variants on the protein structure and function predicted these variants as “tolerated” and “benign,” respectively. None of these variants was annotated as pathogenic for obesity by ClinVar, Mitomaster, and Mitomap databases. Nevertheless, the MT:5460G > A missense variant, which is positively correlated with obesity, has been associated with Alzheimer’s disease and Parkinson’s disease (Lin et al., 1992; Schnopp et al., 1996), and the MT:16362T > C, which was negatively correlated with obesity, was shown to be associated with lower mtRNA expression levels and affect uncoupled mitochondrial respiration (Zhou et al., 2017).

Nine SNPs were detected in only one of the BMI groups. Among the four SNPs that were detected in the obese group only, MT:2758G > A (a non-coding variant of *MT-RNR2*) was observed in eight individuals, MT:8468C > T (a synonymous variant of *MT-ATP8*) was observed in six individuals, MT:16320C > T (an upstream variant from *MT-TP*) was observed in six individuals, and MT:93A > G (an upstream variant from *MT-TF*) was observed in six individuals. Five SNPs were detected in the non-obese group only, including MT:10499A > G (a synonymous variant of *MT-ND4L*) observed in six individuals, MT:10609T > C (a missense variant of *MT-ND4L*) observed in four individuals, MT:3197T > C (a non-coding variant of *MT-RNR2*) observed in five individuals, and MT:16288T > C and MT:16359T > C (upstream variants of *MT-TP*) observed in four individuals.

DISCUSSION

Previous studies have identified mitochondrial haplogroups associated with obesity. Mitochondrial group T was associated with an increased risk of obesity in Austrian (Ebner et al., 2015) and southern Italian (Nardelli et al., 2013) populations. Mitochondrial haplogroups X and H were reported to decrease risk of obesity in Caucasians of northern European origin in the United States (Yang et al., 2011) and Arabs from Kuwait (Eaaswarkhanth et al., 2019), respectively. It should be noted that these significant mitochondrial haplogroup studies did not follow the same approach and that there were differences in the age of the participants (adults versus children), BMI grouping, number of mtDNA variants, and regions studied. Differences in the region studied may explain why some studies, such as those conducted in European–American and African–American populations, found no association between mitochondrial variants and obesity (Grant et al., 2012).

In the present study cohort, the distribution of maternal linage frequency was 75% Western Eurasian, 12.5% African, and 12.5% Asian. This distribution is consistent with previous published distributions in Kuwait (Scheible et al., 2011) as well as neighboring countries, such as Iraq (Al-Zahery et al., 2003) and Saudi Arabia (Abu-Amero et al., 2008).

The frequency of mitochondrial haplogroup R in non-obese group was significantly higher than that in obese group. In the present study, most individuals (85%) in haplogroup R belonged to the R0a clade (**Figure 4A** and **Supplementary Table 2**), which is defined by the mutations MT:64C > T, MT:2442T > C, MT:3847T > C, MT:13188C > T, MT:16126T > C, and MT:16362T > C (Abu-Amero et al., 2007). Univariate and

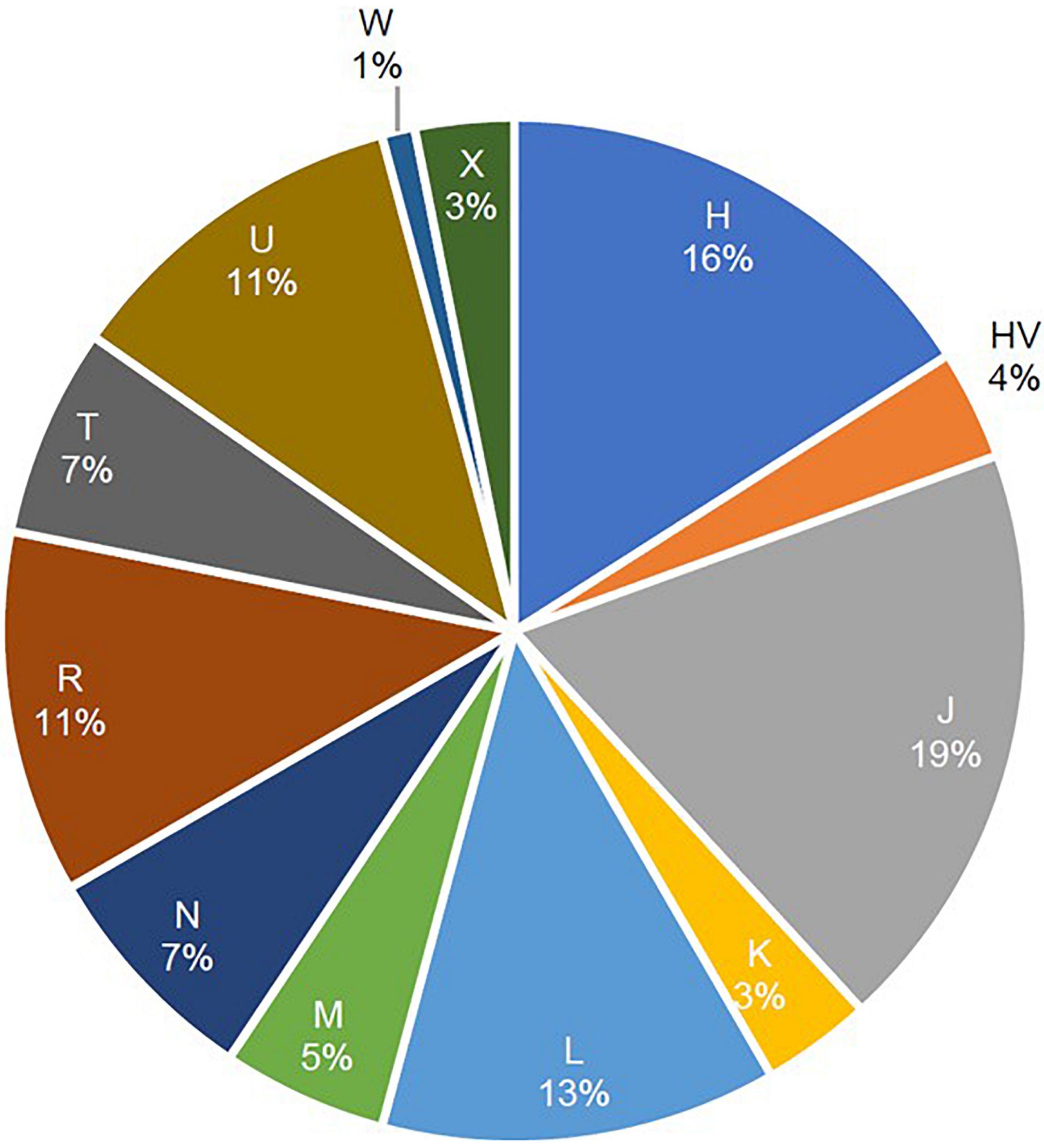


FIGURE 1 | Frequencies of mitochondrial haplogroups in the study cohort of 288 Arab individuals from Kuwait.

multivariate analyses showed that these defining mutations were also negatively correlated with obesity. The same was also observed for MT:11719A > G, which is the defining mutation for an ancestor haplogroup R0. It is important to note that R0a is the most frequent sub-haplogroup in the Arabian Peninsula, with frequency of 5–30%, and it has been speculated that several founders of R0a are present in southern Arabia (Cerny et al., 2011; Scheible et al., 2011). The overall frequency of the R0a haplogroup in our samples was 10%, which is in agreement with the frequency range in the Arabian Peninsula.

The frequency of mitochondrial haplogroup L in the obese group was significantly higher than that in the non-obese group after adjusting for age and sex. Half of the individuals in haplogroup L belonged to the L3 clade (**Figure 4B** and **Supplementary Table 2**), which is associated with out-of-Africa migration into Asia (Cabrera et al., 2018). Within the human mtDNA tree, haplogroup L3 encompasses not only many sub-Saharan Africans but also all ancient non-African lineages. The similarity of the age of L3 to its two non-African daughter haplogroups, M and N, suggested that the same process was likely responsible for both the L3 expansion in Eastern Africa and the dispersal of a small group of modern humans out of Africa to settle the rest of the world (Soares et al., 2012). The defining mutations for African subclade L3, MT:769G > A and MT:1018G > A (van Oven and Kayser, 2009), were positively correlated with risk of obesity after adjusting for age, sex, and maternal haplogroup. Mutations from other subclades of haplogroup L were also positively correlated with

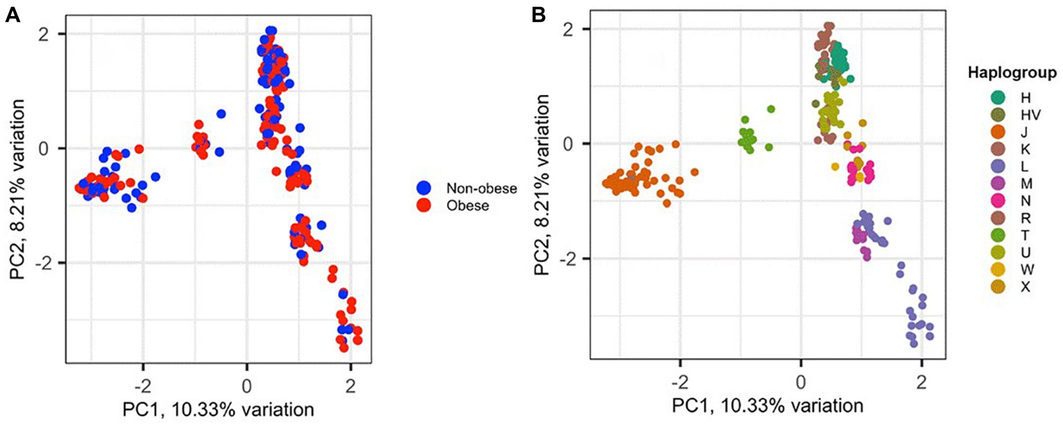


FIGURE 2 | Principal component analysis (PCA) of the 288 Kuwaiti samples based on their mtDNA. **(A)** The two colors represent obese and non-obese samples, and the colors on **(B)** represent haplogroup origin of each sample. PC1 and PC2 on the *x*- and *y*-axes represent principal component 1 and principal component 2 and their variations in percentage, respectively.

TABLE 2 | Mitochondrial haplogroups associated with obesity in the Kuwaiti population.

Haplogroup	Obese	Non-obese	OR	*P*-value	OR (95%CI)* after adjusting the model for age and sex	*P*-value* after adjusting the model for age and sex
	***N* (152)**	***N* (136)**				
H	22 (14.47%)	24 (17.65%)	0.79	0.463	0.786 (0.396–1.562)	0.492
HV	5 (3.29%)	5 (3.68%)	0.89	0.858	0.804 (0.211–3.064)	0.749
J	28 (18.42%)	26 (19.12%)	0.96	0.88	1.027 (0.539–1.958)	0.936
K	6 (3.95%)	4 (2.94%)	1.36	0.641	2.11 (0.504–8.829)	0.307
L	24 (15.79%)	12 (8.82%)	1.94	0.074	2.285 (1.015–5.141)	0.046
M	7 (4.61%)	8 (5.88%)	0.77	0.626	0.449 (0.143–1.415)	0.172
N	12 (7.89%)	9 (6.62%)	1.21	0.677	1.173 (0.434–3.173)	0.753
R	11 (7.24%)	22 (16.18%)	0.4	0.017	0.311 (0.135–0.717)	0.006
T	13 (8.55%)	6 (4.41%)	2.03	0.158	1.906 (0.657–5.532)	0.235
U	16 (10.53%)	16 (11.76%)	0.88	0.738	1.166 (0.509–2.672)	0.717
W	3 (1.97%)	0 (0%)	–	0.1	–	0.996
X	5 (3.29%)	4 (2.94%)	1.12	0.865	1.109 (0.255–4.835)	0.89

**Values after adjustment for age and gender.*
Abbreviations: N, number of individuals; OR, odds ratio; CI, confidence intervals for OR as calculated using logistic regression model using PLINK.

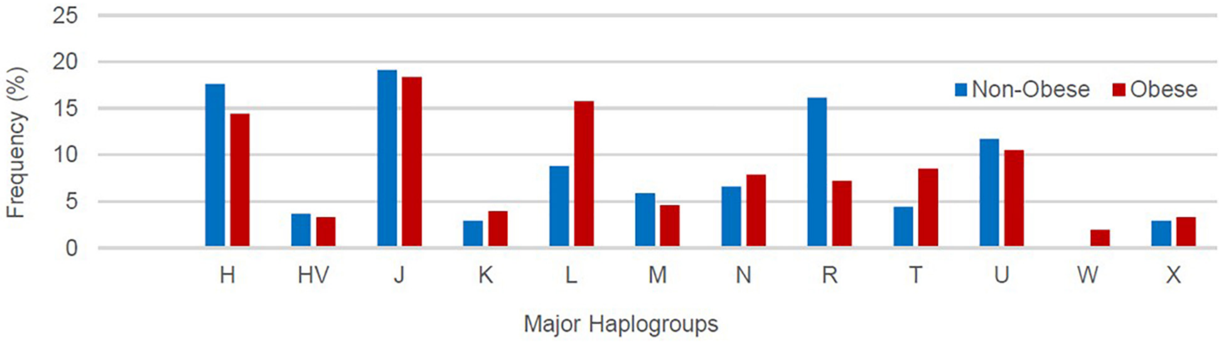


FIGURE 3 | Frequency distribution of major mitochondrial DNA (mtDNA) haplogroups in obese and non-obese groups.

risk of obesity, including MT:709G > A, MT:8468C > T, MT:3594C > T, MT:13650C > T, MT:825T > A, MT:5460G > A, MT:16320C > T, and MT:93A > G (van Oven and Kayser, 2009).

We observed that mitochondrial haplogroup T, which is known to increase risk of obesity (Nardelli et al., 2013; Ebner et al., 2015), showed a higher frequency in the obese

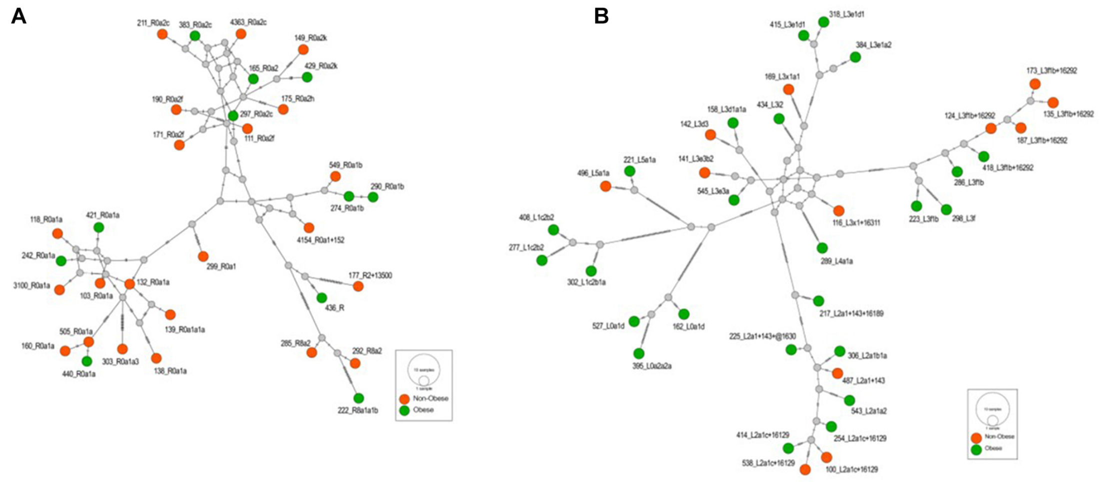


FIGURE 4 | (A) The Median-Joining network of the haplogroup R that is associated with the reduced risk of obesity in the Kuwaiti population. The hatch marks on the edges denote nucleotide positions. **(B)** The Median-Joining network of the haplogroup L that is associated with the increased risk of obesity in the Kuwaiti population. The hatch marks on the edges denote nucleotide positions.

group compared with the non-obese group, but this was not significant. However, its defining mutations, MT:11812A > G and MT:14233A > G, correlated positively with risk of obesity (*P* = 0.029 and *P* = 0.032, respectively). Examination of the mitomap database (Lott et al., 2013) revealed another mutation, MT:10609T > C, which is a marker for a subclade of haplogroup F, which was negatively correlated with risk of obesity. Interestingly, this SNP was associated with athlete status and sprint performance in a Korean population (Hwang et al., 2019).

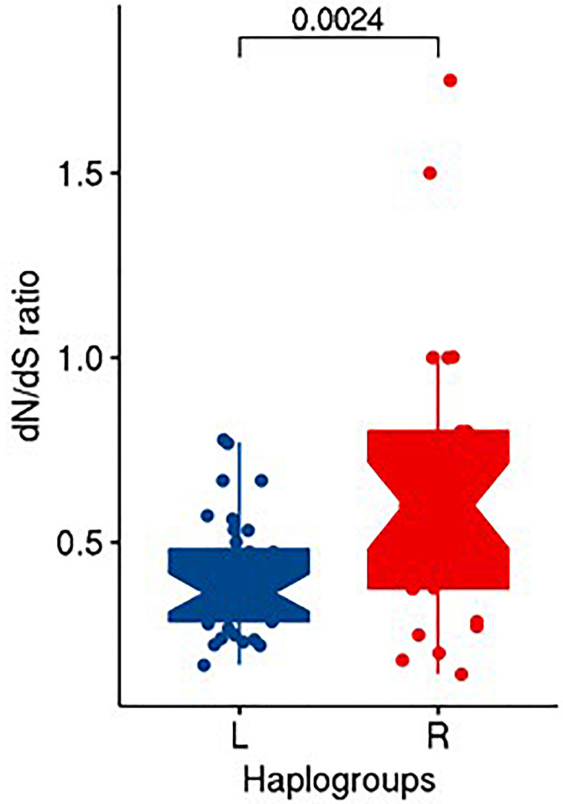


FIGURE 5 | Distribution of dN/dS ratio in R and L haplogroups.

The metric of evolutionary rate ratio *dN/dS* (ratio of non-synonymous to synonymous substitution rates) indicates how quickly a protein's constituent amino acids change relative to synonymous changes. A value of <1 indicates purifying selection, =1 indicates evolving neutrally, and >1 indicates positive (diversifying) selection (Spielman and Wilke, 2015). For both the R and L haplogroups that we observed in our study as associated with obesity, the median *dN/dS* ratio was <1 (0.600 and 0.364, respectively), indicating that both the haplogroups undergo purifying selection in the Kuwaiti population; however, the lower ratio in L haplogroup suggested that the L haplogroup (risk effect on obesity) experienced more purifying selection (or negative selection) than the R haplogroup (protective effect on obesity) by purging deleterious mutations in the process of evolution.

We found that several variants located in nicotinamide adenine dinucleotide (NADH) dehydrogenase subunit (*MT-ND1*, *MT-ND4*, and *MT-ND5*) genes, respiratory complex I, and mitochondrial 12S and 16S ribosomal RNA (*MT-RNR1* and *MT-RNR2*) genes were significantly positively or negatively correlated with risk of obesity. In the *MT-RNR2* gene, MT:2758G > A was only identified in the obese group, whereas MT:3197T > C was only identified in the non-obese group. NADH dehydrogenase is required for energy generation in the cell; therefore, variants within its seven encoding genes could result in metabolic disorders including obesity (Flaquer et al., 2014). The *MT-RNR1* gene encodes MOTS-C protein that regulates insulin sensitivity and metabolic homeostasis and plays a protective role against diet-induced obesity (Lee et al., 2015). Furthermore, *MT-RNR2* encodes Humanin, which plays a protective role against oxidative

TABLE 3 | Mitochondrial variants associated with obesity in the Kuwaiti population.

mtDNA variants	Gene	Consequence	Obese F	Non-obese F	OR (95% CI)	*P*-value	OR (95% CI)*	*P*-value*
MT:709G > A	*RNR1*	Non-coding	0.185	0.075	2.8 (1.304–6.012)	0.008	2.4 (1.048–5.495)	0.038
MT:11812A > G	*ND4*	Synonymous	0.065	0.007	9.366 (1.183–74.17)	0.011	10.53 (1.261–87.88)	0.029
MT:14233A > G	*ND6*	Synonymous	0.065	0.007	9.296 (1.174–73.61)	0.012	10.15 (1.212–84.9)	0.032
MT:16362T > C	***TP***	**Upstream**	**0.118**	**0.227**	**0.456 (0.241–0.864)**	**0.017**	**0.377 (0.185–0.766)**	**0.007**
MT:13188C > T	*ND5*	Synonymous	0.053	0.127	0.384 (0.16–0.922)	0.035	0.287 (0.107–0.769)	0.013
MT:16294C > T	*TP*	Upstream	0.144	0.066	2.388 (1.059–5.385)	0.036	2.08 (0.861–5.024)	0.103
MT:13392T > C	*ND5*	Synonymous	0.052	0.007	7.5 (0.925–60.76)	0.038	8.894 (1.014–78.02)	0.048
MT:58T > C	*RNR1*	Upstream	0.02	0.076	0.25 (0.067–0.929)	0.042	0.238 (0.058–0.978)	0.046
MT:64C > T	*RNR1*	Upstream	0.066	0.143	0.421 (0.188–0.943)	0.047	0.327 (0.132–0.809)	0.015
MT:3594C > T	*ND1*	Synonymous	0.086	0.029	3.085 (0.98–9.703)	0.048	3.519 (1.042–11.88)	0.042
MT:13650C > T	*ND5*	Synonymous	0.086	0.029	3.084 (0.98–9.701)	0.048	3.517 (1.043–11.87)	0.042
MT:1018G > A	*RNR1*	Non-coding	0.087	0.03	3.083 (0.979–9.699)	0.048	3.561 (1.054–12.03)	0.04
MT:8292G > A	*TK*	Upstream	0.013	0.06	0.209 (0.043–1.005)	0.049	0.205 (0.038–1.097)	0.064
MT:13500T > C	*ND5*	Synonymous	0.013	0.059	0.21 (0.043–1.007)	0.049	0.228 (0.041–1.269)	0.091
MT:16193C > T	TP	Upstream	0.073	0.022	3.456 (0.943–12.66)	0.056	4.527 (1.105–18.55)	0.035
MT:11719A > G	ND4	Synonymous	0.24	0.343	0.603 (0.358–1.015)	0.064	0.516 (0.273–0.975)	0.041
MT:825T > A	*RNR1*	Non-coding	0.046	0.007	6.372 (0.773–52.48)	0.071	8.891 (1.008–78.42)	0.049
MT:5460G > A	***ND2***	**Missense**	**0.111**	**0.052**	**2.285 (0.916–5.693)**	**0.087**	**3.762 (1.375–10.29)**	**0.009**
MT:2442T > C	*RNR2*	Non-coding	0.06	0.123	0.454 (0.193–1.067)	0.091	0.338 (0.13–0.88)	0.026
MT:769G > A	*RNR1*	Non-coding	0.092	0.037	2.636 (0.923–7.527)	0.093	3.214 (1.045–9.885)	0.041
MT:3847T > C	*ND1*	Synonymous	0.066	0.123	0.508 (0.222–1.165)	0.147	0.359 (0.142–0.91)	0.03
MT:3537A > G	*ND1*	Synonymous	0.006	0.03	0.213 (0.023–1.933)	0.188	0.088 (0.008–0.925)	0.042
MT:7853G > A	*CO2*	Missense	0.006	0.022	0.289 (0.029–2.814)	0.343	0.091 (0.008–0.945)	0.044

**Values after adjustment for age, gender and maternal haplogroup.*
Abbreviations: F, Frequency; N, number of individuals; OR, odds ratio; CI, confidence intervals.
The rows in bold font indicate the most significant (adjusted P-values < 0.01) associations with obesity.

stress (Voigt and Jelinek, 2016). Thus, variants within these genes could potentially interfere with their function, resulting in an increased or decreased risk of obesity.

To prioritize the significant variants identified in our study, we focused on missense mutations leading to amino acid substitutions that were unique to either the obese or non-obese group. The missense variant MT:5460G > A from the *MT-ND2* gene was only positively correlated with obesity. This finding was in agreement with findings from other studies that reported that variants within the *MT-ND2* gene were associated with body fat mass (Yang et al., 2011) and increased BMI (Flaquer et al., 2014). The *MT-ND4L* gene has been associated with obesity (Flaquer et al., 2014) and is a mitochondrial encoding subunit of respiratory complex I. In the present study, the missense mutation MT:10609T > C in the *MT-ND4L* gene was negatively correlated with risk of obesity. Cytochrome *c* oxidase subunit gene 2 (*MT-CO2*), which is an important regulator of the OXPHOS system, was also negatively correlated with risk of obesity. We found that *MT-CO2* harbored a missense mutation, MT:7853G > A, which exhibited a protective role for obesity. A previous study reported that *MT-CO2* (Kraja et al., 2019) and variants within this gene were associated with obesity, but not after adjusting for multiple testing (Liu et al., 2012).

Off-target whole-exome sequencing for the entire mitochondrial genome revealed a good variable coverage depending on the exome capture kit used. The non-uniformity of mitochondrial coverage between the two exome kits could have been due to differences in design and target sequences. This may explain why we observed a higher overlap in variants between sequencing using Sanger technology and sequencing by Nextera Rapid Capture Exome kit compared with sequencing with the TruSeq Exome Enrichment kit (obsolete). Nevertheless, mtDNA variants from both the exome capture kits detected almost all the mtDNA variants identified using indirect whole-genome sequencing of replicated samples. Thus, whole-exome sequencing is a cost- and time-effective alternative for mitochondrial monogenic (Griffin et al., 2014) and association studies (Li et al., 2014) compared with whole-genome sequencing. The reasons for the difference in detection of variants between Sanger and exome variants include the following: (1) low read coverage of exome data at the start and end of the mitochondrial genome (especially when the average mtDNA coverage is <10); (2) repeated poly-C sequencing error using exome data; and (3) the Sanger variant identification pipeline (Eaaswarkhanth et al., 2019) that uses a predicted mtDNA sequence from a sequence browser with manual adjustment could result in a number of false-positive variants.

PCA with the mitochondrial haplogroup profiling from the 288 whole-exome study samples showed a good clustering of haplogroups, which validates the bioinformatics pipeline used in the present study. Moreover, we observed a high concordance (71%) of mitochondrial haplogroup profiling between variants from whole-exome data and the D-loop region data from conventional Sanger sequencing. One sample that was sequenced with whole-genome and exome kits displayed the same assignment of major haplogroup; however, the D-loop Sanger sequencing of the same sample predicted a different haplogroup. This could have an impact on the significant haplogroups identified in previous studies for complex disorders including obesity due to lower resolution or low number of variants used in the studies.

High-throughput NGS of the mitochondrial genome has advantages compared with Sanger sequencing. However, a technical comparison of both the technologies is required to fully understand and unify their results. The present study compared both technologies and found that some variants were only detected by Sanger sequencing and not NGS; however, this discrepancy could be due to low coverage, the whole-exome capture kit design, target sequences, and machine-specific and human error. Nevertheless, other studies observed the same phenomena despite good mitochondrial sequence coverage. For example, variants at positions MT:16183 (Griffin et al., 2014) and MT:523–524 (Park et al., 2017) were only detected by Sanger sequencing. This could have been due to INDEL alignment errors, as its corresponding position in NGS is MT:513. We also found variants that were incorrectly reported between Sanger sequencing and NGS on the same samples at positions MT:302, MT:309, and MT:310 (Park et al., 2017), which could be also due to alignment errors resulting from the complexity of the region. Interestingly, we also observed a sequencing error variant at position MT:3492 that was only detected by whole-genome sequencing and not exome sequencing. This discrepancy may have been an NGS whole-genome sequencing error (Li et al., 2010).

The present study has some limitations. First, the study did not explore mtDNA heteroplasmic variants within the obese and non-obese groups, as these require high coverage sequences. Second, in order to increase the number of study samples, the study utilized data (generated in our previous studies) obtained using two different exome capture kits from Illumina; the Nextera Rapid Capture Exome kit gave a coverage of 50×, while the TruSeq Exome Enrichment kit gave a coverage of mere 8×; having the second one with a very low coverage can weaken the results by not capturing the variants. Third, we divided our study population into obese and nonobese non-obese groups, and the resultant subcohorts were small in size. Despite these limitations, this study paves the way for a larger study to investigate common complex disorders including obesity using whole mtDNA extracted from whole-exome data with greater coverage exome capture kit.

CONCLUSION

Indirect whole-exome sequencing of 288 Kuwaiti individuals revealed negative and positive associations of mitochondrial haplogroups R and L, respectively, with obesity. We also identified significantly distributed mtDNA variants among the obese and non-obese groups that were mostly haplogroup-defining mutations. We identified several variants of the NADH dehydrogenase subunit that were significantly positively or negatively correlated with risk of obesity. The present study is the first to utilize whole-exome data to extract entire mitochondrial haplogroups and determine their association with obesity in the Arabian Peninsula.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Dasman Diabetes Institute. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MD, FA-M, and TT designed and performed the study and wrote the manuscript. RN and MM performed the sequencing. HA and PS performed the statistical analyses. ME, SJ, and PH participated in data analysis. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors thank the Kuwait Foundation for Advancement in Sciences for the institutional funding.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.626260/full#supplementary-material

Supplementary Figure 1 | The complete phylogeny of the R haplogroup in the Kuwaiti population.

Supplementary Figure 2 | The complete phylogeny of the L haplogroup in the Kuwaiti population.

Supplementary Table 1 | Haplogroup assignments using exome sequence data, Sanger sequencing data and whole genome sequence data.

Supplementary Table 2 | Mitochondrial variants used to assign haplogroups in the study cohort of 288 Kuwaiti individuals.

REFERENCES

Abu-Amero, K. K., Gonzalez, A. M., Larruga, J. M., Bosley, T. M., and Cabrera, V. M. (2007). Eurasian and african mitochondrial DNA influences in the saudi arabian population. *BMC Evol. Biol.* 7:32. doi: 10.1186/1471-2148-7-32

Abu-Amero, K. K., Larruga, J. M., Cabrera, V. M., and Gonzalez, A. M. (2008). Mitochondrial DNA structure in the Arabian Peninsula. *BMC Evol. Biol.* 8:45. doi: 10.1186/1471-2148-8-45

Alsmadi, O., John, S. E., Thareja, G., Hebbar, P., Antony, D., Behbehani, K., et al. (2014). Genome at juncture of early human migration: a systematic analysis of two whole genomes and thirteen exomes from kuwaiti population subgroup of inferred saudi arabian tribe ancestry. *PLoS One* 9:e99069. doi: 10.1371/journal.pone.0103691

Alsmadi, O., Thareja, G., Alkayal, F., Rajagopalan, R., John, S. E., Hebbar, P., et al. (2013). Genetic substructure of kuwaiti population reveals migration history. *PLoS One* 8:e74913. doi: 10.1371/journal.pone.0074913

Al-Zahery, N., Semino, O., Benuzzi, G., Magri, C., Passarino, G., Torroni, A., et al. (2003). Y-Chromosome and MtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-neolithic migrations. *Mol. Phylogenet. Evol.* 28, 458–472. doi: 10.1016/s1055-7903(03)00039-3

Anderson, S., Bankier, A. T., Barrell, B. G., Debruijn, M. H. L., Coulson, A. R., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465.

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M., and Howell, N. (1999). Reanalysis and revision of the cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* 23, 147–147. doi: 10.1038/13779

Bouchard, C., and Perusse, L. (1993). "Genetic-Aspects of obesity," in *Prevention and Treatment of Childhood Obesity*, eds C. L. Williams, and S. Y. S. Kimm, (New York, NY: Annals of the New York Academy of Sciences), 26–35.

Cabrera, V. M., Marrero, P., Abu-Amero, K. K., and Larruga, J. M. (2018). Carriers of mitochondrial DNA macrohaplogroup L3 basal lineages migrated back to africa from Asia around 70,000 years ago. *BMC Evol. Biol.* 18:98. doi: 10.1186/s12862-018-1211-4

Calvo, S. E., Compton, A. G., Hershman, S. G., Lim, S. C., Lieber, D. S., Tucker, E. J., et al. (2012). Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. *Sci. Transl. Med.* 4:118ra10.

Cerny, V., Mulligan, C. J., Fernandes, V., Silva, N. M., Alshamali, F., Non, A., et al. (2011). Internal diversification of mitochondrial haplogroup r0a reveals post-last glacial maximum demographic expansions in South Arabia. *Mol. Biol. Evol.* 28, 71–78. doi: 10.1093/molbev/msq178

Chang, C. C., Chow, C. C., Lcam, T., 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.

Chaubey, G., Metspalu, M., Kivisild, T., and Villems, R. (2007). Peopling of South Asia: investigating the caste-tribe continuum in india. *Bioessays* 29, 91–100. doi: 10.1002/bies.20525

Chinnery, P. F., Elliott, H. R., Hudson, G., Samuels, D. C., and Relton, C. L. (2012). Epigenetics, epidemiology and mitochondrial DNA diseases. *Int. J. Epidemiol.* 41, 177–187. doi: 10.1093/ije/dyr232

Clayton, D. A. (2000). Transcription and replication of mitochondrial DNA. *Hum. Reprod.* 15(Suppl. 2), 11–17. doi: 10.1093/humrep/15.suppl_2.11

Delmiro, A., Rivera, H., Garcia-Silva, M. T., Garcia-Consuegra, I., Martin-Hernandez, E., Quijada-Fraile, P., et al. (2013). Whole-Exome sequencing identifies a variant of the mitochondrial Mt-Nd1 gene associated with epileptic encephalopathy: west syndrome evolving to lennox-gastaut syndrome. *Hum. Mutat.* 34, 1623–1627. doi: 10.1002/humu.22445

Diroma, M. A., Calabrese, C., Simone, D., Santorsola, M., Calabrese, F. M., Gasparre, G., et al. (2014). Extraction and annotation of human mitochondrial genomes from 1000 genomes whole exome sequencing data. *BMC Genom.* 15:S2. doi: 10.1186/1471-2164-15-S3-S2

Eaaswarkhanth, M., Melhem, M., Sharma, P., Nizam, R., Al Madhoun, A., Chaubey, G., et al. (2019). Mitochondrial DNA D-Loop sequencing reveals obesity variants in an arab population. *Appl. Clin. Genet.* 12, 63–70. doi: 10.2147/tacg.s198593

Ebner, S., Mangge, H., Langhof, H., Halle, M., Siegrist, M., Aigner, E., et al. (2015). Mitochondrial haplogroup t is associated with obesity in austrian juveniles and adults. *PLos One* 10:e0135622. doi: 10.1371/journal.pone.0135622

Fernandes, V., Alshamali, F., Alves, M., Costa, M. D., Pereira, J. B., Silva, N. M., et al. (2012). The arabian cradle: mitochondrial relicts of the first steps along the Southern route out of Africa. *Am. J. Hum. Genet.* 90, 347–355. doi: 10.1016/j.ajhg.2011.12.010

Fernandez-Sanchez, A., Madrigal-Santillan, E., Bautista, M., Esquivel-Soto, J., Morales-Gonzalez, A., Esquivel-Chirino, C., et al. (2011). Inflammation, oxidative stress, and obesity. *Int. J. Mol. Sci.* 12, 3117–3132.

Flaquer, A., Baumbach, C., Kriebel, J., Meitinger, T., Peters, A., Waldenberger, M., et al. (2014). Mitochondrial genetic variants identified to be associated with BMI in adults. *PLos One* 9:e105116. doi: 10.1371/journal.pone.0105116

Grant, S. F. A., Glessner, J. T., Bradfield, J. P., Zhao, J., Tirone, J. E., Berkowitz, R. I., et al. (2012). Lack of relationship between mitochondrial heteroplasmy or variation and childhood obesity. *Int. J. Obesity* 36, 80–83. doi: 10.1038/ijo.2011.206

Greenberg, B. D., Newbold, J. E., and Sugino, A. (1983). Intraspecific nucleotide-sequence variability surrounding the origin of replication in human mitochondrial-DNA. *Gene* 21, 33–49. doi: 10.1016/0378-1119(83)90145-2

Griffin, H. R., Pyle, A., Blakely, E. L., Alston, C. L., Duff, J., Hudson, G., et al. (2014). Accurate mitochondrial DNA sequencing using off-target reads provides a single test to identify pathogenic point mutations. *Genet. Med.* 16, 962–971. doi: 10.1038/gim.2014.66

Hosgood, H. D., Liu, C. S., Rothman, N., Weinstein, S. J., Bonner, M. R., Shen, M., et al. (2010). Mitochondrial DNA copy number and lung cancer risk in a prospective cohort study. *Carcinogenesis* 31, 847–849. doi: 10.1093/carcin/bgq045

Hwang, I. W., Kim, K., Choi, E. J., and Jin, H. J. (2019). Association of mitochondrial haplogroup F with physical performance in korean population. *Genom. Informat.* 17:e11. doi: 10.5808/gi.2019.17.1.e11

John, S. E., Antony, D., Eaaswarkhanth, M., Hebbar, P., Channanath, A. M., Thomas, D., et al. (2018). Assessment of coding region variants in kuwaiti population: implications for medical genetics and population genomics. *Sci. Rep.* 8:16583.

Kivisild, T. (2015). Maternal ancestry and population history from whole mitochondrial genomes. *Investigat. Genet.* 6:3. doi: 10.1186/s13323-015-0022-2

Knoll, N., Jarick, I., Volckmar, A. L., Klingenspor, M., Illig, T., Grallert, H., et al. (2014). Mitochondrial DNA variants in obesity. *PLos One* 9:e94882. doi: 10.1371/journal.pone.0094882

Kraja, A. T., Liu, C. Y., Fetterman, J. L., Graff, M., Have, C. T., Gu, C., et al. (2019). Associations of mitochondrial and nuclear mitochondrial variants and genes with seven metabolic traits. *Am. J. Hum. Genet.* 104, 112–138.

Lee, C., Zeng, J., Drew, B. G., Sallam, T., Martin-Montalvo, A., Wan, J. X., et al. (2015). The mitochondrial-derived peptide Mots-c promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metab.* 21, 443–454. doi: 10.1016/j.cmet.2015.02.009

Leigh, J. W., and Bryant, D. (2015). Popart: full-feature software for haplotype network construction. *Methods Ecol. Evol.* 6, 1110–1116. doi: 10.1111/2041-210x.12410

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv* [preprint] 3:13033997.

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, M. K., Schonberg, A., Schaeferd, M., Schroeder, R., Nasidze, I., and Stoneking, M. (2010). Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.* 87, 237–249. doi: 10.1016/j.ajhg.2010.07.014

Li, S. T., Besenbacher, S., Li, Y. R., Kristiansen, K., Grarup, N., Albrechtsen, A., et al. (2014). Variation and association to diabetes in 2000 full mtdna sequences mined from an exome study in a danish population. *Eur. J. Hum. Genet.* 22, 1040–1045. doi: 10.1038/ejhg.2013.282

Lin, F. H., Lin, R., Wisniewski, H. M., Hwang, Y. W., Grundkeiqbal, I., Healylouie, G., et al. (1992). Detection of point mutations in codon-331 of mitochondrial nadh dehydrogenase subunit-2 in alzheimer brains.

Biochem. Biophys. Res. Commun. 182, 238–246. doi: 10.1016/s0006-291x(05)80136-6

Liu, C. Y., Yang, Q., Hwang, S. J., Sun, F. Z., Johnson, A. D., Shirihai, O. S., et al. (2012). Association of genetic variation in the mitochondrial genome with blood pressure and metabolic traits. *Hypertension* 60:949. doi: 10.1161/hypertensionaha.112.196519

Loogvali, E. L., Roostalu, U., Malyarchuk, B. A., Derenko, M. V., Kivisild, T., Metspalu, E., et al. (2004). Disuniting uniformity: a pied cladistic canvas of mtdna haplogroup h in eurasia. *Mol. Biol. Evol.* 21, 2012–2021. doi: 10.1093/molbev/msh209

Lott, M. T., Leipzig, J. N., Derbeneva, O., Xie, H. M., Chalkia, D., Sarmady, M., et al. (2013). mtDNA variation and analysis using mitomap and mitomaster. *Curr. Protocols Bioinform.* 1, 1–6.

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

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., et al. (2016). The ensembl variant effect predictor. *Genome Biol.* 17:122.

Nardelli, C., Labruna, G., Liguori, R., Mazzaccara, C., Ferrigno, M., Capobianco, V., et al. (2013). Haplogroup T is an obesity risk factor: mitochondrial dna haplotyping in a morbid obese population from Southern Italy. *Biomed Res. Int.* 2013:631082.

Naukkarinen, J., Heinonen, S., Hakkarainen, A., Lundbom, J., Vuolteenaho, K., Saarinen, L., et al. (2014). Characterising metabolically healthy obesity in weight-discordant monozygotic twins. *Diabetologia* 57, 167–176. doi: 10.1007/s00125-013-3066-y

Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., et al. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the global burden of disease study 2013. *Lancet* 384, 766–781.

Park, S., Cho, S., Seo, H. J., Lee, J. H., Kim, M. Y., and Lee, S. D. (2017). Entire mitochondrial DNA sequencing on massively parallel sequencing for the Korean population. *J. Korean Med. Sci.* 32, 587–592. doi: 10.3346/jkms.2017.32.4.587

Patowary, A., Nesbitt, R., Archer, M., Bernier, R., and Brkanac, Z. (2017). Next generation sequencing mitochondrial dna analysis in autism spectrum disorder. *Autism Res.* 10, 1338–1343. doi: 10.1002/aur.1792

Picardi, E., and Pesole, G. (2012). Mitochondrial genomes gleaned from human whole-exome sequencing. *Nat. Methods* 9, 523–524. doi: 10.1038/nmeth.2029

Ramachandrappa, S., and Farooqi, I. S. (2011). Genetic approaches to understanding human obesity. *J. Clin. Invest.* 121, 2080–2086. doi: 10.1172/jci46044

Samuels, D. C., Han, L., Li, J., Sheng, Q. H., Clark, T. A., Shyr, Y., et al. (2013). Finding the lost treasures in exome sequencing data. *Trends Genet.* 29, 593–599. doi: 10.1016/j.tig.2013.07.006

Scheible, M., Alenizi, M., Sturk-Andreaggi, K., Coble, M. D., Ismael, S., and Irwin, J. A. (2011). Mitochondrial DNA control region variation in a kuwaiti population sample. *Forensic Sci. Int. Genet.* 5, E112–E113.

Schnopp, N. M., Kosel, S., Egensperger, R., and Graeber, M. B. (1996). Regional heterogeneity of Mtdna heteroplasmy in parkinsonian brain. *Clin. Neuropathol.* 15, 348–352.

Soares, P., Achilli, A., Semino, O., Davies, W., Macaulays, V., Bandelt, H. J., et al. (2010). The archaeogenetics of Europe. *Curr. Biol.* 20, R174–R183.

Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., et al. (2012). The expansion of Mtdna haplogroup L3 within and out of Africa. *Mol. Biol. Evol.* 29, 915–927. doi: 10.1093/molbev/msr245

Sorensen, T. I. A., Holst, C., and Stunkard, A. J. (1998). Adoption study of environmental modifications of the genetic influences on obesity. *Int. J. Obesity* 22, 73–81. doi: 10.1038/sj.ijo.0800548

Spielman, S. J., and Wilke, C. O. (2015). The relationship between Dn/Ds and scaled selection coefficients. *Mol. Biol. Evol.* 32, 1097–1108.

Stunkard, A. J., Harris, J. R., Pedersen, N. L., and McClearn, G. E. (1990). The body-mass index of twins who have been reared apart. *New Engl. J. Med.* 322, 1483–1487.

Tang, S., Wang, J., Zhang, V. W., Li, F. Y., Landsverk, M., Cui, H., et al. (2013). Transition to next generation analysis of the whole mitochondrial genome: a summary of molecular defects. *Hum. Mutat.* 34, 882–893.

Thareja, G., John, S. E., Hebbar, P., Behbehani, K., Thanaraj, T. A., and Alsmadi, O. (2015). Sequence and analysis of a whole genome from kuwaiti population subgroup of persian ancestry. *BMC Genom.* 16:92. doi: 10.1186/s12864-015-1233-x

van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* 30, E386–E394.

Veronese, N., Stubbs, B., Koyanagi, A., Vaona, A., Demurtas, J., Schofield, P., et al. (2018). Mitochondrial genetic haplogroups and incident obesity: a longitudinal cohort study. *Eur. J. Clin. Nutrit.* 72, 587–592.

Voigt, A., and Jelinek, H. F. (2016). Humanin: a mitochondrial signaling peptide as a biomarker for impaired fasting glucose-related oxidative stress. *Physiol. Rep.* 4:e12796.

Wagner, M., Berutti, R., Lorenz-Depiereux, B., Graf, E., Eckstein, G., Mayr, J. A., et al. (2019). Mitochondrial DNA mutation analysis from exome sequencing-a more holistic approach in diagnostics of suspected mitochondrial disease. *J. Inherit. Metab. Dis.* 42, 909–917.

Weissensteiner, H., Pacher, D., Kloss-Brandstatter, A., Forer, L., Specht, G., Bandelt, H. J., et al. (2016). Haplogrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* 44, W58–W63.

Wong, L. J. C. (2013). Next generation molecular diagnosis of mitochondrial disorders. *Mitochondrion* 13, 379–387.

World Health Organization [WHO], (2018). *Noncommunicable Diseases Country Profiles*. Geneva: World Health Organization.

Wortmann, S. B., Zweers-Van Essen, H., Rodenburg, R. J. T., Van Den Heuvel, L. P., De Vries, M. C., et al. (2009). Mitochondrial energy production correlates with the age-related BMI. *Pediatric Res.* 65, 103–108.

Yang, T. L., Guo, Y., Shen, H., Lei, S. F., Liu, Y. J., Li, J., et al. (2011). Genetic association study of common mitochondrial variants on body fat mass. *PLoS One* 6:e21595. doi: 10.1371/journal.pone.0021595

Zhou, H., Nie, K., Qiu, R., Xion, J., Shao, X., Wang, B., et al. (2017). Generation and bioenergetic profiles of cybrids with East Asian mtDNA haplogroups. *Oxid Med. Cell Longev.* 2017:1062314. doi: 10.1155/2017/1062314

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: 30 March 2021
doi: 10.3389/fgene.2021.647288

Evolutionary Changes in the Interaction of miRNA With mRNA of Candidate Genes for Parkinson's Disease

*Saltanat Kamenova[1], Assel Aralbayeva[2], Aida Kondybayeva[1], Aigul Akimniyazova[1,3], Anna Pyrkova[3] and Anatoliy Ivashchenko[3]**

[1] Faculty of Medicine and Health Care, Al-Farabi Kazakh National University, Almaty, Kazakhstan, [2] Department of Neurology, Kazakh Medical University, Almaty, Kazakhstan, [3] Faculty of Biology and Biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan

OPEN ACCESS

Edited by:
Elvira Galieva,
Novosibirsk State University, Russia

Reviewed by:
Changwei Shao,
Chinese Academy of Fishery Sciences, China
Mikhail P. Ponomarenko,
Institute of Cytology and Genetics (RAS), Russia

****Correspondence:***
Anatoliy Ivashchenko
a.iavashchenko@gmail.com;
a_ivashchenko@mail.ru

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *29 December 2020*
Accepted: *11 March 2021*
Published: *30 March 2021*

Citation:
Kamenova S, Aralbayeva A, Kondybayeva A, Akimniyazova A, Pyrkova A and Ivashchenko A (2021) Evolutionary Changes in the Interaction of miRNA With mRNA of Candidate Genes for Parkinson's Disease. Front. Genet. 12:647288. doi: 10.3389/fgene.2021.647288

Parkinson's disease (PD) exhibits the second-highest rate of mortality among neurodegenerative diseases. PD is difficult to diagnose and treat due to its polygenic nature. In recent years, numerous studies have established a correlation between this disease and miRNA expression; however, it remains necessary to determine the quantitative characteristics of the interactions between miRNAs and their target genes. In this study, using novel bioinformatics approaches, the quantitative characteristics of the interactions between miRNAs and the mRNAs of candidate PD genes were established. Of the 6,756 miRNAs studied, more than one hundred efficiently bound to mRNA of 61 candidate PD genes. The miRNA binding sites (BS) were located in the 5′-untranslated region (5′UTR), coding sequence (CDS) and 3′-untranslated region (3′UTR) of the mRNAs. In the mRNAs of many genes, the locations of miRNA BS with overlapping nucleotide sequences (clusters) were identified. Such clusters substantially reduced the proportion of nucleotide sequences of miRNA BS in the 5′UTRs, CDSs, and 3′UTRs. The organization of miRNA BS into clusters leads to competition among miRNAs to bind mRNAs. Differences in the binding characteristics of miRNAs to the mRNAs of genes expressed at different rates were identified. Single miRNA BS, polysites for the binding for one miRNA, and multiple BS for two or more miRNAs in one mRNA were identified. Evolutionary changes in the BS of miRNAs and their clusters in 5′UTRs, CDSs and 3′UTRs of mRNA of orthologous candidate PD genes were established. Based on the quantitative characteristics of the interactions between miRNAs and mRNAs candidate PD genes, several associations recommended as markers for the diagnosis of PD.

Keywords: gene, phylogeny, Parkinson's disease, miRNA, mRNA, association, marker

INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disease with a high mortality rate (Sadlon et al., 2019; Zhang et al., 2019; Zhao and Wang, 2019; Salamon et al., 2020). The development of the disease occurs over several years, which raises the possibility that diagnostic methods may be developed to facilitate subsequent therapy. Unfortunately, at present, there are no effective methods for the early diagnosis of this disease, which significantly reduces the effects of treatment (Arshad et al., 2017; Patil et al., 2019). The difficulty of diagnosing PD is attributable to the many genes that participate in the development of this disease (candidate genes), the expression of which changes with the development of several types of neurodegenerative diseases (Behbahanipour et al., 2019). To date, several dozen candidate PD genes have been identified, and their roles in PD must be investigated. Some of these genes are candidate genes for Alzheimer's disease, PD, dementia, Huntington's disease, frontotemporal dementia, and other neurodegenerative diseases (Brennan et al., 2019; Dong et al., 2020; Yan et al., 2020). Several dozen genes encode proteins containing polyglutamine, the number of which ranges from 30 to 100 or more (Cao et al., 2017). It is believed that both the type of neurodegenerative disease and its severity are associated with the number of glutamine residues (Chen et al., 2018). Previous studies have attempted to link single nucleotide polymorphisms with the probability of PD and other neurodegenerative diseases (D'Anca et al., 2019; Hu et al., 2019; Kakati et al., 2019). Several candidate genes for neurodegenerative diseases contain miRNA binding sites (BS) encoding oligopeptides (Niyazova et al., 2015; Kondybayeva et al., 2018) that are believed to be responsible for the development of Alzheimer's disease, PD, dementia, and Huntington's disease. Therefore, it is necessary to establish which miRNAs can interact in such mRNA regions. In recent years, there has been an increased interest in miRNAs, which can selectively alter gene expression and, to varying degrees, regulate it in different tissues (Jurjević et al., 2017; Marques et al., 2017; Starhof et al., 2019; Uwatoko et al., 2019; Chen et al., 2020; Fejes et al., 2020; Nie et al., 2020; Ozdilek and Demircan, 2020; Ravanidis et al., 2020). Considering the possibility that miRNAs may be synthesized in one tissue and subsequently transferred through the bloodstream to other tissues, the issue of their regulation of the expression of candidate disease genes is complex (Ravanidis et al., 2020; Thomas et al., 2020; Wang et al., 2020; Xie et al., 2020). The human genome encodes more than seven thousand miRNAs, some of which can interact with mRNAs of several genes (Ivashchenko et al., 2014a,b,c; Niyazova et al., 2015; Atambayeva et al., 2017), and some genes are the potential targets of many miRNAs (Kondybayeva et al., 2018, 2020; Aisina et al., 2019; Mukushkina et al., 2020), which also makes it difficult to identify selective markers of the disease. The use of well-known bioinformatic approaches did not lead to the identification of reliable miRNA markers of diseases. In this work, we studied the quantitative characteristics of the interactions of known miRNAs with the mRNAs of candidate PD genes. Quantitative characteristics are a necessary and important parameter for assessing the effectiveness of the interactions between miRNAs and mRNAs. The competition among miRNAs to suppress the expression of one gene by positioning their BS with overlapping nucleotide sequences in regions of the mRNAs called clusters has been demonstrated (Aisina et al., 2019; Mukushkina et al., 2020). Additionally, it should be noted that approximately half of the miRNAs are derived from the introns of the host genes, while the rest of the miRNAs are encoded in intergenic regions (Berillo et al., 2013). Consequently, the host gene can be a source of miRNA and, at the same time, a target of miRNA. Since miRNA can be quickly transferred between tissues through the bloodstream, this characteristic considerably complicates the establishment of the origin of miRNA circulating in the blood (Chen et al., 2018; Rosas-Hernandez et al., 2018; Brennan et al., 2019; Ramaswamy et al., 2020). Therefore, it is necessary to reveal the quantitative characteristics of the interactions of all known miRNAs with candidate genes and subsequently investigate the most effective associations of miRNAs and potential target genes. This approach eliminates many artifacts and enables us to increase the reliability of establishing effective associations of miRNAs and candidate genes.

The expression of candidate genes depends on several factors, including miRNAs that regulate gene expression at the posttranscriptional stage (Leggio et al., 2017; Lu et al., 2017; Martinez and Peplow, 2017; Quinlan et al., 2017; Li et al., 2018; Liu et al., 2019; Patil et al., 2019; Qin et al., 2019). It has been established that some miRNAs can interact with several or even hundreds of genes (Atambayeva et al., 2017), and the reverse situation is also observed: one gene can be the target of many miRNAs (Niyazova et al., 2015; Kondybayeva et al., 2018; Aisina et al., 2019). These properties greatly complicate the identification of miRNA associations and genes that can serve as markers of diseases. Many researchers have studied the changes in the concentrations of several miRNAs or the manipulation of the expression levels of several genes related to PD, and based on these studies, correlations were established between the expression levels of miRNAs and genes (Recabarren and Alarcón, 2017; Ren et al., 2019). Correlation does not allow establishing a direct dependence between miRNA and potential target genes. Consequently, the identification of such correlations does not enable us to establish specific relationships between miRNAs and target genes. Therefore, after over two decades of studying miRNA, no method has been developed for diagnosing various diseases using miRNA. Given the above circumstances, using the MirTarget program, we searched for associations of known human miRNAs with candidate PD genes. To confirm the reliability of these associations, it is necessary to show their presence in orthologous genes.

It is necessary to examine the expression of candidate genes in tissues affected by diseases. The level of miRNA expression in tissues with potential target genes and the possibility of delivering miRNA *via* blood to such tissues must be determined. Many studies have shown that miRNAs can circulate in the blood as components of exosomes and can enter almost any cell (Singh and Sen, 2017; Viswambharan et al., 2017; Wang et al., 2017; Rostamian Delavar et al., 2018; Titze-de-Almeida and Titze-de-Almeida, 2018; Tolosa et al., 2018; Yang et al., 2019; Ozdilek and Demircan, 2020; Wang and Zhang, 2020). Even in one tissue, the transfer of miRNA by diffusion was less effective than the transfer of miRNAs through the bloodstream to organ tissues. Bioinformatic approaches have been employed to identify associations between miRNAs and candidate genes (Zhang et al., 2019; Zhao and Wang, 2019). Such approaches enable us to study a substantial number of options for the interactions of known human miRNAs with all human protein-encoding genes. In this paper, we searched for miRNAs that bind to candidate PD genes to establish effective associations between miRNAs and genes that may be employed for the diagnosis of PD.

MATERIALS AND METHODS

The nucleotide (nt) sequences of candidate genes of PD were downloaded from the NCBI[1]. These specific candidate genes are shown in **Supplementary Table 1**. The nucleotide sequences of 2,565 miRNAs (old miRNAs) were obtained from miRBase, and 3,707 miRNAs (new miRNAs) were obtained from a previous study (Londin et al., 2015). The reads per kilobase million (RPKM) value (Mortazavi et al., 2008) was provided in the Human Protein Atlas data[2]. Orthologous genes of the following objects were used in the work: *Bos mutus* (bta), *Bubalus bubalis (bbu), Callithrix jacchus* (cja), *Capra hircus (chi), Delphinapterus leucas* (dle), *Felis catus (*fca*), Gorilla gorilla* (ggo), *Homo sapiens* (hsa), *Macaca fascicularis* (mfa), *Macaca mulatta* (mml), *Macaca nemestrina* (mne), *Mus musculus* (mmu), *Nomascus leucogenys* (nle), *Odobenus rosmarus divergens* **(**ord*), Orcinus orca* (oor), *Ovis aries (oar), Pongo abelii* (pab), *Papio anubis* (pab), *Pan paniscus* (ppa), *Panthera pardus (*ppr*), Pan troglodytes* (ptr), *Saimiri boliviensis* (sbo), and *Sus scrofa* (ssc). The miRNA BS in the 5′-untranslated region (5′UTR), coding sequence (CDS), and 3′-untranslated region (3′UTR) of several genes were predicted using the MirTarget program (Ivashchenko et al., 2014b, 2016). This program defines the following features of miRNA binding to mRNAs: (a) the start of the initiation of miRNA binding to mRNAs; (b) the localization of miRNA BSs in the 5′UTRs, CDSs and 3′UTRs of mRNAs; (c) the free energy of the interaction between miRNAs and mRNAs (ΔG, kJ/mole); and (d) the schemes of nucleotide interactions between miRNAs and mRNAs. The ratio ΔG/ΔGm (%) was determined for each site (ΔGm equals the free energy of miRNA binding with its fully complementary nucleotide sequence). The miRNA BSs located in mRNAs had ΔG/ΔGm ratios of 90% or more. The ΔG/ΔGm ratios were determined on the assumption that the members of one miRNA family generally differed by no more than one to three nucleotides, and along with a miRNA length of 22 nt, the ΔG/ΔGm value was determined to be 90% (20 nt/22 nt = 90%)±96% (21 nt/22 nt = 96%). With a larger difference in the number of mismatched nucleotides, the probability of two or more miRNAs binding to one site increases, despite the natural ability of miRNAs to interact selectively with the mRNAs of the target gene. The MirTarget program identifies the positions of the BSs on the mRNA, beginning with the first nucleotide of the mRNA's 5′UTR. The MirTarget program identifies hydrogen bonds between adenine (A) and uracil (U), guanine (G) and cytosine (C), G and U, A and C. The distance between A and C was 1.04 nanometers; the distance between G and C and between A and U was 1.03 nanometers; and the distance between G and U was 1.02 nanometers (Leontis et al., 2002). The numbers of hydrogen bonds in the G-C, A-U, G-U, and A-C interactions were 3, 2, 1, and 1, respectively (Kool, 2001; Lemieux and Major, 2002; Leontis et al., 2002; Garg and Heinemann, 2018). Taking into account the formation of non-canonical pairs significantly increases the reliability of establishing the interaction of miRNAs with mRNAs. The MirTarget program determines single miRNA BSs in mRNAs and miRNA BSs in clusters (arranged in series with overlapping nucleotide sequences of the same or several miRNAs). In this study, we suppose that the miRNA BSs in mRNAs were organized into clusters, which can be used as effective PD markers.

RESULTS

An analysis of the interactions between miRNAs and mRNAs was performed with candidate PD genes with an RPKM expression, considering the location of miRNA BSs in the 5′UTRs, CDSs and 3′UTRs. This approach enabled us to determine which miRNAs bound to different mRNA sites and which miRNAs preferred to interact with genes with different expression levels, since the results of the interactions between miRNAs and mRNAs are dependent on the ratio of the miRNA and mRNA concentrations. For example, the strong interaction of miRNAs with mRNAs slightly inhibits translation at miRNA concentrations that are tens of times lower than the mRNA concentrations. Conversely, the average interaction of miRNA with mRNA at substantially higher concentrations of miRNA over mRNA leads to significant suppression of translation. It is important to quantify the interactions between miRNAs and mRNAs to comparatively evaluate competition among miRNAs when they bind to mRNA.

Characteristics of the Interactions Between miRNAs and the 5′UTRs of mRNAs of Candidate PD Genes

Table 1 shows the data on the characteristics of the interactions between miRNAs and the mRNAs of the *GSK3B, PPARGC1A, ZFAND4*, and *CCNY* genes, depicting the cluster organization of the BSs of many miRNAs. The *GSK3B* gene serves as the potential target of 22 miRNAs, which distinguishes it from other candidate PD genes. The cluster of 22 miRNA BSs was located between the third and thirty-nine nucleotides (**Table 1**). The beginnings of these BSs were located over three nucleotides, which corresponded to their connection with the reading frame. In the mRNA of the *MANF* and other genes, paired miRNA BSs were also located over three nucleotides (**Supplementary Table 2**). The total length of the 22 BSs of *GSK3B* mRNA, considering multiple BSs, was 624 nt, which was 16 times greater than the length of the cluster. Gene *GSK3B* has BS for ID00296.3p-miR, ID00756.3p-miR, ID01804.3p-miR, ID02064.5p-miR with ΔG value more than −130 kJ/mole. ID01804.3p-miR, ID00457.3p-miR, ID00061.3p-miR, ID03151.3p-miR, ID02064.5p-miR, and miR-3960 have two BS, which significantly increased the effect of these miRNAs on the expression of the *GSK3B* gene.

Orthologous genes can be used as evidence of the reality of miRNA BS with the potential target gene mRNA. **Figure 1** shows the nucleotide sequences of the BS of several miRNAs included in the mRNA cluster orthologs of the *GSK3B* gene. The obtained results show that the nucleotide sequences of the

[1] http://www.ncbi.nlm.nih.gov

[2] https://www.proteinatlas.org/ENSG00000150093-ITGB1/tissue/

TABLE 1 | Characteristics of miRNA interactions with 5′UTR mRNAs of candidate PD genes.

Gene; RPKM	miRNA	Start of site, nt	ΔG, kJ/mole	ΔG/ΔGm, %	Length, nt
GSK3B; 8.3	ID02187.5p-miR	3	−123	89	23
	ID03229.5p-miR	4	−123	92	22
	ID01804.3p-miR	5, 12	−134	91	23
	ID00756.3p-miR	8	−123	89	23
	ID01041.5p-miR	8	−132	90	24
	ID02294.5p-miR	8	−127	87	24
	ID00457.3p-miR	8, 11	−123, −129	91, 95	22
	ID03367.5p-miR	11	−119	95	20
	ID00061.3p-miR	8,11	−125÷ −136	91÷98	22
	ID00296.3p-miR	9	−138	88	25
	ID01641.3p-miR	9	−127	86	24
	ID03151.3p-miR	9, 12	−115	93	20
	ID03229.5p-miR	10	−123	92	22
	ID01702.3p-miR	13	−140	93	24
	ID02064.5p-miR	10, 13	−129, −136	90, 94	23
	ID01873.3p-miR	11	−123	94	21
	miR-3960	11, 14	−115	92	20
	ID02522.3p-miR	12	−127	91	23
	ID02499.3p-miR	13	−119	92	21
	ID02429.3p-miR	14	−125	92	23
	ID01652.3p-miR	15	−125	89	23
	ID02538.3p-miR	15	−121	90	22
PPARGC1A; 2.5	ID00470.5p-miR	18÷47 (5)	−108÷ −110	89÷91	23
	[1.5] miR-574-5p	20÷31 (5)	−108 ÷ −113	89–93	23
	ID02299.5p-miR	30	−98	92	21
	ID02732.3p-miR	36, 42	−121	89	23
	ID03332.3p-miR	71	−134	90	24
	ID01310.3p-miR	135÷144 (4)	−121÷ −123	92÷94	22
	ID03332.3p-miR	143, 146	−134 −140	90, 94	24
	ID02761.3p-miR	149	−132	89	24
ZFAND4; 0.5	ID03418.3p-miR	109	−123	87	23
	ID00296.3p-miR	112	−134	85	25
	ID03206.5p-miR	114	−115	92	20
	ID01190.5p-miR	114	−136	89	24
	ID00030.3p-miR	114	−125	94	22
	ID02294.5p-miR	114	−125	86	24
	ID01574.5p-miR	116	−121	86	23
	ID01804.3p-miR	118	−125	86	23
	ID01702.3p-miR	118	−129	86	24
	ID03367.5p-miR	118	−113	90	20
	ID03073.3p-miR	128	−129	94	23
CCNY; 19.7	ID01041.5p-miR	1	−136	93	24
	ID01873.3p-miR	1	−123	94	21
	ID00296.3p-miR	4	−140	89	25
	ID01702.3p-miR	4	−134	89	24
	ID01641.3p-miR	4	−132	89	24
	ID01106.5p-miR	7	−132	89	24
	ID01879.5p-miR	8	−129	95	22
	ID02229.3p-miR	9	−121	92	21
	ID02499.3p-miR	9	−123	95	21
	ID03027.3p-miR	11	−121	85	24

In ***Table 1*** *and below, the number of miRNA binding sites is indicated in oval brackets. The value of miRNAs RPKM is indicated in square brackets.*

clusters decrease from 33 nt in the hsa-mRNA of the *GSK3B* gene to 22 nt in the ptr-mRNA. Therefore, starting from ptr-mRNA, the cluster contains miRNAs BS of 21 nt or more, which can bind miRNAs of orthologous genes. Note that changes in the nucleotide composition of BS occur according to the principle of replacement of purine for purine (A ↔ G), or pyrimidine for pyrimidine (U ↔ C). With such substitutions, non-canonical G-U and A-C pairs are formed, or the canonical G-C and A-U pairs are formed (**Figure 1**). Clusters of miRNA BS in the mRNA of all objects are located between the conserved oligonucleotides UGCGGG and CCGAG. All cluster regions in orthologous genes of *GSK3B* include the same pentanucleotide CGGGC.

Figure 2 shows the location of the miRNA BS within the cluster, which demonstrate competition between miRNAs when they bind in the mRNA cluster of the *GSK3B* gene. Binding of any of the miRNAs in the cluster interferes with the binding of other miRNAs.

The efficiency of miRNA binding in a cluster is shown on schemes in **Figure 3**. Due to the formation of non-canonical A-C and G-U pairs, the structure of the miRNA-mRNA complex has a double-stranded helix without the formation of "bubbles," which increases the free binding energy of RNA strands due to stacking interactions.

The mRNA of the *PPARGC1A* gene contains two clusters of miRNA BSs, from 18 to 70 nt and from 135 to 172 nt (**Table 1**). Both clusters contain the BSs of several miRNAs with multiple BSs; that is, several of their BSs for one miRNA are located sequentially over two to three nucleotides. For example, the start of the miR-574-5p and ID00470.5p-miR have five and eight BSs, respectively, that are located over two nucleotides. In the second cluster for ID01310.3p-miR and ID03332.3p-miR four and five BSs respectively. Gene *PPARGC1A* has two BSs for ID03332.3p-miR and one for ID02761.3p-miR with a ΔG value greater than −130 kJ/mole. The association of the *PPARGC1A* gene with these miRNAs can be used as a marker for the diagnosis of PD.

With weak gene expression (value RPKM is 10), it is highly probable that several miRNAs can strongly suppress the synthesis of the corresponding proteins and have a decisive influence on the manifestation of their function. In addition, the presence of multiple BSs for the miRNAs ID00470.5p-miR, miR-574-5p, and ID01310.3p-miR in the mRNA of certain genes, such as *PPARGC1A*, significantly increases the probability that their expression will be suppressed.

There are clusters of miRNA BS in the mRNA of orthologous genes of monkeys (**Figure 4**). The first cluster, 53 nt long, is highly conserved (**Figure 4A**), while the second cluster differs from species to species (**Figure 4B**).

In the first and second clusters, the flanking oligonucleotides from 5-end (GCUCUGC and UUGAGAA) and 3-end (GGCACAG and GCAUCC) are conserved.

There was a cluster in the mRNA of the *ZFAND4* gene containing the BSs for 11 miRNAs (**Table 1** and **Figure 5**). The nucleotide sequences of clusters of mRNA BSs orthologous gene and flanking sequences were highly conserved. Gene *ZFAND4* has miRNA BS for ID00296.3p-miR, ID01190.5p-miR with a ΔG value of more than −130 kJ/mole (**Table 1**). This association is recommended for use in PD diagnosis.

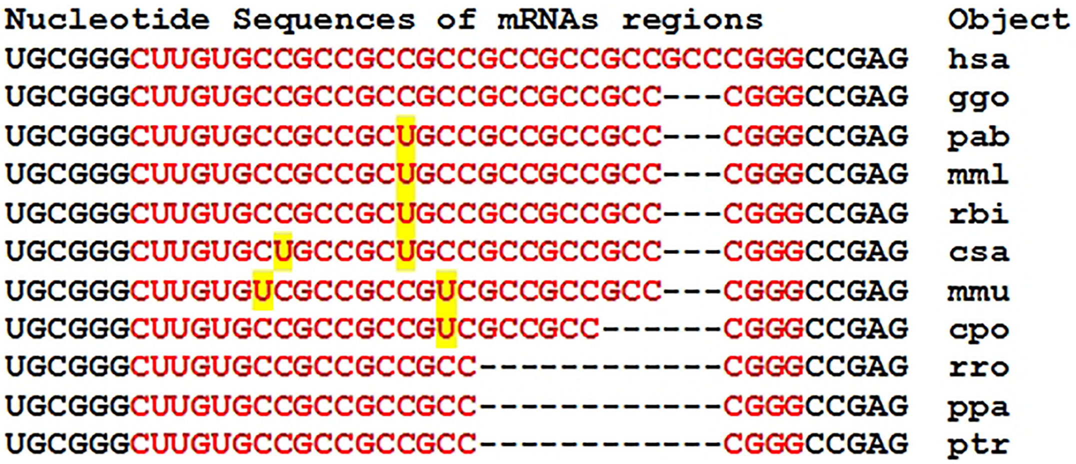
The miRNA BSs in the mRNA of some genes formed clusters in which these sites featured partially overlapping nucleotides. The mRNA of the *CCNY* gene contained two clusters (**Table 1**). The first cluster from 1 to 30 nt included the BSs of nine miRNAs with a total length equal to 206 nt, which was 6.9 times greater than the length of the cluster. The organization of the miRNA BSs into clusters has the following consequences. The length of the 5′UTR is 180 nt, and the BSs of nine miRNAs with a length of 206 nt cannot be sequentially located in the 5′UTR without nucleotide overlap. Therefore, compaction of the miRNA BSs is necessary. However, the compaction of the BSs leads to competition between miRNAs for binding to a 30-nt region in which only one miRNA can bind. In this case, miRNA predominantly binds the mRNA with the highest free interaction energy. For example, ID01041.5p-miR, ID00296.3p-miR, ID01702.3p-miR, ID01641.3p-miR, and ID01106.5p-miR preferably bind to the mRNA of the *CCNY* gene. In addition, it must be considered that the concentration of each miRNA can differ by a factor of several tens, and as a result, the quantitative characteristics of the interactions of mRNAs with different miRNAs in combination with their concentrations determines the duration of the miRNA complex with mRNA. For this reason, it is necessary to control the concentration of all miRNAs and mRNAs, which results in the miRNA determining the primary inhibition of translation. The second cluster of BSs for ID02971.3p-miR, ID02128.5p-miR, and ID01976.5p-miR in the mRNA of the *CCNY* gene had a smaller compaction of 1.4-fold. However, in this case, competition also was observed among the three miRNAs for binding to mRNA.

Given in **Figure 6**, the nucleotide sequences of clusters of BS in the mRNA of the orthologous *CCNY* genes are flanked by conserved oligonucleotides UGGCG and CCGGC.

Of the 15 candidate genes with RPKM values less than 10, eleven genes each had a miRNA BSs in the 5′UTR (**Supplementary Table 2**). The *KANSL1* gene was the potential target of two miRNAs, and the *CRHR1* and *ERBB2* genes were the potential targets of three miRNAs. In the mRNA of the *LRP10* gene, the ID03064.3p-miR and ID01106.5p-miR BSs formed a cluster. The mRNA of the *LRP10* gene contains a cluster from 406 to 434 nt, 28 nt long (**Figure 7**). The flanking pentanucleotides GCGCC and CCGGC are the same in all objects.

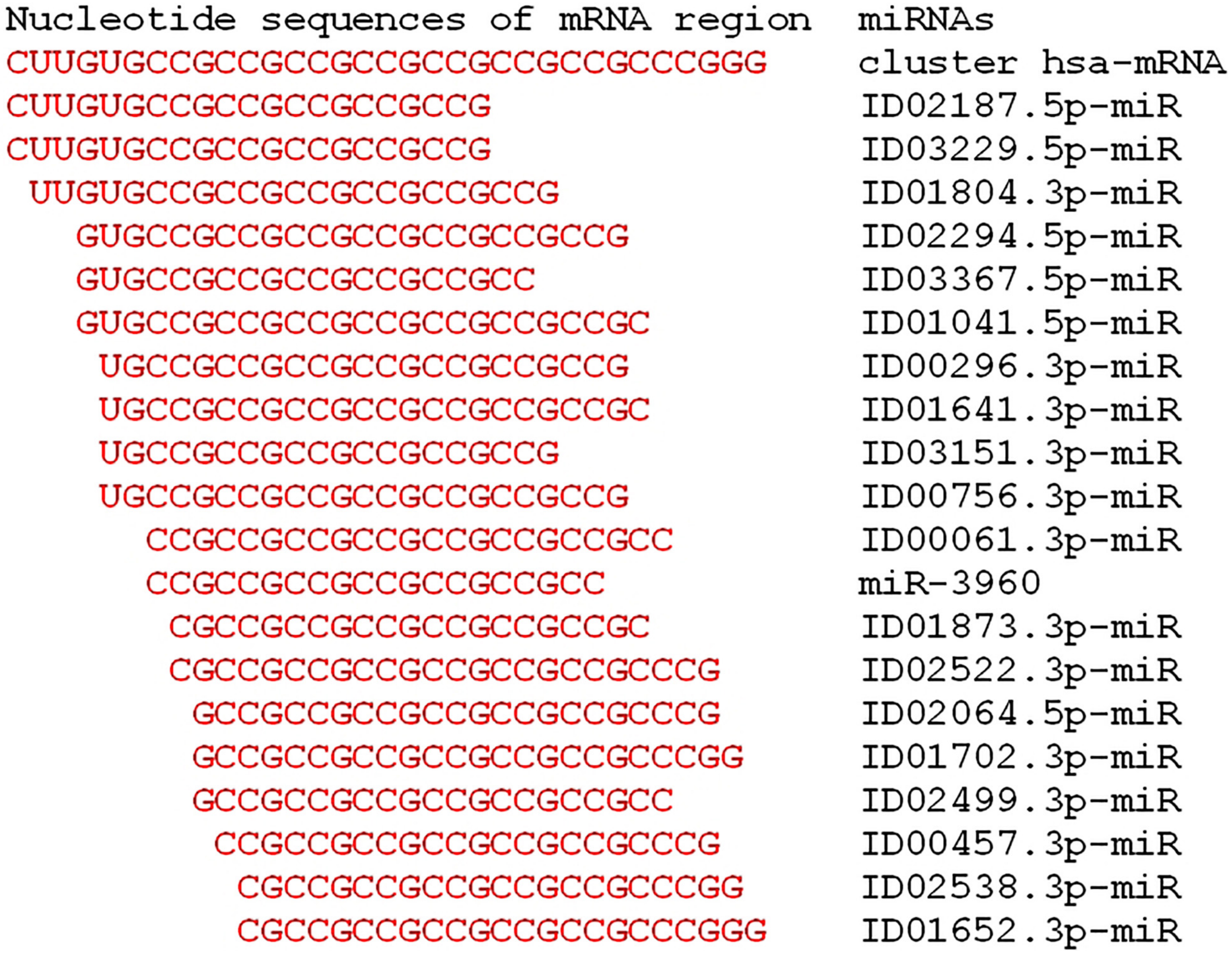
In the mRNA of the *MANF* gene, two miRNAs had sites in the cluster from 56 to 97 nt (**Supplementary Table 2**). The total length of the BSs of two miRNAs was 69 nt, which was 1.7 times greater than the length of the cluster. Therefore, when organizing BSs into clusters, the sites were compacted to reduce the length of the 5′UTR. Another consequence of this compaction is the emergence of competition among miRNAs for binding to mRNA, since only one miRNAs can interact with a 41-nt-long cluster.

The *RAB5A* gene is the potential target of six miRNAs, the BSs of which form three clusters (**Supplementary Table 2**). ID03445.3p-miR has two BSs with overlapping nucleotide sequences, which increases the likelihood of its interaction with the mRNA of the *RAB5A* gene. Compared with other miRNAs, the association of ID02930.3p-miR has a large free energy of interaction with the mRNA of the *RAB5A* gene and can be recommended as a marker for PD diagnosis.



```
Nucleotide Sequences of mRNAs regions        Object
UGCGGGCUUGUGCCGCCGCCGCCGCCGCCGCCCGGGCCGAG    hsa
UGCGGGCUUGUGCCGCCGCCGCCGCCGCC---CGGGCCGAG    ggo
UGCGGGCUUGUGCCGCCGCUGCCGCCGCC---CGGGCCGAG    pab
UGCGGGCUUGUGCCGCCGCUGCCGCCGCC---CGGGCCGAG    mml
UGCGGGCUUGUGCCGCCGCUGCCGCCGCC---CGGGCCGAG    rbi
UGCGGGCUUGUGCUGCCGCUGCCGCCGCC---CGGGCCGAG    csa
UGCGGGCUUGUGUCGCCGCCGUCGCCGCC---CGGGCCGAG    mmu
UGCGGGCUUGUGCCGCCGCCGUCGCC------CGGGCCGAG    cpo
UGCGGGCUUGUGCCGCCGCC------------CGGGCCGAG    rro
UGCGGGCUUGUGCCGCCGCC------------CGGGCCGAG    ppa
UGCGGGCUUGUGCCGCCGCC------------CGGGCCGAG    ptr
```

FIGURE 1 | Nucleotide sequences of 5′UTR regions of mRNAs of orthologous *GSK3B* genes containing clusters of miRNAs binding sites.



```
Nucleotide sequences of mRNA region   miRNAs
CUUGUGCCGCCGCCGCCGCCGCCGCCGCCCGGG     cluster hsa-mRNA
CUUGUGCCGCCGCCGCCGCCG                 ID02187.5p-miR
CUUGUGCCGCCGCCGCCGCCG                 ID03229.5p-miR
 UUGUGCCGCCGCCGCCGCCGCCG              ID01804.3p-miR
   GUGCCGCCGCCGCCGCCGCCGCCG           ID02294.5p-miR
   GUGCCGCCGCCGCCGCCGCC               ID03367.5p-miR
   GUGCCGCCGCCGCCGCCGCCGCCGC          ID01041.5p-miR
    UGCCGCCGCCGCCGCCGCCGCCG           ID00296.3p-miR
    UGCCGCCGCCGCCGCCGCCGCCGC          ID01641.3p-miR
    UGCCGCCGCCGCCGCCGCCG              ID03151.3p-miR
    UGCCGCCGCCGCCGCCGCCGCCG           ID00756.3p-miR
      CCGCCGCCGCCGCCGCCGCCGCC         ID00061.3p-miR
      CCGCCGCCGCCGCCGCCGCC            miR-3960
       CGCCGCCGCCGCCGCCGCCGC          ID01873.3p-miR
       CGCCGCCGCCGCCGCCGCCGCCCG       ID02522.3p-miR
        GCCGCCGCCGCCGCCGCCGCCCG       ID02064.5p-miR
        GCCGCCGCCGCCGCCGCCGCCCGG      ID01702.3p-miR
        GCCGCCGCCGCCGCCGCCGCC         ID02499.3p-miR
         CCGCCGCCGCCGCCGCCGCCCG       ID00457.3p-miR
          CGCCGCCGCCGCCGCCGCCCGG      ID02538.3p-miR
          CGCCGCCGCCGCCGCCGCCCGGG     ID01652.3p-miR
```

FIGURE 2 | Schemes of the location of miRNAs binding sites in the cluster located in the 5′UTR of the mRNA of the candidate hsa-*GSK3B* gene of Parkinson’s disease.

```
miRNA, position of the miRNA binding site    miRNA, position of the miRNA binding site
```

```
ID02187.5p-miR,4
5'-GCCGCCGCCGCCGCCGCCGCCGC-3'
   |  || |||||||||||||||||
3'-CCUCAUUGGCGGCGGCGGCGGCG-5'
```

```
ID03229.5p-miR,4
5'-GCUUGUGCCGCCGCCGCCGCCG-3'
   | | ||||||||||||||||||
3'-CUACCACGGCGGCGGCGGCGGC-5'
```

```
ID01804.3p-miR,5
5'-UGCCGCCGCCGCCGCCGCCGCCG-3'
   ||  |||||||||||||||||||
3'-GCCCCGGCGGCGGCGGCGGCGGC-5'
```

```
ID02294.5p-miR,8
5'-GUGCCGCCGCCGCCGCCGCCGCCG-3'
   ||| | ||||||||||||||||||
3'-CGCCGAGGCGACGGCGACGGCGGC-5'
```

```
ID00756.3p-miR,8
5'-UGCCGCCGCCGCCGCCGCCGCCG-3'
     |||| ||||||||||||||||
3'-CAGGCAUCAGCGGCGGCGGCGGC-5'
```

```
ID01641.3p-miR,9
5'-UGCCGCCGCCGCCGCCGCCGCCGC-3'
   | ||||| ||||||||||||||||
3'-GGGGUGGGGGCGGCGGCGGCGGUG-5'
```

```
ID00296.3p-miR,9
5'-UGCCGCCGCCGCCGCCGCCGCCGCC-3'
   | ||||| |||||||||||||||||
3'-GGGGUGGGGGCGGCGGCGGCGGCGG-5'
```

```
ID03151.3p-miR,9
5'-UGCCGCCGCCGCCGCCGCCG-3'
   |||| |||||||||||||||
3'-AUGGGGGCGGCGGCGGCGGC-5'
```

```
ID01702.3p-miR,12
5'-CGCCGCCGCCGCCGCCGCCGCCCG-3'
   | |||||||||||||||||||||
3'-GGGGCGGCGGCGGCGGCGGCGGGA-5'
```

```
ID01873.3p-miR,12
5'-CGCCGCCGCCGCCGCCGCCGC-3'
   || | ||||||||||||||||
3'-GCCAGGGCGGCGGCGGCGGCG-5'-5'
```

```
ID02064.5p-miR,13
5'-GCCGCCGCCGCCGCCGCCGCCCG-3'
   |||||||||||||||||| ||||
3'-CGGCGGCGGCGGCGGCGGGAGGC-5'
```

```
ID00457.3p-miR,14
5'-CCGCCGCCGCCGCCGCCGCCCG-3'
     |||||||||||||||||||
3'-CUCGGCGGCGGCGGCGGCGGGA-5'
```

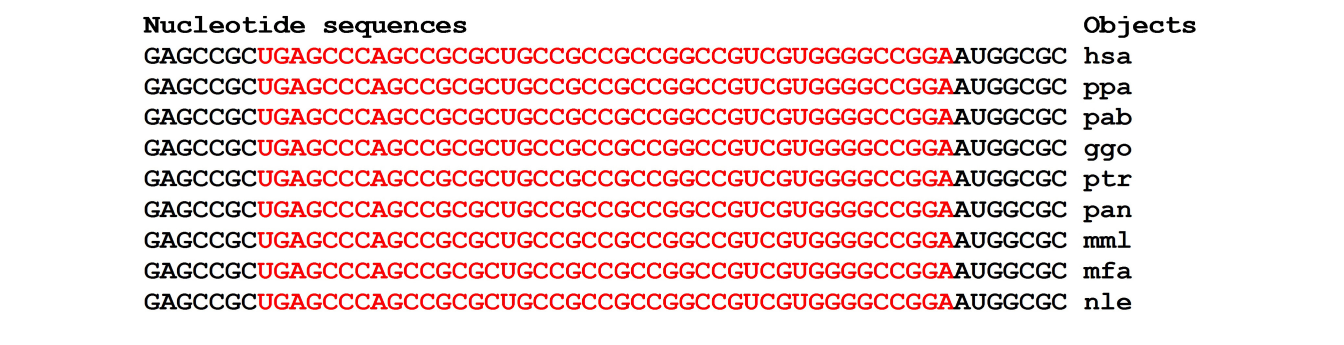
```
ID02499.3p-miR,14
5'-CCGCCGCCGCCGCCGCCGCCC-3'
   |||||||||||||||| ||||
3'-GGCGGCGGCGGCGGUGCUGGG-5'
```

```
ID02538.3p-miR,15
5'-CGCCGCCGCCGCCGCCGCCCGG-3'
   |||||||||||||||| |||||
3'-GCGGCGGCGACGGCGGGAGGCU-5'
```

FIGURE 3 | Schemes of miRNA interaction in cluster 5′UTR mRNA of the candidate *GSK3B* gene for Parkinson's disease.

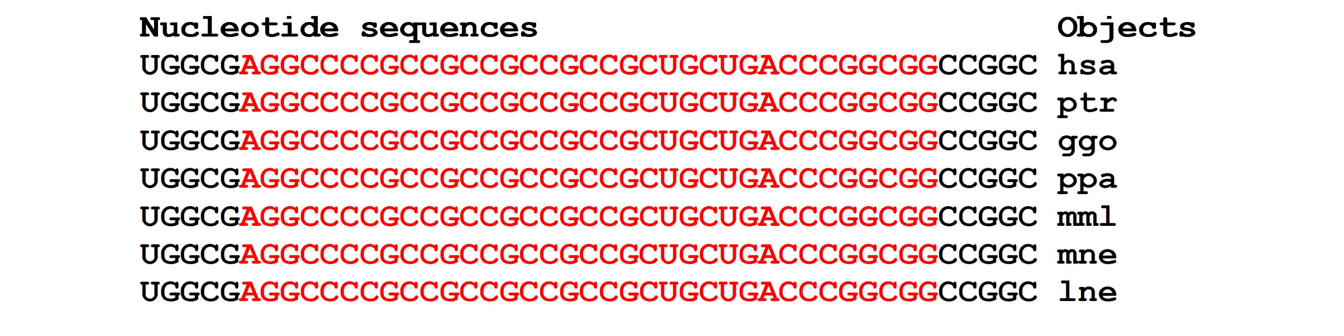
```
A
Nucleotide sequences                                                 Objects
GCUCUGCGCGCACACACCACACACACGCACACGCACACACACGCGCGCACACACGCAGCCGGCACAG hsa
GCUCUGCGCGCACACACCACACACACGCACACGCACACACACGCGCGCACACACGCAGCCGGCACAG pab
GCUCUGCGCGCACACACCUCACACACGCACACGCACACACACGCGCGCACACACGCAGCCGGCACAG pan
GCUCUGCGCGCACACACCUCACACACGCACACGCACACACACGCGCGCACACACGCAGCCGGCACAG mml
GCUCUGCGCGCACACACCUCACACACGCACACGCACACACACGCGCGCACACACGCAGCCGGCACAG mne


B
Nucleotide sequences                                          Objects
UUGAGAAGGCGGCAGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCAGCCCGAGCAUCC pan
UUGAGAAGGCGGCAGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCAGCCCGAGCAUCC ppa
UUGAGAAGGCGGCAGCGGCGGCGGCGGCGGCGGCGGCGGC---AGCCCGAGCAUCC ggo
UUGAGAAGGCGGCAGCGGCGGCGGCGGCGGCGGCGGC------AGCCCGAGCAUCC hsa
UUGAGAAGGCGGCAGCGGCGGCGGCGGUGGCGGCGGC------AGCCCGAGCAUCC pab
UUGAGAAGGCGGCAGCGGCGGCAGCGGCGGCGGCGGC------AGCCCGAGCAUCC mml
UUGAGAAGGCGGCAGCGGCGGCAGCGGCGGCGGCGGC------AGCCCGAGCAUCC mne
```

FIGURE 4 | Nucleotide sequences of 5′UTR regions of mRNAs of orthologous *PPARGC1A* genes containing the first cluster **(A)** and the second cluster **(B)** of miRNAs binding sites.



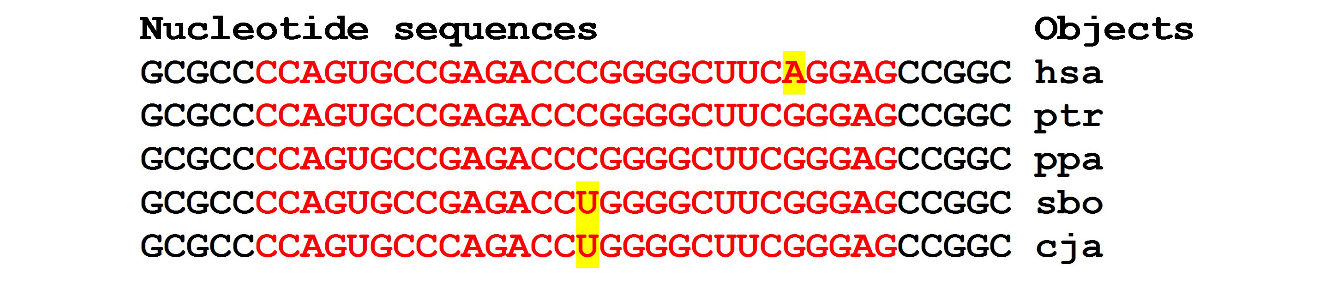
Nucleotide sequences	Objects
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	hsa
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	ppa
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	pab
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	ggo
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	ptr
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	pan
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	mml
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	mfa
GAGCCGCUGAGCCCAGCCGCGCUGCCGCCGCCGGCCGUCGUGGGGCCGGAAUGGCGC	nle

FIGURE 5 | Nucleotide sequences of 5′UTR regions of mRNAs of orthologous ZFAND4 genes containing a cluster of miRNAs binding sites.



Nucleotide sequences	Objects
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	hsa
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	ptr
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	ggo
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	ppa
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	mml
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	mne
UGGCGAGGCCCCGCCGCCGCCGCCGCUGCUGACCCGGCGGCCGGC	lne

FIGURE 6 | Nucleotide sequences of 5′UTR regions of mRNAs of orthologous *CCNY* genes containing clusters of miRNAs binding sites.



Nucleotide sequences	Objects
GCGCCCCAGUGCCGAGACCCGGGGCUUCAGGAGCCGGC	hsa
GCGCCCCAGUGCCGAGACCCGGGGCUUCGGGAGCCGGC	ptr
GCGCCCCAGUGCCGAGACCCGGGGCUUCGGGAGCCGGC	ppa
GCGCCCCAGUGCCGAGACCUGGGGCUUCGGGAGCCGGC	sbo
GCGCCCCAGUGCCCAGACCUGGGGCUUCGGGAGCCGGC	cja

FIGURE 7 | Nucleotide sequences of 5′UTR regions of mRNAs of orthologous *LRP10* genes containing clusters of miRNAs binding sites.

In 20 genes with low expression levels, 14 miRNA and mRNA associations were identified (**Table 1** and **Supplementary Table 2**). Six miRNA and mRNA associations were identified in 15 genes with high expression in the 5′UTR mRNA. Fourteen associations between miRNAs [ID00061.3p-miR, ID00296.3p-miR, ID01041.5p-miR, ID01106.5p-miR, ID01190.5p-miR, ID01702.3p-miR, ID01804.3p-miR, ID02064.5p-miR, ID02761.3p-miR, ID03047.3p-miR, ID03064.3p-miR (two sites), and ID03332.3p-miR] and the 5′UTRs of mRNAs of candidate PD target genes (*ERBB2, GSK3B, LRCH1, LRP10, PPARGC1A, ZFAND4,* and *CCNY*) have free energy interactions of more than −130 kJ/mole (**Table 1** and **Supplementary Table 2**) and are recommended as markers for PD.

Of the 15 candidate genes, seven had one BS for different miRNAs (**Table 1** and **Supplementary Table 2**). Each of the *CTNNB1, MANF, MAPT, RTN1,* and *VSNL1* genes were targets of two miRNAs (**Supplementary Table 2**).

A small number of BSs with the formation of clusters (**Supplementary Table 2**) is characteristic of most genes with high expression in addition to those shown in **Table 1**. Only in the mRNA of the *CDK5R1, MART,* and *VSNL1* genes were the clusters of two miRNA BSs identified. The free energy of the interactions of these miRNAs with mRNAs of candidate PD genes was a ΔG value below −130 kJ/mole. In 13 genes with high expression, there were no such associations (**Supplementary Table 2**).

Characteristics of the Interactions Between miRNAs and the CDSs of mRNAs of Candidate PD Genes

In the CDSs of the mRNAs of six genes, there was one BS, and in the mRNAs of nine genes, there were two BSs (**Supplementary Table 3**). Only in the mRNAs of the *AXIN1, CD5,* and *ERBB2* genes were clusters of BSs for two miRNAs detected. The *FOXO1* gene was the potential target of seven miRNAs, the BSs of which were distributed across two clusters (**Table 2**). From 655 to 695 nt, there were five BSs with a total length of 137 nt, which was 3.3 times greater than the length of this cluster. The value of the free energy of the interactions between the miRNAs and mRNAs for the three associations of miRNAs and mRNAs of *FOXO1* was above −130 kJ/mole.

The nucleotide sequences of mRNAs binding site clusters of orthologous *FOXO1* genes are highly conserved (**Figure 7**). The high GC-content of miRNAs BS determines the high free energy of miRNAs binding to mRNAs of the *FOXO1* gene. The oligonucleotides flanking the clusters are conserved.

Clusters of miRNA BSs were identified only in the mRNAs of the orthologous *APOE* genes. The nucleotide sequences of the cluster and the oligonucleotides flanking them were conserved (**Table 2** and **Figure 8**). The protein regions encoded by BS of ID01030.3p-miR and ID03261.3p-miR in mRNA orthologous *APOE* genes were conserved, respectively (**Supplementary Figure 1**).

The miRNA BS in the mRNA of 11 orthologous *FOXO1* genes of some mammals formed clusters encoding longer oligonucleotides (**Figure 9**). However, the flanking amino acid sequences from the C-terminus were identical and from the N-terminus differed by one amino acid (**Supplementary Figure 2**).

The clusters of miRNAs BS in mRNAs of the *FOXO1* gene encoded the same oligopeptides AAAVAAAAAAAA, with the exception of the nle-mRNAs of the *FOXO1* gene (**Supplementary Figure 2**). The amino acid sequences flanking the oligopeptides encoded by the cluster of BS were completely conserved.

The miRNA BS in the mRNA of 11 orthologous *FOXO1* genes of some mammals formed clusters encoding longer oligonucleotides (**Supplementary Figure 3**). However, the flanking amino acid sequences from the C-terminus were identical and from the N-terminus differed by one amino acid.

Among the associations of miRNAs and mRNA of *FOXO1*, ID02761.3p-miR, ID03332.3p-miR, ID01804.3p-miR stand out, which are recommended as markers of PD as interacting with a ΔG value of more than −130 kJ/mole (**Table 2**).

The mRNA of the *SETD1A* gene contained 17 miRNA BSs (**Table 2**). ID03324.3p-miR and ID01641.3p-miR each had two BSs in different clusters. The cluster of BSs from 4,877 to 4,928 nt was four times less than the total length of miRNA BSs, which was 205 nt (**Figure 10**). Highly conserved nucleotide sequences of cluster encode conserved amino acids in orthologous proteins (**Supplementary Figure 4**). Six associations between miRNAs (ID00296.3p-miR, ID01641.3p-miR, ID01702.3p-miR target genes *SETD1A* have a free energy of the miRNA interaction with the CDS of more than −130 kJ/mole (**Table 2**) and are recommended as markers for PD. In 34 genes with low expression, only two genes with clusters of BSs had six associations with a ΔG value higher than −130 kJ/mole.

Each mRNA of the *ATN1* and *ATP13A2* genes had BSs for seven miRNAs (**Supplementary Table 3**). Only in the mRNA of the *ATP13A2* gene was a cluster of two BSs for ID01157.5p-miR and ID01377.3p-miR identified. Therefore, in the CDSs of mRNAs of the candidate PD genes, there were no clusters of BSs of more than two miRNAs. BS of ID01047.3p-miR is conserved in the mRNA of *ATN1* orthologous genes (**Figure 11**). Corresponding amino acid sequences were conserved along with flanking oligopeptides (**Supplementary Figure 5**). Of the 16 genes with high expression in the protein-encoding region, no miRNA BSs with free interaction energies higher than −130 kJ/mole were found.

Characteristics of miRNA Interactions With 3′UTRs of mRNAs of Candidate PD Genes

Each of the mRNAs of ten candidate genes bound to only one miRNA (**Supplementary Table 4**). The mRNAs of the *LRP10, PRKN, RBBP5,* and *SLC14A1* genes were potential targets for two or three miRNAs, containing a cluster of miR-5095 and miR-619-5p BSs located six nucleotides apart (**Table 3** and **Supplementary Table 4**). The mRNA of the *GSK3B* gene, in addition to the miRNA BSs in the 5′UTR, contained miRNA BSs in the 3′UTR (**Table 3**), which made up the cluster of BSs for miR-466, ID01030.3p-miR, and ID00436.3p-miR, and together with ID01727.5p-miR, these miRNAs could bind to the mRNA of the *PPARGC1A* gene (**Table 3**).

The data shown in **Figure 12** indicate a difference in the size of the cluster of miRNAs BS in the 3′UTR of mRNA of *GSK3B* orthologous genes. At the same time, the flanking oligonucleotides remain highly conserved. These results prove the emergence of a connection between miRNA and mRNA of target genes many millions of years ago. The organization of miRNA BS into clusters also has a long history. The existing changes in the nucleotide composition of BS occur according to the principle of replacement of purine for purine (A↔G), or pyrimidine for pyrimidine (U↔C). Such substitutions result in non-canonical pairs G-U and A-C. MirTarget takes these interactions into account and predicts the formation of these pairs.

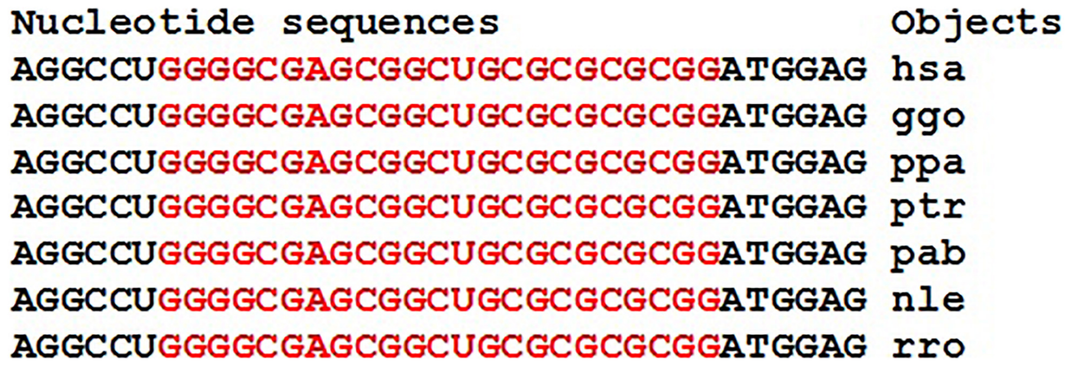
Note that the preservation of the oligonucleotide composition of the cluster-flanking BS in the 3′UTR of orthologous genes during evolution is probably necessary to preserve the interactions of miRNAs with mRNAs. Note that the flanking nucleotides contain the same CUUGGU hexanucleotides (**Supplementary Figure 7**).

The *LRP10* gene is the potential target of nine miRNAs. The miR-5095 and miR-619-5p BSs form a cluster, and the beginnings of their BSs differed by six nucleotides. This relationship between miR-5095 and miR-619-5p BSs is not accidental, since the identical arrangement of their BSs was determined in the mRNAs of the *PRKN* (**Supplementary Table 4**), *RBBP5,* and *SLC14A1* genes (**Table 3** and **Figure 13**).

In addition, the difference between the miR-5096 and miR-619-5p BSs was the same (74 nt) in the mRNAs of the *PDP2, RBBP5,* and *SLC14A1* genes. The beginnings of the miR-5585-3p and miR-1285-5p BSs differed by 99 nt in the mRNAs of the *LRP10* and *PDP2* genes. The miR-619-5p bound to the mRNA of the *PRKN* gene fully complementarily among the 201 genes that are the target of this miRNA (Atambayeva et al., 2017). Candidate

TABLE 2 | Characteristics of the interactions between miRNAs and CDS mRNAs of candidate PD genes.

Gene; RPKM	miRNA	Start of site, nt	ΔG, kJ/mole	ΔG/ΔGm, %	Length, nt
APOE; 269.2	ID03402.5p-miR	758	−121	95	22
	ID03398.5p-miR	881	−115	93	20
	ID03261.5p-miR	883	−115	93	20
FOXO1; 2.5	ID03332.3p-miR	655, 658	−136, −140	91, 94	24
	ID02761.3p-miR	661	−132	89	24
	ID02611.3p-miR	660	−125	91	22
	ID00171.3p-miR	666	−117	93	20
	ID01804.3p-miR	672	−136	93	23
	ID01057.5p-miR	745	−123	91	23
	ID02429.3p-miR	749	−123	91	23
SETD1A; 4.5	miR-6824-5p	2,062	−113	90	22
	[0.2] miR-1207-5p	2,064	−115	93	21
	ID00850.3p-miR	2,495	−117	90	22
	ID01321.5p-miR	2,498	−113	91	21
	[19.1] miR-762	4,098	−125	92	22
	miR-6891-3p	4,759	−106	93	21
	ID03324.3p-miR	4,764, 4,788	−115	90	22
	ID03238.3p-miR	4,769	−117	90	23
	ID01545.3p-miR	4,776	−113	93	21
	ID02538.3p-miR	4,877	−121	90	22
	ID01641.3p-miR	4,894, 4,900	−132, −140	89, 94	24
	ID01323.3p-miR	4,898	−123	91	22
	miR-3960	4,899	−115	92	20
	ID00296.3p-miR	4,900	−140	89	25
	ID01702.3p-miR	4,900	−134	89	24
	ID01959.3p-miR	4,905	−117	92	21
	ID00962.3p-miR	4,905	−117	89	23

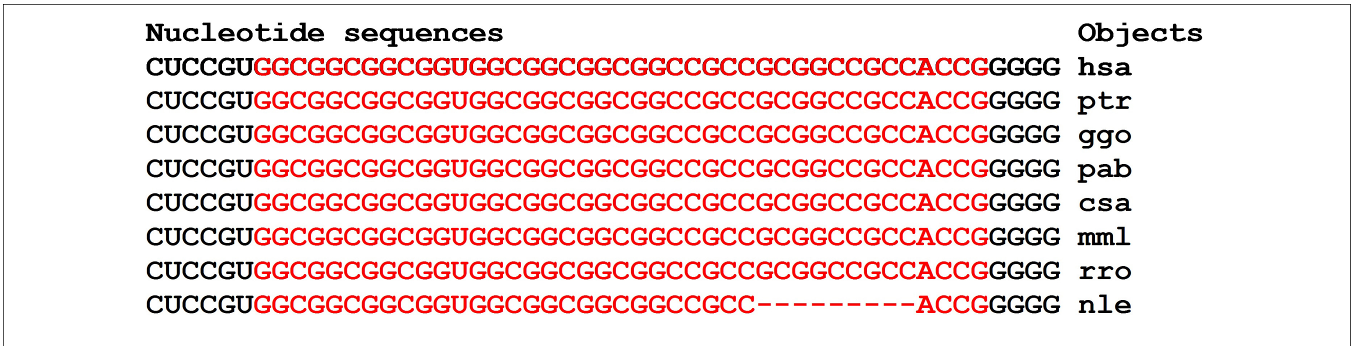


Nucleotide sequences	Objects
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	hsa
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	ggo
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	ppa
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	ptr
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	pab
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	nle
AGGCCUGGGGCGAGCGGCUGCGCGCGGATGGAG	rro

FIGURE 8 | Nucleotide sequences of the CDSs mRNAs regions of orthologous *APOE* genes containing binding sites of ID01030.3p-miR and ID03261.3p-miR.

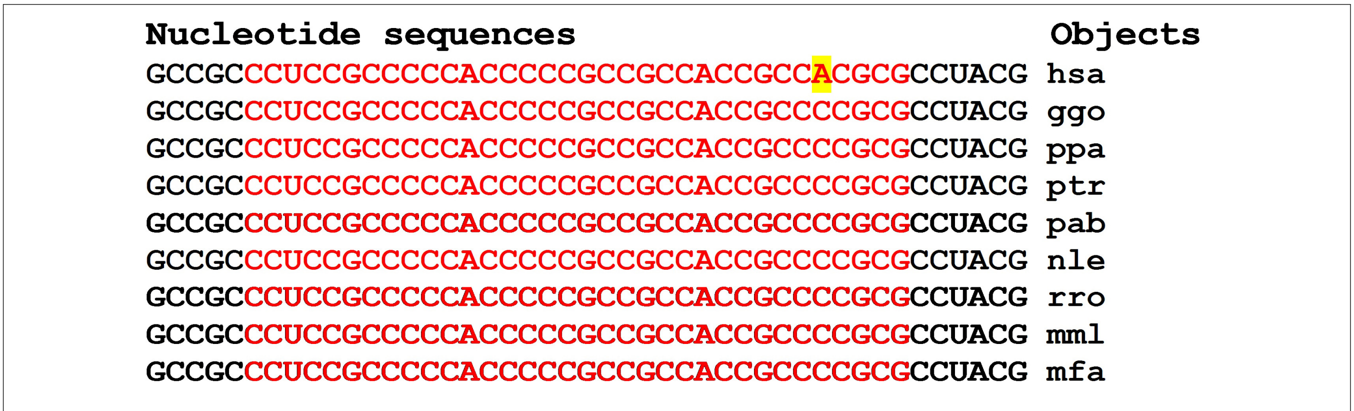
PD genes that are targets of miRNAs that bind to the 3′UTRs of mRNAs significantly differ from other candidate genes in the number of BSs for miR-619-5p, miR-5095, miR-5096, miR-5585-3p, and miR-1285-5p. Another feature of candidate PD genes is the association of the *GSK3B* and *PPARGC1A* genes with miR-466, ID00436.3p-miR, ID01030.3p-miR, and ID01727.5p-miR, the BSs of which form one cluster (**Table 3**). The interactions between miRNAs and the 3′UTRs of mRNAs occur with less free energy than those between miRNAs and the 5′UTRs and CDSs of mRNAs because these miRNAs have reduced GC contents. For example, only ID02732.3p-miR was associated with the mRNA of the *PRKN* gene, exhibiting a ΔG value of −132 kJ/mole.

The cluster of miR-5095 and miR-619-5p BS in the mRNA of the *SLC14A1* gene is conserved in part of miR-619-5p binding (GGCUCACACCUGUAAUCCCAGC) and is variable in the six nucleotide segment that binds to miR-5095 (**Figure 13**). Flanking nucleotides from the 3-end of the cluster are the same for all objects (ACUUUGGG) and coincide with the flanking nucleotides of the 3-end of the cluster in the mRNA of the *PRKN* gene of most objects (**Figure 14**).



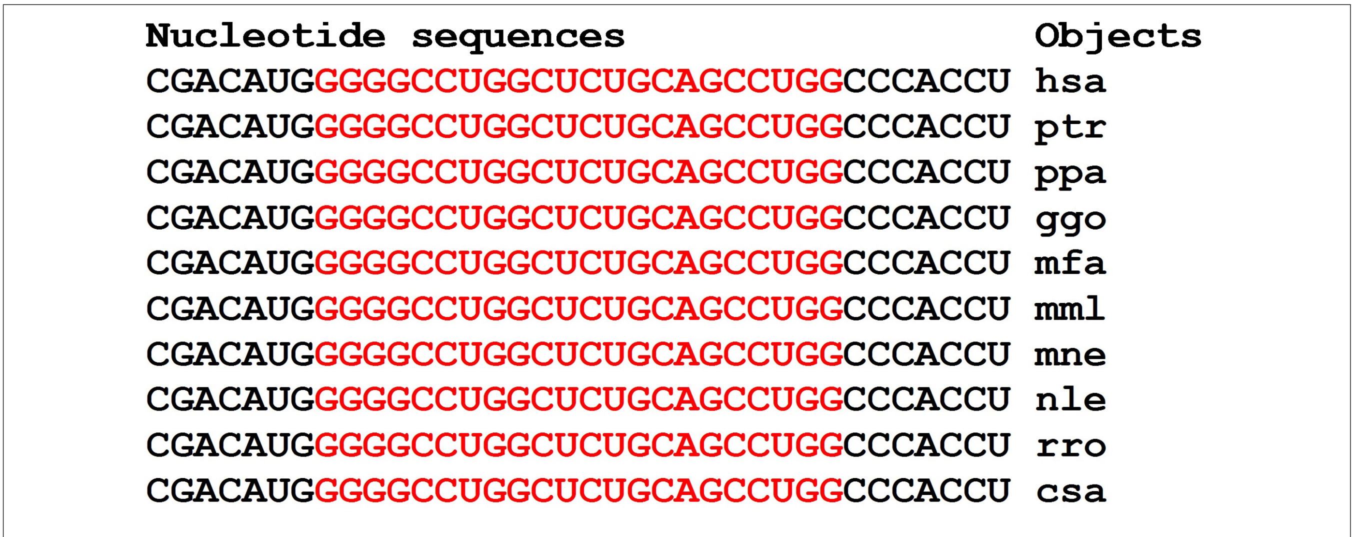
```
Nucleotide sequences                                      Objects
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG hsa
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG ptr
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG ggo
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG pab
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG csa
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG mml
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCCGCGGCCGCCACCGGGGG rro
CUCCGUGGCGGCGGCGGUGGCGGCGGCGGCCGCC---------ACCGGGGG nle
```

FIGURE 9 | Nucleotide sequences of CDS regions of mRNAs of orthologous *FOXO1* genes containing clusters of miRNAs binding sites.



```
Nucleotide sequences                            Objects
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCACGCGCCUACG hsa
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG ggo
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG ppa
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG ptr
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG pab
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG nle
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG rro
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG mml
GCCGCCCUCCGCCCCCACCCCCGCCGCCACCGCCCCGCGCCUACG mfa
```

FIGURE 10 | Nucleotide sequences of the CDSs mRNAs regions of orthologous *SETD1A* genes containing clusters of miRNAs BSs.



```
Nucleotide sequences               Objects
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU hsa
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU ptr
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU ppa
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU ggo
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU mfa
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU mml
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU mne
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU nle
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU rro
CGACAUGGGGGCCUGGCUCUGCAGCCUGGCCCACCU csa
```

FIGURE 11 | Nucleotide sequences of the CDSs mRNAs regions of orthologous *ATN1* genes containing the ID01047.3p-miR BSs.

The schemes of miRNA and mRNA nucleotide interactions are a clear illustration of the effectiveness of the MirTarget program in establishing miRNA BSs in mRNA of candidate PD genes (**Supplementary Figure 6**). These schemes demonstrate the important role of non-canonical U and G, A and C pairs in maintaining the double-stranded structure of the miRNA-mRNA complex while maintaining the stacking interaction between miRNA and mRNA nucleotides, which gives their complex increased stability.

In 21 genes with high and low expression, no miRNA-miRNA associations with a ΔG value of more than −130 kJ/mole were found. However, miRNA associations with multiple BSs in the mRNA of candidate target genes can be proposed as associations for diagnostics. These miRNAs and the target gene may include miR-574-5p, ID00470.5p-miR, and *VSNL1* (**Table 3** and **Figure 15**).

Of the large family of miR-1273a,c,d,e,f,g-5p or -3p (Ivashchenko et al., 2014a) only a few candidate genes have been targeted by some miR-1273 (**Table 3**). A cluster of BS miR-1273a, miR-1273c, and miR-1273g-3p with an efficiency of their binding $\Delta G/\Delta Gm$ of more than 90% was detected in the mRNA of the hsa-PDP2 gene. For mRNA of ptr-*PDP2*, ppa-*PDP2*, and ggo-*PDP2* orthologous genes, the $\Delta G/\Delta Gm$ value was 85% or more. mRNA of orthologous genes pab-*PDP2*, mfa-*PDP2*, mml-*PDP2*, mne-*PDP2*, mmu-*PDP2*, rno-*PDP2* interacted with miR-1273a, miR-1273c, and miR-1273g-3p with a $\Delta G/\Delta Gm$ value of less than 80%, which indicates a weak interaction of these miRNA and mRNA. Diagrams of the corresponding associations for hsa, ggo, ppa, ptr are shown in **Supplementary Figure 8** and demonstrate the interaction of miRNA and mRNA without bubbles. Note that in the clusters of BS, purine for purine and pyrimidine for pyrimidine is replaced, which insignificantly affects the free energy of interaction between miRNA and mRNA. Oligonucleotides before and after the cluster of BS miR-1273a, miR-1273c, and miR-1273g-3p were conserved (**Table 3**), which reflects the need to maintain the position of the cluster of BS for these miRNAs. Based on the results presented, the association of miR-1273a, miR-1273c, miR-1273g-3p, and the PDP2 gene can be proposed as a marker for the diagnosis of PD.

The mRNA of the *CCNY* gene, in addition to the 5′UTR BSs, had BSs for six miRNAs in the 3′UTR (**Supplementary Table 4**). The miR-1273a, ID03224.5p-miR, and miR-1273g-3p BSs formed a cluster 45 nt long with a total length of 69 nt BSs. The mRNA of the *DIRAS1* gene had two clusters of miRNA BSs that started at 929 and at 3,443 nt. This placement of clusters led to the competition of miRNA in each of the clusters for binding to the mRNA of the *DIRAS1* gene (**Supplementary Table 4**). Consequently, the highly expressed *CCNY*, *DIRAS1*, and *VSNL1* genes have clusters of miRNA BSs in their mRNAs. In the mRNA of the *WDR82* gene, a cluster of miR-5095 and miR-619-5p BSs was detected with a difference of six nucleotides in the start sites of the BSs. Candidate PD genes with high expression levels did not have miRNA BSs with free energy greater than −130 kJ/mole in the 3′UTRs of their mRNAs (**Table 3** and **Supplementary Table 4**).

Note that genes expressed with RPKM values from 0.1 to 10 had an average RPKM value of 3.5 ± 2.9 and the host genes of intronic miRNAs had an average RPKM value of 4.6 ± 2.8 (**Supplementary Table 5**). The correlation coefficient between the RPKM values of the host genes and 51 target genes of their miRNA was equal to 0.26, that is, there was no strong relationship between the expression of miRNA and potential target genes.

The expression of miRNA and the expression of their target genes were comparable, which indicates the need to maintain close concentrations of miRNA and corresponding mRNA in the norm.

For target genes with a high RPKM value of 43.1–322.5 (average value 116.8), the RPKM value of miRNA host genes varied in the range 1.2–22.4 (average value 7.8). Therefore, these miRNAs normally only slightly alter the expression of target genes, since the expected miRNA concentration will be about 15 times less than the mRNA concentration. However, in pathology, the concentration of miRNA can increase tens of times, or the expression of the gene can decrease many times, and then their significant interaction will occur. This analysis should be taken into account when interpreting the experimental results of determining the concentrations of miRNA and mRNA target genes that make up the association for the diagnosis of the disease. Most of the miRNAs that act on candidate genes for PD are new miRNAs. Unfortunately, we have no information about which of them are of intronic origin. However, the significant similarity between the properties of old and new miRNAs (Aisina et al., 2019; Kondybayeva et al., 2020; Mukushkina et al., 2020) allows us to hope that the relationships revealed in this work between the expression of old miRNAs and their potential target genes are also characteristic of new miRNAs.

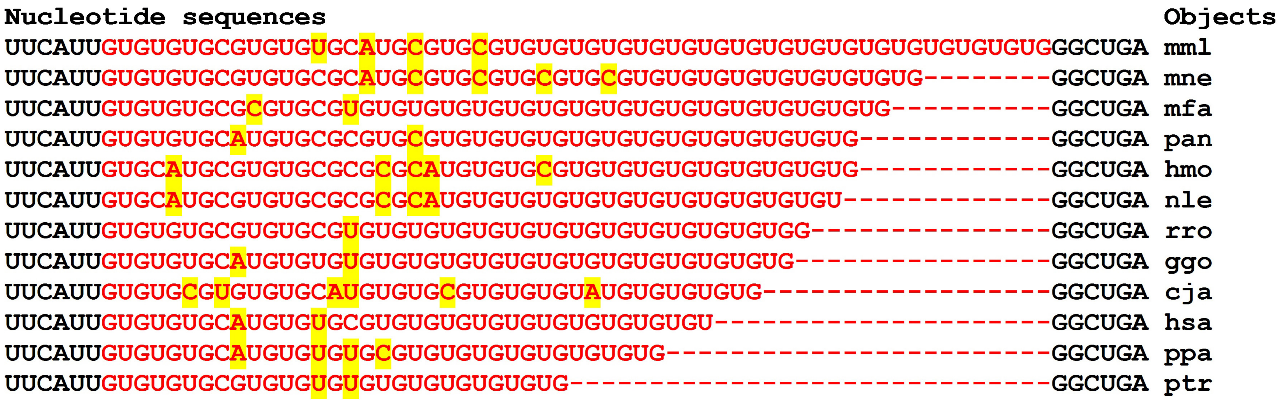
DISCUSSION

Our studies have shown that for many known PD candidate genes, their mRNAs are effectively targeted by miRNAs. The *in silico* characteristics of the interactions between miRNAs and mRNAs can be used in calculating the inhibition efficiency of the translation process at different ratios of miRNA and mRNA concentrations. Thus, using the kinetic equations of the analysis of the interaction of the inhibitor and the enzyme, we can interpret the effect of miRNA by changing the ratio of the mRNA and miRNA concentrations.

The correlations established in many published reports between the changes in the concentration of one or several miRNAs and the changes in the expression of putative target genes involved in the development of PD are not very reliable. This lack of reliability is observed because in most studies of PD and other diseases, the concentration of miRNA was not controlled simultaneous to the expression of the putative target genes. The results of their interactions strongly depend on the ratio of the miRNA concentrations and mRNA concentrations of the target genes. For example, even with strong binding of miRNA to mRNA, the suppression of gene expression is negligible if the concentration of miRNA is significantly lower than the concentration of mRNA. Conversely, with an average interaction of miRNA with mRNA and an excess of the miRNA concentration over the mRNA concentration, strong translation inhibition is observed. Therefore, in the tables, we also present low characteristics of the binding of miRNAs to mRNAs of candidate genes. These associations of miRNAs and genes can be markers with increasing concentrations of miRNAs relative to mRNAs. Considering the competition between miRNAs upon binding to mRNAs, the problem of establishing an effective miRNA for a particular gene becomes even more complicated, since the expression of several or even tens of miRNAs and genes needs to be controlled. Bioinformatics approaches make it possible to select from several thousand miRNAs that are likely to interact with candidate PD genes, which significantly reduces the material costs of searching for miRNAs and target gene associations.

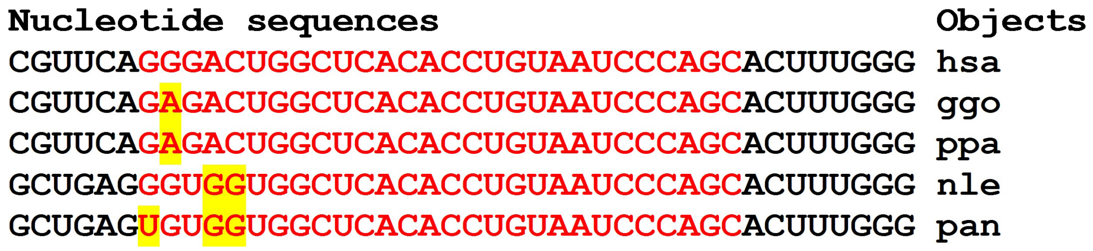
TABLE 3 | Characteristics of the interactions between miRNAs and the 3′UTR mRNAs of candidate PD genes.

Gene; RPKM	miRNA	Start of site, nt	ΔG, kJ/mole	ΔG/ΔGm, %	Length, nt
GSK3B; 8.3	ID01030.3p-miR	4,705÷4,719 (4)	−108÷ −113	89÷93	23
	miR-466	4,709÷4,721 (6)	−104÷ −106	89÷91	23
	ID00436.3p-miR	4,713÷4,723 (3)	−104	89	23
	ID01727.5p-miR	4,722	−106	91	23
LRP10; 7.8	(1.6) miR-5096	3,237	−104	92	21
	ID02175.3p-miR	3,353	−110	91	22
	[3.6] miR-5585-3p	3,305	−115	98	22
	[5.3] miR-1285-5p	3,404	−102	91	21
	[4.3] miR-619-5p	3,497	−110	91	22
	[19.7] miR-4452	3,544	−108	94	23
	[8.3] miR-5095	3,788	−106	91	21
	[4.3] miR-619-5p	3,794	−115	95	22
	ID00913.5p-miR	3,814	−117	92	23
PDP2; 1.2	ID00047.3p-miR	3,220	−110	93	21
	[1.6] miR-5096	3,920	−108	96	21
	[4.3] miR-619-5p	3,980	−113	93	22
	[3.6] miR-5585-3p	3,987	−106	91	22
	[5.3] miR-1285-5p	4,086	−102	91	21
	ID01200.3p-miR	4,511	−102	91	21
	[0.2] miR-1273a	4,639	−119	90	25
	[3.3] miR-1273c	4,641	−110	91	22
	[8.3] miR-1273g-3p	4,661	−106	91	21
	ID01360.3p-miR	5,493	−104	91	21
	miR-3159	5,861	−106	91	22
	[4.3] miR-619-5p	5,863	−113	93	22
	[4.3] miR-619-5p	5,988	−110	91	22
	[4.3] miR-619-5p	6,173	−119	98	22
	[1.6] miR-5096	6,247	−108	96	21
	[4.3] miR-619-5p	6,308	−117	96	22
	ID01836.5p-miR	6,398	−113	90	23
	[1.6] miR-5096	6,413	−102	91	21
PPARGC1A; 2.5	miR-466	3,321, 3,337	−106	91	23
	ID00436.3p-miR	3,323÷3,339 (3)	−104÷ −108	89÷93	23
	ID01030.3p-miR	3,325	−115	95	23
	ID01727.5p-miR	3,338	−104	89	23
RBBP5; 3.9	(8.3) miR-5095	3,065	−108	93	21
	[4.3] miR-619-5p	3,071	−113	93	22
	[1.6] miR-5096	3,145	−106	94	21
	ID03006.5p-miR	4,015	−121	89	24
	[4.3] miR-619-5p	4,030	−115	95	22
	[1.6] miR-5096	4,104	−106	94	21
	miR-3159	4,163	−106	91	22
	ID02175.3p-miR	4,220	−113	93	22
	ID01237.3p-miR	4,271	−117	92	24
SLC14A1; 8.5	(8.3) miR-5095	2,771	−110	95	21
	[4.3] miR-619-5p	2,777	−119	98	22
	[1.6] miR-5096	2,851	−102	91	21
	[4.3] miR-619-5p	3,215	−115	95	22
	ID01836.5p-miR	3,003	−113	90	23
VSNL1; 206.5	(1.5) miR-574-5p	1,021÷1,045 (13)	−108 ÷ −113	89–93	23
	ID00470.5p-miR	1,023÷1,045 (12)	−108	89	23



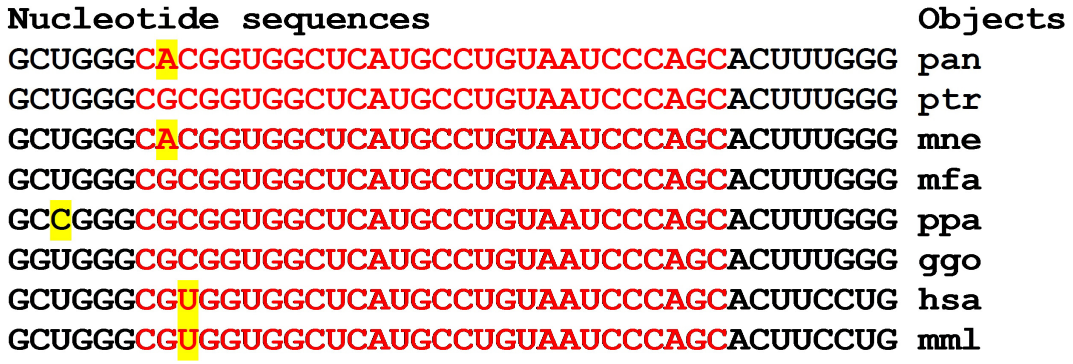
```
Nucleotide sequences                                                    Objects
UUCAUUGUGUGUGCGUGUGUGCAUGCGUGCGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGGGCUGA mml
UUCAUUGUGUGUGCGUGUGCGCAUGCGUGCGUGCGUGCGUGUGUGUGUGUGUGUGUG--------GGCUGA mne
UUCAUUGUGUGUGCGCGUGCGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUG----------GGCUGA mfa
UUCAUUGUGUGUGCAUGUGCGCGUGCGUGUGUGUGUGUGUGUGUGUGUGUGUG------------GGCUGA pan
UUCAUUGUGCAUGCGUGUGCGCGCGCAUGUGUGCGUGUGUGUGUGUGUGUGUG------------GGCUGA hmo
UUCAUUGUGCAUGCGUGUGCGCGCGCAUGUGUGUGUGUGUGUGUGUGUGUGU-------------GGCUGA nle
UUCAUUGUGUGUGCGUGUGCGUGUGUGUGUGUGUGUGUGUGUGUGUGUGG---------------GGCUGA rro
UUCAUUGUGUGUGCAUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUG----------------GGCUGA ggo
UUCAUUGUGUGCGUGUGUGCAUGUGUGCGUGUGUGUAUGUGUGUGUG------------------GGCUGA cja
UUCAUUGUGUGUGCAUGUGUGCGUGUGUGUGUGUGUGUGUGUGU---------------------GGCUGA hsa
UUCAUUGUGUGUGCAUGUGUGUGCGUGUGUGUGUGUGUGUG------------------------GGCUGA ppa
UUCAUUGUGUGUGCGUGUGUGUGUGUGUGUGUGUG------------------------------GGCUGA ptr
```

FIGURE 12 | Nucleotide sequences of 3′UTR regions of mRNAs of orthologous *GSK3B* genes containing clusters of miRNAs BSs.



```
Nucleotide sequences                           Objects
CGUUCAGGGACUGGCUCACACCUGUAAUCCCAGCACUUUGGG hsa
CGUUCAGAGACUGGCUCACACCUGUAAUCCCAGCACUUUGGG ggo
CGUUCAGAGACUGGCUCACACCUGUAAUCCCAGCACUUUGGG ppa
GCUGAGGGUGGUGGCUCACACCUGUAAUCCCAGCACUUUGGG nle
GCUGAGUGUGGUGGCUCACACCUGUAAUCCCAGCACUUUGGG pan
```

FIGURE 13 | Nucleotide sequences of 3′UTR regions of mRNAs of orthologous *SLC14A1* genes containing clusters of miR-5095 and miR-619-5p BSs.



```
Nucleotide sequences                           Objects
GCUGGGCACGGUGGCUCAUGCCUGUAAUCCCAGCACUUUGGG pan
GCUGGGCGCGGUGGCUCAUGCCUGUAAUCCCAGCACUUUGGG ptr
GCUGGGCACGGUGGCUCAUGCCUGUAAUCCCAGCACUUUGGG mne
GCUGGGCGCGGUGGCUCAUGCCUGUAAUCCCAGCACUUUGGG mfa
GCCGGGCGCGGUGGCUCAUGCCUGUAAUCCCAGCACUUUGGG ppa
GGUGGGCGCGGUGGCUCAUGCCUGUAAUCCCAGCACUUUGGG ggo
GCUGGGCGUGGUGGCUCAUGCCUGUAAUCCCAGCACUUCCUG hsa
GCUGGGCGUGGUGGCUCAUGCCUGUAAUCCCAGCACUUCCUG mml
```

FIGURE 14 | Nucleotide sequences of 3′UTR regions of mRNAs of orthologous *PRKN* genes containing clusters of binding sites miR-5095 and miR-619-5p.

Based on the results obtained in this study, the following generalizations can be made. Not all of the more than 200 candidate PD genes were targets of miRNA. Out of 6,756 miRNAs, only 150 miRNAs were identified that efficiently bound to the mRNA of 61 candidate PD genes. The miRNA BSs were located in the 5′UTRs, CDSs and 3′UTRs of the mRNAs of candidate PD genes. Each of more than half of the candidate genes was the potential target of one miRNA. The mRNAs of the remaining genes could bind two or more miRNAs. The BSs of most miRNAs were located along the entire length of the mRNA without overlapping nucleotide sequences. In the mRNA of some genes, miRNA BSs located in overlapping nucleotide sequences (clusters) were detected. Such clusters reduced by several times the proportion of nucleotide sequences of miRNA BSs in the 5′UTRs, CDSs and 3′UTRs. The miRNAs with BSs in the cluster compete with one another, and only one of these miRNAs can bind to mRNA. The miRNA with a large free energy of interaction with the mRNA and present in a higher concentration compared to other miRNAs has the advantage for binding. The start of the miR-619-5p and miR-5095 BSs are located over six nucleotides, thereby forming a cluster. As a rule, the free energy of the interaction of the mRNA with miR-619-5p is greater than that with miR-5095 (**Table 3** and **Supplementary Table 3**). However, if the concentration of miR-5095 is two to three times higher than

```
Nucleotide sequences                                                       Objects
GCUAUUCACACACACACCCACACACACACACCCACACACACACACACACACACACACACACACACACAAAUAUU csa
GCUAUUCACACACACACACACACACACACACACACACACACACACACACACACACAC--------AAAUAUU ptr
GCUAUUCACACACACACACACACACACACACACACACACACACACAC----------------AAAUAUU hsa
GCUAUUCACACAUACACACACACACACACACACACACACACACACAC----------------AAAUAUU cja
GCUAUUCAUACACACACACACACACACACACACACACACACACACAC----------------AAAUAUU rro
GCUAUUCACACACACACACACACACACACACACACACACACACACAC----------------AAAUAUU ggo
GCUAUUCACACACACACACACACCCACACACACACACACACACACAC--------------------AAAUAUU pan
GCUAUUCACACACACACACACACACACACACACACACACACACAC--------------------AAAUAUU pab
GCUAUUCACACACACACACACCCACACCCACACACACACACACACAC--------------------AAAUAUU mfa
GCUAUUCACACACCCACACACCCACACACACACACACACACACACAC--------------------AAAUAUU mne
GCUAUUCACACACACACACACACACACACACACACACAC--------------------------AAAUAUU nle
GCUAUUCACACACACACACCCACACCCACACACACACAC----------------------------AAAUAUU mml
GCUAUUCACACACACACACACACACACACACACACAC----------------------------AAAUAUU ppa
```

FIGURE 15 | Nucleotide sequences of 3′UTR regions of mRNAs of orthologous *VSNL1* genes containing clusters of miRNAs binding sites.

the concentration of miR-619, then it is more likely to suppress translation. If the cluster contains the BSs of many miRNAs, then more complex calculations are required to establish the miRNAs with the greatest influence on the translation process.

In the CDSs of mRNAs of almost all low and highly expressed genes, there were miRNA BSs that were not repeated in other genes (**Table 2** and **Supplementary Table 3**). In other words, these associations of miRNA and candidate PD target genes are specific, which gives them preference for use in diagnosis. A feature of some PD candidate genes is the presence of clusters containing BSs for the same set of miRNAs in their mRNAs (**Tables 1**, **3** and **Supplementary Table 4**). Differences in the binding characteristics of miRNAs to the mRNAs of genes expressed at different rates have been established. Multiple BSs of miR-466, ID01030.3p-miR, ID00436.3p-miR, and ID01727.5p-miR were located in the 3′UTRs of the mRNAs of the *PPARGC1A* and *GSK3B* genes with low expression (**Table 3**). The miR-5095 and miR-619-5p BSs formed a cluster, and the beginnings of their BSs differed by six nucleotides. This connection of miR-5095 and miR-619-5p BSs is not random, since it is observed in the mRNA of the *LRP10, PDP2, PRKN, RBBP5, SLC14A1*, and *WDR82* genes (**Table 3** and **Supplementary Table 4**). In addition, the difference between the miR-5096 and miR-619-5p BSs (74 nt) was the same in the mRNAs of the *PDP2, RBBP5*, and *SLC14A1* genes. The start of the miR-5585-3p and miR-1285-5p BSs differed by 99 nt in the mRNA of the *LRP10* and *PDP2* genes. miR-619-5p binds to the mRNA of the *PRKN* gene completely complementarily among 201 genes that are the potential target of this miRNA (Atambayeva et al., 2017). Candidate PD genes that are targets of miRNAs that bind in 3′UTRs of mRNAs significantly differed from other candidate genes by the number of miR-619-5p, miR-5095, miR-5096, miR-5585-3p, and miR-1285-5p BSs (**Table 3** and **Supplementary Table 4**; Ivashchenko et al., 2014a).

In the 5′UTRs of the mRNAs of genes, miRNA BSs were more frequently organized into clusters (**Table 1**). The identified features of the interactions between miRNAs and the mRNAs of candidate PD genes should be taken into account when selecting miRNA associations and potential target genes for diagnosing the disease.

Based on the quantitative characteristics of the interactions between miRNAs and mRNAs, the associations of miRNAs and candidate genes with a high free energy of interaction were identified, which are recommended as markers for the diagnosis of PD. For the diagnosis of diseases, it is preferable to use miRNA associations with BS in the 5′UTRs of candidate genes, since the free energy of interaction between miRNAs and mRNAs has a higher value than in CDSs and 3′UTRs (**Tables 1–3**). Let us consider examples of the association of miRNAs and genes on which the development of PD can depend to a greater extent. Clusters of ID00296.3p-miR and ID01702.3p-miR BSs were detected in the mRNAs of the *GSK3B, SETD1A*, and *CCNY* genes. Therefore, it is necessary to control the expression of both miRNAs and the three genes to evaluate the role of these associations in the development of the disease. This approach is necessary in elucidating the role of other associations of miRNAs and genes in the development of PD. For example, the association of several miRNAs (ID01041.5p-miR, ID00457.3p-miR, ID03367.5p-miR, and ID02770.5p-miR) and the *GSK3B* gene shows the need to control the binding of these miRNAs in two clusters in the mRNA of the *GSK3B* gene (**Table 1**). In addition to the two considered examples of miRNAs and gene associations, other associations will be considered below, which generally demonstrate the need to control a large number of miRNAs and candidate gene expression levels. There is no other approach to determine which miRNAs out of the currently known 6,266 miRNAs can regulate the development of PD. The bioinformatics approach enables only dozens of effective associations to be selected from many millions of associations between miRNAs and mRNAs.

In the 5′UTR of the mRNA of the *PPARGC1A* gene, there was a cluster of ID00470.5p-miR and miR-574-5p BSs, each of which had five sequentially located BSs (**Table 1**). In the 3′UTR of the mRNA of the *VSNL1* gene, these miRNAs had more than ten multiple BSs (**Table 3**). Therefore, these miRNAs can have a greater effect on the expression of these genes than miRNAs with one BS. With point mutations of nucleotides (e.g., single nucleotide polymorphisms) in a cluster with multiple BSs, the effectiveness of these miRNAs does not change substantially.

Analysis of the role of the expression of candidate genes in the form of potential miRNA targets leads to the following conclusions. It is expected that for the regulation of highly expressed genes, comparably high concentrations of miRNAs are required; otherwise, if the miRNAs are present at lower concentrations than the mRNAs, they will not significantly regulate the translation process. Based on the above considerations, the concentrations of miRNA and mRNA should be comparable. Therefore, there is a conserved relationship between the nucleotide sequences of miRNAs and miRNA BSs in the mRNA (Davis et al., 2005; Wang et al., 2016; Yurikova et al., 2019). While recommending the association of miRNAs and potential target genes for disease diagnosis, we emphasize that miRNA and mRNA concentrations must be simultaneously recorded. Without these quantitative indicators, it is difficult to draw conclusions regarding the significance of the data obtained.

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

AI and AAr conceived of the study, drafted the manuscript, and gave final approval of the version to be published. SK and AK conceived of the study and drafted the manuscript. AAk analyzed the effect of miR-619-5p and miR-5095 on genes, conceived of the study, and drafted the manuscript. All authors made substantial contributions to acquisition of data, to interpretation and modification of the data, were involved in subsequent rounds of revisions, and read and approved the final manuscript.

FUNDING

The work was carried out with the financial support of the Ministry of Education and Science of the Republic of Kazakhstan within the framework of the grant no. AP05132460.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.647288/full#supplementary-material

REFERENCES

Aisina, D., Niyazova, R., Atambayeva, S., and Ivashchenko, A. (2019). Prediction of clusters of miRNA binding sites in mRNA candidate genes of breast cancer subtypes. *PeerJ* 7:e8049. doi: 10.7717/peerj.8049

Arshad, A. R., Sulaiman, S. A., Saperi, A. A., Jamal, R., Ibrahim, M. N., and Murad, N. A. A. (2017). MicroRNAs and target genes as biomarkers for the diagnosis of early onset of Parkinson disease. *Front. Mol. Neurosci.* 10:352. doi: 10.3389/fnmol.2017.00352

Atambayeva, S., Niyazova, R., Ivashchenko, A., Pyrkova, A., Pinsky, I., Akimniyazova, A., et al. (2017). The binding sites of miR-619-5p in the mRNAs of human and orthologous genes. *BMC Genomics* 18:428. doi: 10.1186/s12864-017-3811-6

Behbahanipour, M., Peymani, M., Salari, M., Hashemi, M. S., Nasr-Esfahani, M. H., and Ghaedi, K. (2019). Expression profiling of blood microRNAs 885, 361, and 17 in the patients with the Parkinson's disease: integrating interaction data to uncover the possible triggering age-related mechanisms. *Sci. Rep.* 9:13759. doi: 10.1038/s41598-019-50256-3

Berillo, O., Régnier, M., and Ivashchenko, A. (2013). Binding of intronic miRNAs to the mRNAs of host genes encoding intronic miRNAs and proteins that participate in tumourigenesis. *Comput. Biol. Med.* 43, 1374–1381. doi: 10.1016/j.compbiomed.2013.07.011

Brennan, S., Keon, M., Liu, B., Su, Z., and Saksena, N. K. (2019). Panoramic visualization of circulating MicroRNAs across neurodegenerative diseases in humans. *Mol. Neurobiol.* 56, 7380–7407. doi: 10.1007/s12035-019-1615-1

Cao, X. Y., Lu, J. M., Zhao, Z. Q., Li, M. C., Lu, T., An, X. S., et al. (2017). MicroRNA biomarkers of Parkinson's disease in serum exosome-like microvesicles. *Neurosci. Lett.* 644, 94–99. doi: 10.1016/j.neulet.2017.02.045

Chen, L., Yang, J., Lü, J., Cao, S., Zhao, Q., and Yu, Z. (2018). Identification of aberrant circulating miRNAs in Parkinson's disease plasma samples. *Brain Behav.* 8:e00941. doi: 10.1002/brb3.941

Chen, Y., Zheng, J., Su, L., Chen, F., Zhu, R., Chen, X., et al. (2020). Increased salivary microRNAs that regulate DJ-1 gene expression as potential markers for Parkinson's disease. *Front. Aging Neurosci.* 12:210. doi: 10.3389/fnagi.2020.00210

D'Anca, M., Fenoglio, C., Serpente, M., Arosio, B., Cesari, M., Scarpini, E. A., et al. (2019). Exosome determinants of physiological aging and age-related neurodegenerative diseases. *Front. Aging Neurosci.* 11:232. doi: 10.3389/fnagi.2019.00232

Davis, E., Caiment, F., Tordoir, X., Cavaillé, J., Ferguson-Smith, A., Cockett, N., et al. (2005). RNAi-mediated allelic trans-interaction at the imprinted Rtl1/Peg11 locus. *Curr. Biol.* 15, 743–749. doi: 10.1016/j.cub.2005.02.060

Dong, X., Zheng, D., and Nao, J. (2020). Circulating exosome microRNAs as diagnostic biomarkers of dementia. *Front. Aging Neurosci.* 12:580199. doi: 10.3389/fnagi.2020.580199

Fejes, Z., Erdei, J., Pócsi, M., Takai, J., Jeney, V., Nagy, A., et al. (2020). Elevated pro-inflammatory cell-free MicroRNA levels in cerebrospinal fluid of premature infants after intraventricular hemorrhage. *Int. J. Mol. Sci.* 21:6870. doi: 10.3390/ijms21186870

Garg, A., and Heinemann, U. (2018). A novel form of RNA double helix based on G·U and C·A+ wobble base pairing. *RNA* 24, 209–218. doi: 10.1261/rna.064048.117

Hu, Y. B., Zhang, Y. F., Wang, H., Ren, R. J., Cui, H. L., Huang, W. Y., et al. (2019). miR-425 deficiency promotes necroptosis and dopaminergic neurodegeneration in Parkinson's disease. *Cell Death Dis.* 10:589. doi: 10.1038/s41419-019-1809-5

Ivashchenko, A., Berillo, O., Pyrkova, A., and Niyazova, R. (2014a). Binding sites of miR-1273 family on the mRNA of target genes. *Biomed Res. Int.* 2014:620530. doi: 10.1155/2014/620530

Ivashchenko, A., Berillo, O., Pyrkova, A., Niyazova, R., and Atambayeva, S. (2014b). miR-3960 binding sites with mRNA of human genes. *Bioinformation* 10, 423–427. doi: 10.6026/97320630010423

Ivashchenko, A., Berillo, O., Pyrkova, A., Niyazova, R., and Atambayeva, S. (2014c). The properties of binding sites of miR-619-5p, miR-5095, miR-5096 and miR-5585-3p in the mRNAs of human genes. *Biomed Res. Int.* 2014:720715. doi: 10.1155/2014/720715

Ivashchenko, A. T., Pyrkova, A. Y., Niyazova, R. Y., Alybayeva, A., and Baskakov, K. (2016). Prediction of miRNA binding sites in mRNA. *Bioinformation* 12, 237–240. doi: 10.6026/97320630012237

Jurjević, I., Miyajima, M., Ogino, I., Akiba, C., Nakajima, M., Kondo, A., et al. (2017). Decreased expression of hsa-miR-4274 in cerebrospinal fluid of normal pressure hydrocephalus mimics with Parkinsonian syndromes. *J. Alzheimers Dis.* 56, 317–325. doi: 10.3233/JAD-160848

Kakati, T., Bhattacharyya, D. K., Barah, P., and Kalita, J. K. (2019). Comparison of methods for differential co-expression analysis for disease biomarker prediction. *Comput. Biol. Med.* 113:103380. doi: 10.1016/j.compbiomed.2019.103380

Kondybayeva, À, Akimniyazova, A., Kamenova, S., Duchshanova, G., Aisina, D., Goncharova, A., et al. (2020). Prediction of miRNA interaction with mRNA of stroke candidate genes. *Neurol. Sci.* 41, 799–808. doi: 10.1007/s10072-019-04158-x

Kondybayeva, ÀÌ, Akimniyazova, ÀN., Kamenova, S. U., and Ivashchenko, ÀÒ (2018). The characteristics of miRNA binding sites in mRNA of *ZFHX3* gene and its orthologs. *Vavilov J. Genet. Breed.* 22, 438–444. doi: 10.18699/VJ18.380

Kool, E. T. (2001). Hydrogen bonding, base stacking, and steric effects in DNA replication. *Annu. Rev. Biophys. Biomol. Struct.* 30, 1–22. doi: 10.1146/annurev.biophys.30.1.1

Leggio, L., Vivarelli, S., L'Episcopo, F., Tirolo, C., Caniglia, S., Testa, N., et al. (2017). microRNAs in Parkinson's disease: from pathogenesis to novel diagnostic and therapeutic approaches. *Int. J. Mol. Sci.* 18:2698. doi: 10.3390/ijms18122698

Lemieux, S., and Major, F. (2002). RNA canonical and non-canonical base pairing types: a recognition method and complete repertoire. *Nucleic Acids Res.* 30, 4250–4263. doi: 10.1093/nar/gkf540

Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* 30, 3497–3531. doi: 10.1093/nar/gkf481

Li, L., Xu, J., Wu, M., and Hu, J. M. (2018). Protective role of microRNA-221 in Parkinson's disease. *Bratisl. Lek. Listy* 119, 22–27. doi: 10.4149/BLL_2018_005

Liu, W., Li, L., Liu, S., Wang, Z., Kuang, H., Xia, Y., et al. (2019). MicroRNA expression profiling screen miR-3557/324-targeted CaMK/mTOR in the rat striatum of Parkinson's disease in regular aerobic exercise. *Biomed Res. Int.* 2019:7654798. doi: 10.1155/2019/7654798

Londin, E., Loher, P., Telonis, A. G., Quann, K., Clark, P., Jing, Y., et al. (2015). Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1106–1115. doi: 10.1073/pnas.1420955112

Lu, J., Xu, Y., Quan, Z., Chen, Z., Sun, Z., and Qing, H. (2017). Dysregulated microRNAs in neural system: implication in pathogenesis and biomarker development in Parkinson's disease. *Neuroscience* 365, 70–82. doi: 10.1016/j.neuroscience.2017.09.033

Marques, T. M., Kuiperij, H. B., Bruinsma, I. B., van Rumund, A., Aerts, M. B., Esselink, R. A. J., et al. (2017). MicroRNAs in cerebrospinal fluid as potential biomarkers for Parkinson's disease and multiple system atrophy. *Mol. Neurobiol.* 54, 7736–7745. doi: 10.1007/s12035-016-0253-0

Martinez, B., and Peplow, P. V. (2017). MicroRNAs in Parkinson's disease and emerging therapeutic targets. *Neural Regen. Res.* 12, 1945–1959. doi: 10.4103/1673-5374.221147

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Mukushkina, D., Aisina, D., Pyrkova, A., Ryskulova, A., Labeit, S., and Ivashchenko, A. (2020). In silico prediction of miRNA interactions with candidate atherosclerosis gene mRNAs. *Front. Genet.* 11:605054. doi: 10.3389/fgene.2020.605054

Nie, C., Sun, Y., Zhen, H., Guo, M., Ye, J., Liu, Z., et al. (2020). Differential expression of plasma exo-miRNA in neurodegenerative diseases by next-generation sequencing. *Front. Neurosci.* 14:438. doi: 10.3389/fnins.2020.00438

Niyazova, R., Berillo, O., Atambayeva, S., Pyrkova, A., Alybaeva, A., and Ivashchenko, A. (2015). miR-1322 binding sites in paralogous and orthologous genes. *Biomed Res. Int.* 2015, 1–7. doi: 10.1155/2015/962637

Ozdilek, B., and Demircan, B. (2020). Serum microRNA expression levels in Turkish patients with Parkinson's disease. *Int. J. Neurosci.* 130, 1–9. doi: 10.1080/00207454.2020.1784165

Patil, K. S., Basak, I., Dalen, I., Hoedt, E., Lange, J., Lunde, K. A., et al. (2019). Combinatory microRNA serum signatures as classifiers of Parkinson's disease. *Parkinsonism Relat. Disord.* 64, 202–210. doi: 10.1016/j.parkreldis.2019.04.010

Qin, L. X., Tan, J. Q., Zhang, H. N., Tang, J. G., Jiang, B., Shen, X. M., et al. (2019). Preliminary study of hsa-miR-626 change in the cerebrospinal fluid of Parkinson's disease patients. *J. Clin. Neurosci.* 70, 198–201. doi: 10.1016/j.jocn.2019.08.082

Quinlan, S., Kenny, A., Medina, M., Engel, T., and Jimenez-Mateos, E. M. (2017). MicroRNAs in neurodegenerative diseases. *Int. Rev. Cell Mol. Biol.* 334, 309–343. doi: 10.1016/bs.ircmb.2017.04.002

Ramaswamy, P., Yadav, R., Pal, P. K., and Christopher, R. (2020). Clinical application of circulating microRNAs in Parkinson's disease: the challenges and opportunities as diagnostic biomarker. *Ann. Indian Acad. Neurol.* 23, 84–97. doi: 10.4103/aian.AIAN_440_19

Ravanidis, S., Bougea, A., Papagiannakis, N., Koros, C., Simitsi, A. M., Pachi, I., et al. (2020). Validation of differentially expressed brain-enriched microRNAs in the plasma of PD patients. *Ann. Clin. Transl. Neurol.* 7, 1594–1607. doi: 10.1002/acn3.51146

Recabarren, D., and Alarcón, M. (2017). Gene networks in neurodegenerative disorders. *Life Sci.* 183, 83–97. doi: 10.1016/j.lfs.2017.06.009

Ren, Y., Li, H., Xie, W., Wei, N., and Liu, M. (2019). MicroRNA-195 triggers neuroinflammation in Parkinson's disease in a Rho-associated kinase 1-dependent manner. *Mol. Med. Rep.* 19, 5153–5161. doi: 10.3892/mmr.2019.10176

Rosas-Hernandez, H., Chigurupati, S., Raymick, J., Robinson, B., Cuevas, E., Hanig, J., et al. (2018). Identification of altered microRNAs in serum of a mouse model of Parkinson's disease. *Neurosci. Lett.* 687, 1–9. doi: 10.1016/j.neulet.2018.07.022

Rostamian Delavar, M., Baghi, M., Safaeinejad, Z., Kiani-Esfahani, A., Ghaedi, K., and Nasr-Esfahani, M. H. (2018). Differential expression of miR-34a, miR-141, and miR-9 in MPP+-treated differentiated PC12 cells as a model of Parkinson's disease. *Gene* 662, 54–65. doi: 10.1016/j.gene.2018.04.010

Sadlon, A., Takousis, P., Alexopoulos, P., Evangelou, E., Prokopenko, I., and Perneczky, R. (2019). miRNAs identify shared pathways in Alzheimer's and Parkinson's diseases. *Trends Mol. Med.* 25, 662–672. doi: 10.1016/j.molmed.2019.05.006

Salamon, A., Zádori, D., Szpisjak, L., Klivenyi, P., and Vecsei, L. (2020). Neuroprotection in Parkinson's disease: facts and hopes. *J. Neural Trans.* 127, 821–829. doi: 10.1007/s00702-019-02115-8

Singh, A., and Sen, D. (2017). MicroRNAs in Parkinson's disease. *Exp. Brain Res.* 235, 2359–2374. doi: 10.1007/s00221-017-4989-1

Starhof, C., Hejl, A. M., Heegaard, N., Carlsen, A. L., Burton, M., Lilje, B., et al. (2019). The biomarker potential of cell-free microRNA from cerebrospinal fluid in Parkinsonian syndromes. *Mov. Disord.* 34, 246–254. doi: 10.1002/mds.27542

Thomas, L., Florio, T., and Perez-Castro, C. (2020). Extracellular vesicles loaded miRNAs as potential modulators shared between glioblastoma, and Parkinson's and Alzheimer's diseases. *Front. Cell. Neurosci.* 14:590034. doi: 10.3389/fncel.2020.590034

Titze-de-Almeida, R., and Titze-de-Almeida, S. S. (2018). miR-7 replacement therapy in Parkinson's disease. *Curr. Gene Ther.* 18, 143–153. doi: 10.2174/1566523218666180430121323

Tolosa, E., Botta-Orfila, T., Morató, X., Calatayud, C., Ferrer-Lorente, R., Martí, M. J., et al. (2018). MicroRNA alterations in iPSC-derived dopaminergic neurons from Parkinson disease patients. *Neurobiol. Aging* 69, 283–291. doi: 10.1016/j.neurobiolaging.2018.05.032

Uwatoko, H., Hama, Y., Iwata, I. T., Shirai, S., Matsushima, M., Yabe, I., et al. (2019). Identification of plasma microRNA expression changes in multiple system atrophy and Parkinson's disease. *Mol. Brain* 12:49. doi: 10.1186/s13041-019-0471-2

Viswambharan, V., Thanseem, I., Vasu, M. M., Poovathinal, S. A., and Anitha, A. (2017). miRNAs as biomarkers of neurodegenerative disorders. *Biomark. Med.* 11, 151–167. doi: 10.2217/bmm-2016-0242

Wang, J., Li, Z., Liu, B., Chen, G., Shao, N., Ying, X., et al. (2016). Systematic study of cis-antisense miRNAs in animal species reveals miR-3661 to target PPP2CA in human cells. *RNA* 22, 87–95. doi: 10.1261/rna.052894.115

Wang, L., and Zhang, L. (2020). Circulating exosomal miRNA as diagnostic biomarkers of neurodegenerative diseases. *Front. Mol. Neurosci.* 13:53. doi: 10.3389/fnmol.2020.00053

Wang, X., Zhou, Y., Gao, Q., Ping, D., Wang, Y., Wu, W., et al. (2020). The role of exosomal microRNAs and oxidative stress in neurodegenerative diseases. *Oxid. Med. Cell. Longev.* 2020:3232869. doi: 10.1155/2020/3232869

Wang, Y., Yang, Z., and Le, W. (2017). Tiny but mighty: promising roles of microRNAs in the diagnosis and treatment of Parkinson's disease. *Neurosci. Bull.* 33, 543–551. doi: 10.1007/s12264-017-0160-z

Xie, S., Niu, W., Xu, F., Wang, Y., Hu, S., and Niu, C. (2020). Differential expression and significance of miRNAs in plasma extracellular vesicles of patients with Parkinson's disease. *Int. J. Neurosci.* 2020, 1–16. doi: 10.1080/00207454.2020.1835899

Yan, J. H., Hua, P., Chen, Y., Li, L. T., Yu, C. Y., Yan, L., et al. (2020). Identification of microRNAs for the early diagnosis of Parkinson's disease and multiple system atrophy. *J. Integr. Neurosci.* 19, 429–436. doi: 10.31083/j.jin.2020.03.163

Yang, Z., Li, T., Cui, Y., Li, S., Cheng, C., Shen, B., et al. (2019). Elevated plasma microRNA-105-5p level in patients with idiopathic Parkinson's disease: a potential disease biomarker. *Front. Neurosci.* 13:218. doi: 10.3389/fnins.2019.00218

Yurikova, O. Y., Aisina, D. E., Niyazova, R. E., Atambayeva, S. A., Labeit, S., and Ivashchenko, A. (2019). The interaction of miRNA-5p and miRNA-3p with the mRNAs of orthologous genes. *Mol. Biol.* 53, 692–704. doi: 10.1134/S0026898419040189

Zhang, Y., Xu, W., Nan, S., Chang, M., and Fan, J. (2019). MicroRNA-326 inhibits apoptosis and promotes proliferation of dopaminergic neurons in Parkinson's disease through suppression of KLK7-mediated MAPK signaling pathway. *J. Mol. Neurosci.* 69, 197–214. doi: 10.1007/s12031-019-01349-1

Zhao, L., and Wang, Z. (2019). MicroRNAs: game changers in the regulation of α-synuclein in Parkinson's disease. *Parkinsons Dis.* 2019:1743183. doi: 10.1155/2019/1743183

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.

TECHNOLOGY AND CODE
published: 10 June 2021
doi: 10.3389/fgene.2021.679029

RHIVDB: A Freely Accessible Database of HIV Amino Acid Sequences and Clinical Data of Infected Patients

Olga Tarasova[1], Anastasia Rudik[1]*, Dmitry Kireev[2]* and Vladimir Poroikov[1]**

[1] *Department of Bioinformatics, Institute of Biomedical Chemistry, Moscow, Russia,* [2] *Central Research Institute of Epidemiology, Moscow, Russia*

Edited by:
Tatiana V. Tatarinova,
University of La Verne, United States

Reviewed by:
Shaoqing Wen,
Fudan University, China
Ramon Lorenzo-Redondo,
Northwestern University,
United States

****Correspondence:***
Olga Tarasova
olga.a.tarasova@gmail.com
Anastasia Rudik
rudik_anastassia@mail.ru
Dmitry Kireev
dmitkireev@yandex.ru
Vladimir Poroikov
vladimir.poroikov@ibmc.msk.ru

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *10 March 2021*
Accepted: *14 May 2021*
Published: *10 June 2021*

Citation:
Tarasova O, Rudik A, Kireev D and Poroikov V (2021) RHIVDB: A Freely Accessible Database of HIV Amino Acid Sequences and Clinical Data of Infected Patients. Front. Genet. 12:679029. doi: 10.3389/fgene.2021.679029

Human immunodeficiency virus (HIV) infection remains one of the most severe problems for humanity, particularly due to the development of HIV resistance. To evaluate an association between viral sequence data and drug combinations and to estimate an effect of a particular drug combination on the treatment results, collection of the most representative drug combinations used to cure HIV and the biological data on amino acid sequences of HIV proteins is essential. We have created a new, freely available web database containing 1,651 amino acid sequences of HIV structural proteins [reverse transcriptase (RT), protease (PR), integrase (IN), and envelope protein (ENV)], treatment history information, and CD4+ cell count and viral load data available by the user's query. Additionally, the biological data on new HIV sequences and treatment data can be stored in the database by any user followed by an expert's verification. The database is available on the web at http://www.way2drug.com/rhivdb.

Keywords: antiretroviral therapy, drug exposure, therapy success, database, human immunodeficiency virus, HIV, sequence data analysis, HIV drug resistance

INTRODUCTION

Human immunodeficiency virus (HIV) along with other viruses has a high social impact due its ability to spread from one person to another. According to the latest data[1], in 2020, the estimated number of new infection cases was over 1.5 million, while more than 38 million people are currently living with HIV [see text footnote 1]. All known antiretroviral drugs can only suppress viral replication but it is still impossible to eliminate the virus from human body completely (Geronikaki et al., 2016). Due to its high mutagenicity HIV is capable to develop resistance, to existing antiretroviral drugs (Geronikaki et al., 2016). Data on the amino acid sequences of HIV proteins, including reverse transcriptase (RT), protease (PR), integrase (IN), and envelope protein (ENV), are important for the prediction of HIV drug resistance (Liu and Shafer, 2006; Toor et al., 2011; Raposo and Nobre, 2017; Ramon et al., 2019; Steiner et al., 2020) and the so-called drug exposure, which is considered one of the features potentially associated with HIV drug resistance (Pironti et al., 2017). With data from the (i) amino acid sequences of HIV proteins, (ii) drug combinations used to treat HIV-positive patients, and (iii) clinical data obtained from the patients, it is possible to build

[1] http://www.who.int

models predicting (a) drug exposure and HIV drug resistance and (b) therapeutic effectiveness based on the HIV sequence data and the treatment history (Tarasova et al., 2020).

There are databases of amino acid and nucleotide sequences of HIV freely available for downloading and analysis (Kuiken et al., 2003; Rhee et al., 2003; Shafer, 2006). Particularly, Los Alamos National Laboratory (LANL) HIV sequence database contains over 900,000 sequences of HIV, which can be found by a user's query. Retrieved sequences can be aligned to assess their similarity with resistant samples or to investigate phylogeny. LANL HIV sequence database also contains premade alignments that can be used to investigate frequently occurred mutations, which may cause drug resistance. HIV drug resistance database (Rhee et al., 2003), developed and maintained at Stanford University, includes three main types of data: "genotype-phenotype," "genotype-treatment," and "genotype-clinical." "Genotype-phenotype" relationship includes information about HIV sequences and the data on their drug resistance/susceptibility, including resistance against HIV RT, PR, and IN inhibitors. It includes data on over 15,000 isolates tested on drug resistance in various assays. "Genotype-treatment" data includes over 300,000 sequences retrieved from HIV samples with the set of drugs taken by a patient. "Genotype-clinical" data contains over 1,500 episodes of the particular drug combinations taken by a patient along with some clinical data (CD4+ cell count and viral load at the time). There are statistics on the mutation prevalence, patterns of drug resistance mutations, and a summary of major and minor drug resistance positions. These databases are beneficial for HIV drug resistance analysis.

In addition to the databases that have already been developed, we have created a new, freely available web database, RHIVDB, to provide comprehensive data on HIV amino acid sequences, clinical data, and drug treatment history information. The main feature of RHIVDB is the availability of drug treatment history and clinical data for each record.

RHIVDB is developed based on the clinical data and the data on amino acid sequences of the HIV proteins collected in the Russian Federation in the Central Research Institute of Epidemiology. The database contains information about amino acid sequences of HIV proteins, drug combinations that were taken by a patient during a particular period, and CD4+ cell count and viral load data available for fast downloading on the user's query. The database can be used for determining the effectiveness of particular drug combinations, analysis of HIV sequences for various cohorts of patients, building models for prediction of therapeutic success based on sequence, clinical, and drug history data.

METHODS

Plasma samples were obtained as part of routine drug resistance testing in all federal districts of the Russian Federation. The dates of diagnosis include years from 1997 to 2019. Blood sampling dates ranged from January 2014 to December 2019.

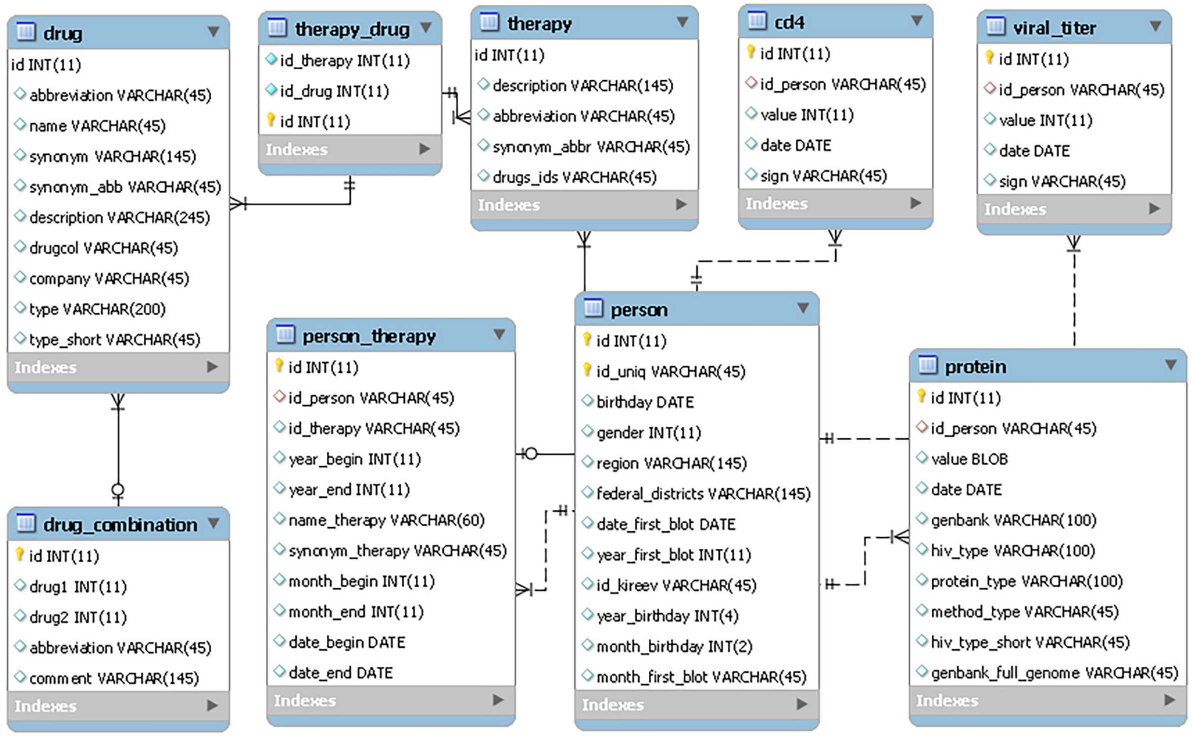


FIGURE 1 | The RHIVDB database schema.

TABLE 1 | Database characteristics. Number of records and values for each quantitative parameter of the database (A); number of records containing data on drug combinations and viral sequences (B).

(A)

Parameter	**Number of records**	**Mean**	**Standard deviation**
CD4+ cell count	1,732	343.3	266
Viral load (copies per ml)	1,823	62,475	57,625
Age	1,093	39	9.96

(B)

Parameter	**Number of records**
HIV RT and PRamino acid sequences	1,653
HIV IN amino acid sequences	281
HIV ENV amino acid sequences	276
Drug combination, total	434
Protease inhibitors	104
Reverse transcriptase inhibitors (NRTIs)	409
Reverse transcriptase inhibitors (NNRTIs)	344
Integrase inhibitors	31

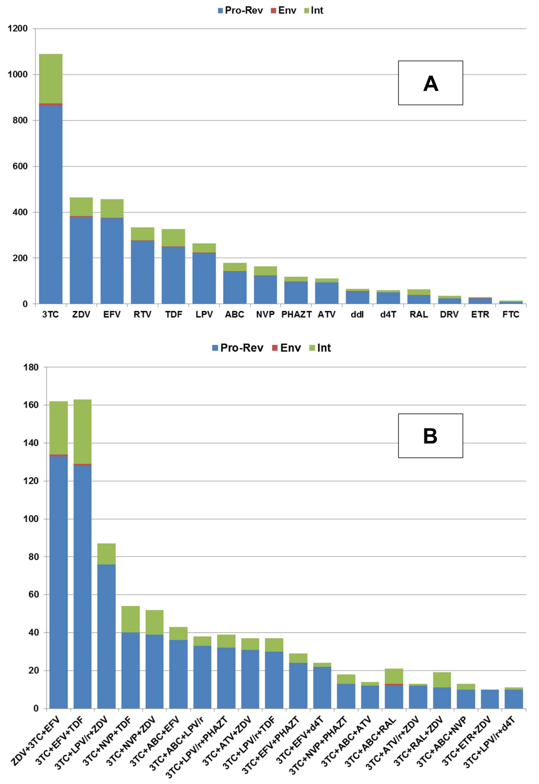


FIGURE 2 | The availability of the amino acid sequences (Y-axis) collected for each HIV protein in association with the particular amino acid sequences obtained from the patients taking a particular drug **(A)** or drug combination **(B)**.

RNA extraction and HIV genome amplification were carried out by ViroSeq HIV-1 Genotyping System (Abbott Molecular, United States) or AmpliSens HIV-Resist-Seq (Central Research Institute of Epidemiology, the Russian Federation). The amplified region of the pol region was at least 1,092 nucleotides length and covered positions 2,253–3,344 with respect to the reference HIV-1 strain HXB2 [GenBank: K03455 (Ratner et al., 1985)][2]. The amplified region of the env region was 420 nucleotides length and covered positions 6,954–7,374 of HXB2 strain. The nucleotide sequences of the pol and env regions were obtained using Sanger sequencing. Nucleotide sequences represented the part of the pol region encoding HIV PR and RT. Therefore amino acid sequences obtained from nucleotide sequences include corresponding PR and RT parts.

Data on HIV sequences with drug combinations used and data on CD4+ and viral load titer were processed for (i) duplicates removal; (ii) standardization of the drug names representation; (iii) verification of amino acid sequences data.

The RHIVDB web database uses the MySQL server to store data. The schema of the database is provided in **Figure 1**.

PHP and HTML codes were used to implement the main interface, and jQuery plugin DataTables for data accessing and manipulating (sorting, paging, and filtering). The scripts for data export and search were developed using PHP scripting language. They are available as a part of Supporting Information.

RESULTS

The RHIVDB database contains data on the amino acid sequences of HIV proteins, including the RT, PR, integrase IN, and HIV envelope proteins. In addition, it includes combinations of antiretroviral drugs taken during a particular time period and CD4+ cell count and viral load data during the periods of treatment in the database. The data stored in the RHIVDB do not contain any personal information about patients. The database is freely available on the web [see text footnote 3].

As of March 2021, the database contains 1,653 records on HIV-1 sequence data collected from different patients. For the web-accessible database, we chose only the records that consisted of both sequence and clinical data. For all 1,094 patients, there is information about CD4+ cell count and the number of HIV RNA copies per one ml. Sequence data on RT and PR are available for all 1,094 patients, while for IN and ENV, the data on sequences are available for 281 and 276 patients, respectively. For 434 records, there are data on the drug combinations taken by patients. The total numbers of records corresponding to each parameter are shown in **Table 1**.

The database interface provides data on therapy, with the periods during which the particular drug or a combination was taken, and the flag indicating therapy change during treatment. CD4+ cell count and viral load parameters measured in a certain

[2] https://www.ncbi.nlm.nih.gov/nuccore/1906382

data are provided in the columns "CD4" and "Viral load." The database includes the patient's information about age, gender, the date of diagnosis.

The user can perform a search using keywords ("Search" tab). Complex queries are available through the Filter option. It is possible to include several simple queries and combine them using Boolean operators "and", "or." Additionally, a user can quickly examine the records satisfying a particular CD4+ cell count or viral load (titer). Retrieved data can be easily exported in Microsoft Excel (CSV), Adobe Acrobat Reader (PDF), and Extensible Markup Language (XML) formats by selecting a particular option. Such options provide an easy way to process the data stored in the database.

Contributions to the database are possible for registered users who are signed in. After data verification by the experts, the information can be added into the database.

If antiviral drug resistance occurs, it is necessary to change a patient's antiretroviral therapy. On average, for each patient from the database, there are two schemas of therapy. The maximum number of therapy regimens per person is 14. The data stored

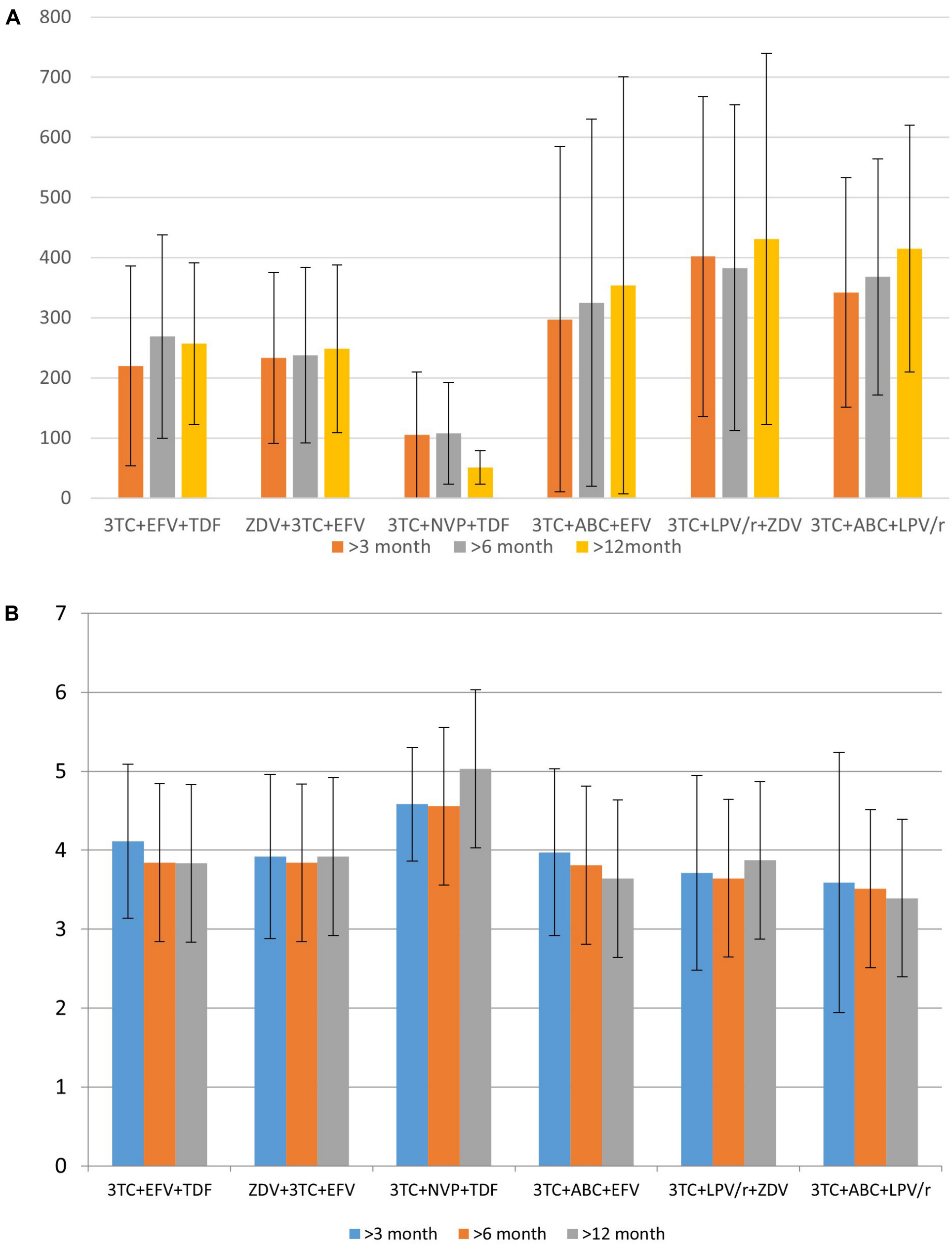


FIGURE 3 | The distribution of mean CD4+ cell count **(A)** and logarithmic values of viral load **(B)** for specific drug combination.

in the database allow the user to collect information about the therapy (a drug combination) and its effects on the viral load and CD4+ cell count.

DISCUSSION

The RHIVDB database information provides basis for the selection of the most effective treatment schema and for building models of treatment effectiveness based on clinical data (CD4+ cell count, viral load). The data on the amino acid sequences can be used along with treatment and clinical data to predict drug exposure or treatment effectiveness (Tarasova et al., 2020).

The correlations of the number of HIV-1 sequence to antiretroviral drug combination (A) and to the individual drugs (B) that the patient was taking before the sequence was determined are shown in **Figure 2**.

The database can help evaluate the therapeutic effectiveness and estimate the mutations' occurrence related to a patient's particular drug or drug combinations. Further, we demonstrate its applicability for two purposes: (i) search for CD4+ count and viral titer for particular drug combinations and (ii) evaluating the mutation frequency associated with nucleoside inhibitor abacavir as a case study.

Based on the data collected in the database, it is possible to identify some associations between drugs taken by a patient and CD4+ lymphocytes count or viral load. These parameters, along with the clinical symptoms, are used for the understanding of therapeutic success. We illustrate the applicability of the database for such purposes. **Figures 3A,B** display the distribution of the mean CD4+ cell count and various viral load values for specific drug combinations, respectively.

Data in **Figure 3** provides information regarding drug combinations characterized by the highest and lowest therapeutic efficacy for the cases included in the database. Additionally, in most cases, the average viral load values are remarkably similar to each other after 3 months, 6 months, and 1 year after the beginning of therapy; the same trend might be observed and for CD4+ cell count. It means that in most cases if a drug combination is effective 3 months after the beginning of the therapy, there is a high chance that it is effective after a year.

To demonstrate the applicability of the database to the estimation of amino acid substitutions prevalence associated with a particular drug, we performed such analysis for abacavir and zidovudine (nucleoside reverse transcriptase inhibitors) as a case study. We performed the search in the database and selected 101 and 247 amino acid sequences associated with abacavir and zidovudine taken by a patient, respectively. Based on the dates of therapy changes, we selected 39 sequences, for which the period of therapy with abacavir exceeded 90 days. The number of sequences associated with zidovudine (for the period over 90 days) was 245. If an exact date of therapy change is unknown, sequences obtained in the same year were excluded. These sequences were aligned using the ClustalW tool (Sievers et al., 2011). As a result, we obtained a set of substitutions associated with therapy schemes included abacavir and zidovudine (**Table 2**).

It is worth noting that some of them are included in the list of major drug mutations associated with therapy schemes included nucleoside reverse transcriptase inhibitors (for instance, 65 K/R, 74 V/L)[3], while other substitutions are not common. Interestingly, some of these sequences are characterized by substitutions at 101 and 103 positions, typically associated with resistance to NNRTIs. This example demonstrates that using RHIVDB it is possible to obtain some new information about substitutions that can be associated with the particular drug taken as a part of therapeutic drug combinations.

TABLE 2 | The substitutions appeared in the amino acid sequences of HIV RT of the samples, retrieved from patients that used nucleoside RT inhibitor abacavir or zidovudine in therapy schemes.

		Abacavir		Zidovudine	
Position	Substitution	N	%	N	%
35	T	23	59	158	64
	I	9	23	59	24
	A (R)	5	2	2	0.8
	V	2	5	19	7
	M	2	5	3	2
	K	2	5	4	2
36	D	22	56	88	36
	E	20	51	157	64
43	**K**	34	87	234	96
	E	2	5	6	2
	A	1	3	2	0.8
	R	2	5	3	1
49	**K**	36	92	239	97
	R	3	8	6	3
60	V	30	77	218	88
	I	9	23	27	12
65	**K**	33	85	195	79
	R	4	10	50	21
	N	2	5	0	0
70	**K**	35	90	232	94
	R	4	10	10	5
	E	0	0	2	0.8
	G	0	0	1	0.2
74	**L**	31	79	230	93
	V	17	44	10	5
90	**V**	33	85	215	87
	I	6	15	30	13
101	**K**	27	69	188	76
	E	7	18	55	23
	Q	4	10	0	0
	R			2	0.8
103	**K**	34	87	217	89
	N	5	13	28	11

The amino acid residue of the consensus HIV reverse transcriptase sequence is provided in bold.

[3] NRTI Resistance Notes – HIV Drug Resistance Database (https://www.stanford.edu/)

The further development of our database will provide an opportunity to collect data on various groups of patients who may have different susceptibilities to HIV infection (Lieberman et al., 2001; Jülg and Goebel, 2005; Gonzalo-Gil et al., 2017; Pironti et al., 2017; Lopez-Galindez et al., 2019; Ivanov et al., 2020). We believe that RHIVDB will help analyzing information about patients who do not develop a high viral load over a long time period. The information about the patients, sequence data, CD4+ cell count, and viral load may be used for developing the models of viremic control based on the patients' data and viral sequences. Therefore, the database can be helpful for developing personalized methods for HIV/AIDS treatment. These methods in particular may include the analysis of gene expression of the HIV-positive patients, analysis of a therapy regimen, allowing identify their individual reply to the particular combination of antiretroviral drugs.

CONCLUSION

We developed the database of HIV amino acid sequences containing the data on the combinations of antiretroviral therapy taken by a patient. Additionally, it contains information on the blood parameters that indicate the severity of HIV infection progress and the effectiveness of antiretroviral drug therapy. RHIVDB can be used by clinical specialists, biologists, bioinformatics for analysis of therapy effectiveness, HIV susceptibility and its resistance to antiretroviral therapy, and the variability of HIV sequences considering drug therapy. This database is available on the Internet for any user, it does not require registering an account. The biological data on new HIV sequences and data of therapy can be stored in the database by any user followed by the verification by an expert in the field of HIV epidemiology.

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: Dataryad and accession name RHIVDB data. https://datadryad.org/stash; https://datadryad.org/stash/share/iqb_UwuHd61_I5_z6bot9ui9PKoeEzBxpxX187vnEy0.

AUTHOR CONTRIBUTIONS

OT: idea, manuscript writing, and review. AR: database realization and manuscript writing. DK: collecting amino acid sequences, clinical data on CD4+ cell count, viral load, and manuscript editing. VP: manuscript review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by Russian Science Foundation grant 19-75-10097.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.679029/full#supplementary-material

REFERENCES

Geronikaki, A., Eleftheriou, P., and Poroikov, V. (2016). "Anti-HIV agents: current status and recent trends," in *Communicable Diseases of the Developing World in Topics in Medicinal Chemistry*, ed. A. K. Saxena (Cham: Springer), 37–95.

Gonzalo-Gil, E., Ikediobi, U., and Sutton, R. E. (2017). Mechanisms of virologic control and clinical characteristics of HIV+ elite/viremic controllers. *Yale J. Biol. Med.* 90, 245–259.

Ivanov, S., Lagunin, A., Filimonov, D., and Tarasova, O. (2020). Network-based analysis of OMICs data to understand the HIV–host interaction. *Front. Microbiol.* 11:1314. doi: 10.3389/fmicb.2020.01314

Jülg, B., and Goebel, F. D. (2005). Susceptibility to HIV/AIDS: an individual characteristic we can measure? *Infection* 33, 160–162. doi: 10.1007/s15010-005-6305-4

Kuiken, C., Korber, B., and Shafer, R. W. (2003). HIV sequence databases. *AIDS Rev.* 5, 52–61.

Lieberman, J. P., Manjunath, S. N., and Andersson, J. (2001). Dressed to Kill? A review of why antiviral CD8 T lymphocytes fail to prevent progressive immunodeficiency in HIV-1 infection. *Blood* 98, 1667–1677. doi: 10.1182/blood.v98.6.1667

Liu, T. F., and Shafer, R. W. (2006). Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin. Infect. Dis.* 42, 1608–1618. doi: 10.1086/503914

Lopez-Galindez, C., Pernas, M., Casado, C., Isabel Olivares, I., and Lorenzo-Redondo, R. (2019). Elite controllers and lessons learned for HIV-1 cure. *Curr. Opin. Virol.* 38, 31–36. doi: 10.1016/j.coviro.2019.05.010

Pironti, A., Pfeifer, N., Walter, H., Jensen, B.-E. O., Zazzi, M., Perpétua Gomes, P., et al. (2017). Using drug exposure for predicting drug resistance — a data-driven genotypic interpretation tool. *PLoS One* 12:e0174992. doi: 10.1371/journal.pone.0174992

Ramon, E., Belanche-Muñoz, L., and Pérez-Enciso, M. H. I. V. (2019). Drug resistance prediction with weighted categorical kernel functions. *BMC Bioinformatics* 20:410. doi: 10.1186/s12859-019-2991-2

Raposo, L. M., and Nobre, F. F. (2017). Ensemble classifiers for predicting HIV-1 resistance from three rule-based genotypic resistance interpretation systems. *J. Med. Syst.* 41:155. doi: 10.1007/s10916-017-0802-8

Ratner, L., Haseltine, W., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., et al. (1985). Complete nucleotide sequence of the AIDS Virus, HTLV-III. *Nature* 313, 277–284. doi: 10.1038/313277a0

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Bradley, B. J., Ravela, J., and Shafer, R. W. (2003). Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 31, 298–303. doi: 10.1093/nar/gkg100

Shafer, R. W. (2006). Rationale and uses of a public HIV drug-resistance database. *J. Infect. Dis.* 194(Suppl. 1), S51–S58. doi: 10.1086/505356

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* 7;539. doi: 10.1038/msb.2011.75

Steiner, M. C., Gibson, K. M., and Crandall, K. A. (2020). Drug resistance prediction using deep learning techniques on HIV-1 sequence data. *Viruses* 12:560. doi: 10.3390/v12050560

Tarasova, O., Biziukova, N., Kireev, D., Lagunin, A., Ivanov, S., Filimonov, D., et al. (2020). A computational approach for the prediction of treatment history and the effectiveness or failure of antiretroviral therapy. *Int. J. Mol. Sci.* 21:748. doi: 10.3390/ijms21030748

Toor, J. S., Sharma, A., Kumar, R., Gupta, P., Garg, P., and Arora, S. K. (2011). Prediction of drug-resistance in HIV-1 subtype C based on protease sequences from ART naive and first-line treatment failures in north india using genotypic and docking analysis. *Antiviral Res.* 92, 213–218. doi: 10.1016/j.antiviral.2011.08.005

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: 16 June 2021
doi: 10.3389/fgene.2021.678374

Ethnic Differences in the Frequency of *CFTR* Gene Mutations in Populations of the European and North Caucasian Part of the Russian Federation

***Nika Petrova*[1]*, *Natalia Balinova*[1], *Andrey Marakhonov*[1], *Tatyana Vasilyeva*[1], *Nataliya Kashirskaya*[1], *Varvara Galkina*[1], *Evgeniy Ginter*[1], *Sergey Kutsev*[1] and *Rena Zinchenko*[1,2]**

[1] Research Centre for Medical Genetics, Moscow, Russia, [2] N.A. Semashko National Research Institute of Public Health, Moscow, Russia

OPEN ACCESS

Edited by:
Elvira Galieva,
Novosibirsk State University, Russia

Reviewed by:
Daria Lichman,
Institute of Cytology and Genetics,
Russian Academy of Sciences, Russia
Felice Amato,
University of Naples Federico II, Italy
Ayman Salah EL-Seedy,
Alexandria University, Egypt
Anatoliy Ivashchenko,
Al-Farabi Kazakh National University,
Kazakhstan

***Correspondence:**
Nika Petrova
npetrova63@mail.ru

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *09 March 2021*
Accepted: *12 May 2021*
Published: *16 June 2021*

Citation:
Petrova N, Balinova N, Marakhonov A, Vasilyeva T, Kashirskaya N, Galkina V, Ginter E, Kutsev S and Zinchenko R (2021) Ethnic Differences in the Frequency of CFTR Gene Mutations in Populations of the European and North Caucasian Part of the Russian Federation. Front. Genet. 12:678374. doi: 10.3389/fgene.2021.678374

Cystic fibrosis (CF) is a common monogenic disease caused by pathogenic variants in the *CFTR* gene. The distribution and frequency of *CFTR* variants vary in different countries and ethnic groups. The spectrum of pathogenic variants of the *CFTR* gene was previously studied in more than 1,500 CF patients from different regions of the European and North Caucasian region of Russia and the spectrum of the most frequent pathogenic variants of the *CFTR* gene and ethnic features of their distribution were determined. To assess the population frequency of *CFTR* gene mutations some of the common variants were analyzed in the samples of healthy unrelated individuals from the populations of the European part of the Russian Federation: 1,324 Russians from four European regions (Pskov, Tver, Rostov, and Kirov regions), representatives of five indigenous ethnic groups of the Volga-Ural region [Mari (n = 505), Udmurts (n = 613), Chuvash (n = 780), Tatars (n = 704), Bashkirs (n = 517)], and six ethnic groups of the North Caucasus [Karachay (n = 324), Nogais (n = 118), Circassians (n = 102), Abazins (n = 128), Ossetians (n = 310), and Chechens (n = 100)]. The frequency of common *CFTR* mutations was established in studied ethnic groups. The frequency of F508del mutation in Russians was found to be 0.0056 on average, varying between four regions, from 0.0027 in the Pskov region to 0.0069 in the Rostov region. Three variants W1282X, 1677delTA, and F508del were identified in the samples from the North Caucasian populations: in Karachay, the frequency of W1282X mutation was 0.0092, 1677delTA mutation – 0.0032; W1282X mutation in the Nogais sample – 0.0127, the frequency of F508del mutations was 0.0098 and 1677delTA – 0.0098 in Circassians; in Abazins F508del (0.0039), W1282X (0.0039) and 1677delTA (0.0117) mutations were found. In the indigenous peoples of the Volga-Ural region, the maximum frequency of the F508del mutation was detected in the Tatar population (0.099), while this mutation was never detected in the Mari and Bashkir populations. The E92K variant was found in Chuvash and Tatar populations. Thus, interethnic differences in the spectra of *CFTR* gene variants were shown both in CF patients and in healthy population of the European and North Caucasian part of Russia.

Keywords: allelic polymorphism, *CFTR* gene variants, population differences, ethnic diversity, European and North Caucasian part of Russia

INTRODUCTION

Cystic fibrosis (CF; OMIM 219700) is a common monogenic disease caused by a mutation of the *CFTR* gene (CFTR, OMIM 602421; reference sequence accession number NM_000492.3). To date, more than 2,100 variants of the *CFTR* gene have been identified (Cystic Fibrosis Mutation Database, 2011), the distribution and frequency of which vary in different regions and ethnic groups (Bobadilla et al., 2002; Lao et al., 2003; Schrijver, 2011; World Health Organization [WHO], 2021]. In CF patients the most common mutations are F508del (66.8%), G542X (2.6%), N1303K (1.6%), G551D (1.5%), W1282X (1.0%), 1717-1G → A (0.83%), R553X (0.75%), 621 + 1G → T (0.54%), and R1162X (0.51%) (Estivill et al., 1997; World Health Organization [WHO], 2021). There is a decreasing proportion of CF patients with F508del from northwestern to southeastern Europe (Lucotte and Hazout, 1995; Bobadilla et al., 2002; Atag et al., 2019; Farrell et al., 2018), the highest frequency in Denmark (87.2%) and the lowest in Algeria (26.3%). Mutation G542X is common in the Mediterranean countries (6.1%). N1303K is found in most of the western and Mediterranean countries with the highest frequency in Tunisia (17.2%). G551D is common in north-west and central Europe. W1282X has the highest frequency in Israel (36.2%), being also common in most Mediterranean countries and North Africa (Estivill et al., 1997; Bobadilla et al., 2002; Lao et al., 2003; World Health Organization [WHO], 2021).

The population of the European part of Russia, represented by more than 70 ethnic groups, is about 109 million people. Previously, we studied the spectrum of pathogenic variants of the *CFTR* gene in more than 1,500 CF patients living in different regions of the European part of Russia and determined the spectrum of the most common pathogenic variants of the *CFTR* gene and the ethnic features of their distribution (Stepanova et al., 2012; Petrova et al., 2016, 2020a; Petrova N. V. et al., 2019). The difference in the spectra of the *CFTR* gene pathogenic variants in CF patients in different populations of Russia was shown. For ethnic Russian CF patients, a significant diversity of the spectrum of *CFTR* variants was shown: up to 98% of mutant alleles were caused by 110 variants, the most common were F508del (55%), CFTRdele2,3 (7.5%), 2143delA (2.7%), 3849 + 10kbC-T (2.3%), 2184insA (2.2%), N1303K (1.7%), G542X (1.5%), W1282X (1.2%), L138ins (1.1%), E92K (1.0%), and W1282R (0.7%) (Petrova et al., 2020a). The low diversity of the *CFTR* gene variant spectra was revealed in the ethnic groups of the North Caucasus region. The W1282X variant accounted for 88% of Karachays (Petrova et al., 2016), the high proportion of 1677delTA (81.5%) and E92K (12.5%) variants in Chechens (Petrova N. V. et al., 2019), and W1282X (50%) and F508del (20%) variants in Ossetians [Petrova et al., 2020b (in Russ.)] were found. The prevalence of E92K (55%) and F508del (30%) variants was noted among Chuvash, one of the ethnic groups of the Volga-Ural region (Stepanova et al., 2012). The distribution of relative frequencies of common *CFTR* variants in studied populations were shown in **Figure 1** (**Supplementary Table 1**).

Data on the mutation spectrum and prevalence of major *CFTR* gene mutations in various ethnic groups are important for the development of molecular diagnostics tools for identifying genetic causes of CF; however, data on the prevalence of common *CFTR* mutations in some populations are still not available.

Here, we present data on frequencies of major *CFTR* mutations in Russian Federation, in 15 populations of European Russia and North Caucasus.

MATERIALS AND METHODS

The total number of 5505 DNA samples of healthy unrelated individuals – representatives from 15 various populations of 12 ethnic groups living in the territory of European Russia have been studied. These are Russians (648 from Rostov, 354 – from Kirov, 182 – from Tver and 140 – from Kirov regions), Udmurts (*N* = 613), Maris (*N* = 505), Bashkirs (*N* = 517), Tatars (*N* = 704), Chuvashes (*N* = 780), Karachays (*N* = = 324), Nogais (*N* = 118), Cherkessians (*N* = 102), Abaza (*N* = 128), Chechens (*N* = 100), and Ossetians (*N* = 310). The mean age was 31 years (range 18–65), gender ratio – 0.40 males: 0.60 females. None of them had discernable symptoms suggestive of CF. The places of location of the populations under study are shown in **Table 1** and **Figure 2**.

Blood samples were collected during research expeditions in 1995–2018 by the staff of Laboratory of Genetic Epidemiology. The ethnic origin (up to the third generation) was defined by direct interview with examined persons. For this research, all DNA samples studied were anonymized. The DNA was extracted from whole blood samples collected in vacutainers with the preservative EDTA using commercial kits (Wizard Genomic DNA Purification Kit, Promega, United States) according to the manufacturer's recommendations.

The *CFTR* gene variants in (c.54-5940_273+10250del21kb (p.Ser18ArgfsX16; CFTRdele2,3), c.262_263delTT (p.Leu88 IlefsX22, 394delTT), c.411_412insCTA (p.Leu138dup; L138ins), c.1521_1523delCTT (p.Phe508del, F508del), c.1519_1521 delATC (p.Ile507del, I507del), c.1545_1546delTA (p.Tyr515X; 1677delTA), c.2012delT (p.Leu671X, 2143delT), c.2051_

Abbreviations: CF, cystic fibrosis; CFTR, cystic fibrosis transmembrane regulator; DNA, deoxyribonucleic acid; PCR, polymerase chain reaction; RFLP, restriction fragment length polymorphism.

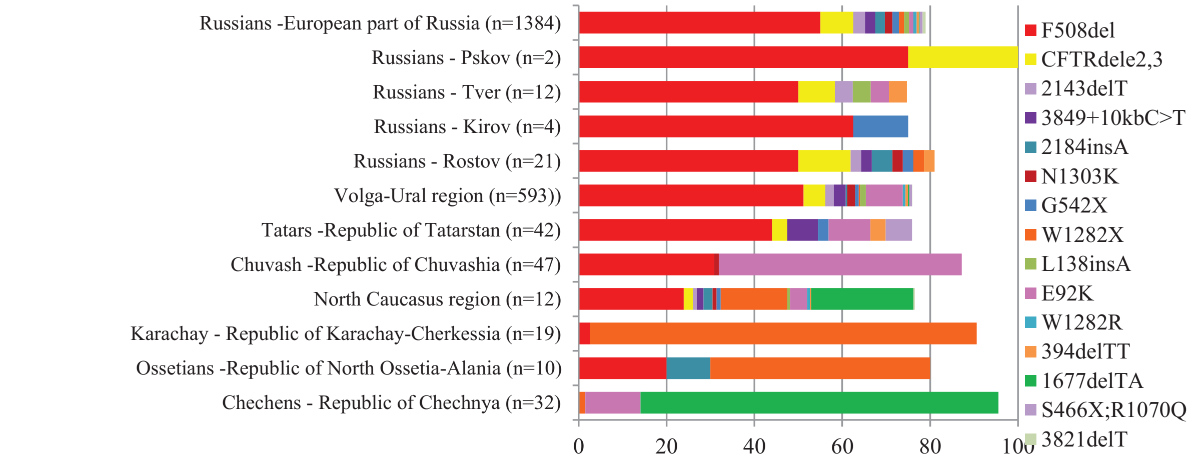


FIGURE 1 | Relative frequencies of common *CFTR* variants in studied regions and ethnic groups of European part and North Caucasus of Russia.

2052delAAinsG (p.Lys684SerfsX38, 2183AA>G), c.2052_2053insA (p.Gln685ThrfsX4; 2184insA), c.3691delT (p.Ser1231ProfsX4; 3821delT) by PCR/AFLP (amplified fragment length polymorphism) analysis, variants c.274G>A (p.Glu92Lys, E92K) and c.3846G > A (p.Trp1282X; W1282X) were tested by PCR/RFLP (restriction fragment length polymorphism) analysis according to previously described protocol (Petrova et al., 2008, 2020a; Petrova N. V. et al., 2019). Further, variant designation is given according to the "legacy" nomenclature.

The frequency of identified alleles was calculated according to the formula: $p_i = n_i/n$, where n_i is the number of i-th alleles, n is the sample size (the number of tested chromosomes) (Zhivotovsky, 1991). The Exact method was used to calculate 95% confidence intervals (95% CI) (Clopper and Pearson, 1934). The comparison of the population frequencies of variants in different samples was carried out using the Fisher test or χ^2-test with Yates correction, according to the generally accepted method (Zhivotovsky, 1991).

Maps of population frequency distribution for variants F508del, 1677delTA, W1282X, and E92K were constructed on the basis of data obtained in our study and on the basis of data on different populations of Europe calculated from the literature sources (Bobadilla et al., 2002; World Health Organization [WHO], 2021) (**Supplementary Table 2**) using the Bing maps add-in for Excel 365.

RESULTS

The frequency of eight *CFTR* mutations, CFTRdele2,3, F508del, 1677delTA, 2143delT, 2183AA > G, 2184insA, 394delTT, 3821delT, L138ins, E92K, and W1282X were analyzed in four Russians samples. In Kirov Russians 4 carriers of F508del mutation were found, in Tver Russians – one F508del carrier and one CFTRdele2,3 carrier, in Pskov Russians – one F508del carrier in Rostov Russians – 9 F508del carriers, one 1677delTA carrier and one carrier of W1282X (**Table 1**).

Population samples of five indigenous peoples of the Volga-Ural region were tested: Mari Udmurts, Bashkirs, Chuvash and Tatars. Only one carrier of E92K variant was found in the sample of Mari people. In the Chuvash and Udmurt samples, three and two carriers of the F508del mutation were found, respectively, and in the Chuvash sample, one carrier of the CFTRdele23 mutation and one carrier of the E92K variant were also found. In the Bashkir sample, the F508del mutation was not detected, but one carrier of the CFTRdele2,3 mutation was detected (**Table 1**). In the Tatar population, five of the tested variants were identified (**Table 1**): including the F508del mutation (9 carriers) with the maximum frequency for the indigenous peoples of the Volga-Ural region (0.0099, the frequency differences are significant) (**Supplementary Table 3**), the L138ins variant was found only in Tatars, and the E92K variant was found in Mari, Chuvash and Tatars.

Six indigenous ethnic populations of the North Caucasus region: two Turkic-speaking (Karachay and Nogais), two Abkhazian-Adyghe peoples (Abaza and Circassians), one Nakh-speaking – Chechens, and Iranian-speaking – Ossetians were studied. In samples from the studied North Caucasus populations carriers of W1282X, F508del and 1677delTA variants were identified. In Karachay, six carriers of W1282X mutation and two of 1677delTA; in Nogais three carriers of W1282X mutation; in Circassians one carrier of F508del and one of 1677delTA mutation; in Abaza – two carriers of F508del, two of W1282X variant, and one carrier of 1677delTA. Chechens had three carriers of 1677delTA variant, Ossetians – one carrier of F508del and two carriers of W1282X (0.0032) (**Table 1**).

DISCUSSION

Russians are the most numerous ethnic group in Russia. Up to 80 million ethnic Russians live in the European part of Russian Federation. For the present study four regions with a predominantly Russian population were selected: Pskov, located

TABLE 1 | Allele frequencies of identified mutations in studied populations of European part of Russia (Eastern European, Volga-Ural and North Caucasus regions).

Ethnic group (region)	Linguistic family/group	Sample size	Mutant/tested chromosomes frequencies (95% CI)					
			F508del	CFTRdele2,3	1677delTA	E92K	L138ins	W1282X
Russians (European part of Russia – 4 regions)	Indo-European/Slavic	1,324	15/2648 0.0057 (0.0032–0.0093)	1/2648 0.0004 (0.0000–0.0021)	1/2648 0.0004 (0.0000–0.0021)	0/810 0.0000 (0.0000–0.0037)	0/810 0.0000 (0.0000–0.0037)	1/890 0.0011 (0.0000–0.0062)
Russians (Rostov)		648	9/1296 0.0069 (0.0032–0.0131)	0/1296 0.0000 (0.0000–0.0023)	1/1296 0.0008 (0.0000–0.0043)	0/210 0.0000 (0.0000–0.0149)	0/210 0.0000 (0.0000–0.0149)	1/290 0.0034 (0.0001–0.0019)
Russians (Kirov)		354	4/708 0.0056 (0.0015–0.0144)	0/708 0.0000 (0.0000–0.0042)	0/708 0.0000 (0.0000–0.0042)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)
Russians (Tver)		182	1/364 0.0027 (0.0001–0.0152)	1/364 0.0027 (0.0001–0.0152)	0/364 0.0000 (0.0000–0.0089)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)
Russians (Pskov)		140	1/280 0.0036 (0.0001–0.0197)	0/280 0.0000 (0.0000–0.0106)	0/280 0.0000 (0.0000–0.0106)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)
Volgo-Ural region								
Mari	Ural-Yukaghir/Finno-Ugric	505	0/1010 0.0000 (0.0000–0.0030)	0/1010 0.0000 (0.0000–0.0.0030)	0/1010 0.0000 (0.0000–0.0030)	1/380 0.0026 (0.0001–0.0146)	0/300 (0.0000 (0.0000–0.0099)	0/380 0.0000 (0.0000–0.0079)
(Republic of Mari El)								
Udmurts (Republic of Udmurtia)	Ural-Yukaghir/Finno-Ugric	613	2/1206 0.0026 (0.0001–0.0146)	0/1206 0.0000 (0.0000–0.0025)	0/1206 0.0000 (0.0000–0.0025)	0/210 0.0000 (0.0000–0.0142)	0/210 0.0000 (0.0000–0.0142)	0/344 0.0000 (0.0000–0.0087)

(Continued)

TABLE 1 | Continued

Ethnic group (region)	Linguistic family/group	Sample size	Mutant/tested chromosomes frequencies (95% CI)					
			F508del	**CFTRdele2,3**	**1677delTA**	**E92K**	**L138ins**	**W1282X**
Chuvash (Republic of Chuvashia)	Altaic/Turkic	780	3/1560 0.0019 (0.0004–0.0056)	1/1560 0.0006 (0.0000–0.0036)	0/1560 0.0000 (0.0000–0.0019)	1/224 0.0045 (0.0001–0.0246)	0/224 0.0000 (0.0000–0.0133)	0/328 0.0000 (0.0000–0.0091)
Bashkirs (Republic of Bashkiria)	Altaic/Turkic	517	0/1034 0.0010 (0.000–0.0031)	1/1034 0.001 (0.0000–0.0054)	0/1034 0.0000 (0.0000–0.0029)	0/510 0.0000 (0.0000–0.0059)	0/510 0.0000 (0.0000–0.0059)	0/534 0.0000 (0.0000–0.0056)
Tatars (Republic of Tatarstan)	Altaic/Turkic	707	14/1414 0.0099 (0.0054–0.0166)	1/1414 0.0007 (0.0000–0.0039)	1/1414 0.0007 (0.0000–0.0039)	3/844 0,0036 (0.0007–0.0104)	4/1414 0.0028 (0.0008–0.0072)	0/400 0.0000 (0.0000–0.0075)
North Caucasus region								
Karachay (Republic of Karachay-Cherkessia)	Altaic/Turkic	324	0/648 0.0000 (0.0000–0.0046)	0/648 0.0000 (0.0000–0.0046)	2/648 0.0031 (0.0004–0.0011)	0/648 0.0000 (0.0000–0.0046)	0/648 0.0000 (0.0000–0.0046)	6/648 0.0093 (0.0034–0.0200)
Nogais (Republic of Karachay-Cherkessia)	Altaic/Turkic	118	0/236 0.0000 (0.0000–0.0126)	0/236 0.0000 (0.0000–0.0126)	0/236 0.0000 (0.0000–0.0126)	0/236 0.0000 (0.0000–0.0126)	0/236 0.0000 (0.0000–0.0126)	3/236 0.0127 (0.0026–0.0367)
Circassians (Republic of Karachay-Cherkessia)	North Caucasian/Abkhazian-Adyghe	102	2/204 0.0098 (0.0012–0.0350)	0/204 0.0000 (0.0000–0.0146)	2/204 0.0098 (0.0012–0.0350)	0/204 0.0000 (0.0000–0.0146)	0/204 0.0000 (0.0000–0.0146)	0/204 0.0000 (0.0000–0.0146)
Abaza (Republic of Karachay-Cherkessia)	North Caucasian/Abkhazian-Adyghe	128	1/256 0.0039 (0.0001–0.0216)	0/256 0.0000 (0.0000–0.0116)	3/256 0.0117 (0.0024–0.0339)	0/256 0.0000 (0.0000–0.0116)	0/256 0.0000 (0.0000–0.0116)	1/256 0.0039 (0.0001–0.0216)
Ossetians [Republic of North Ossetia–Alania)]	Indo-European/Iranian	310	1/620 0.0016 (0.0000–0.0090)	0/620 0.0000 (0.0000–0.0048)	0/620 0.0000 (0.0000–0.0048)	0/620 0.0000 (0.0000–0.0048)	0/620 0.0000 (0.0000–0.0048)	2/620 0.0032 (0.0004–0.0116)
Chechens(Republic of Chechnya)	North Caucasian/Nakh-Dagestan	100	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.149)	3/200 0.015 (0.0031–0.0051)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)	0/200 0.0000 (0.0000–0.0149)

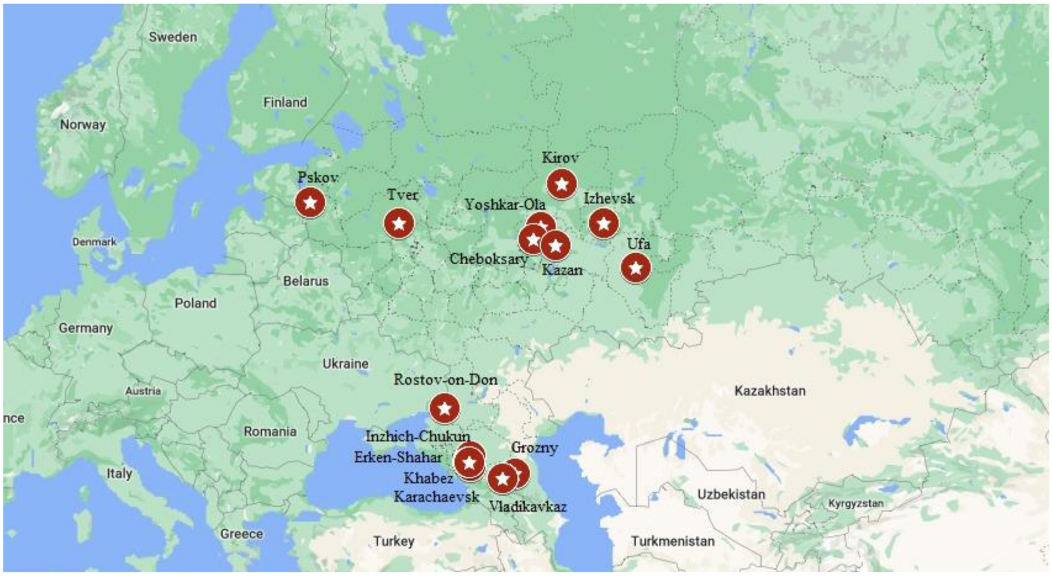


FIGURE 2 | Location of studied populations. Russians – Pskov, Tver, Kirov, Rostov-on-Don. Volga-Ural region: Yoshkar-Ola – Mari; Cheboksary – Chuvashes; Kazan – Bashkirs; Izhevsk – Udmurts; Ufa – Bashkirs. North Caucasus region: Inzhich-Chukum – Abaza; Erken-Shakar – Nogais; Khabez – Circassians; Karachaevsk – Karachay; Vladikavkaz – Ossetians; Grozny – Chechens.

TABLE 2 | F508del mutation frequencies in some populations of the world (comparison on Russians of European part of Russia).

Population (references)	No. mutations/no. chromosomes	Mutation frequency	*P*-value
Russians/European Russia/(Abramov et al., 2015)	15/2,000	0.0075	0.4391
Scotland (Brock et al., 1998)	816/54,322	0.01503	<0.0001
Denmark (Brandt et al., 1994)	172/13,198	0.01303	0.0014
Italy (Gasparini et al., 1999)	90/8,952	0.0101	0.0362
Israel (Jews-Ashkenazi) (Kalman et al., 1994)	35/3,892	0.0089	0.1293
Estonia (Teder et al., 2000)	88/14,792	0.0059	0.8603
India (Kapoor et al., 2006)	4/1,910	0.0021	0.1067

in the west, Tver – in the center, Rostov – in the south, and Kirov – in the north-east of the European part of Russia. In Russian samples from four regions, four different mutations in the CFTR gene were found: F508del, CFTRdele2, 3 (21kb), 1677delTA, and W1282X, but only F508del was found in all regions, varying in frequency from 0.0027 in the Pskov region to 0.0069 in the Rostov region. The differences in the F508del mutation frequency between Russian samples are not significant (**Supplementary Table 4**). The average frequency of the F508del mutation in Russians is 0.0056.

To calculate a more accurate value of the F508del mutation frequency in Russians of the European part of Russia, all samples can be combined. The frequency of F508del revealed in Russians was 0.0056, which is comparable to the data obtained by other researchers studying individuals from Russian populations of central regions of Russia (Abramov et al., 2015), but significantly lower than in a number of European populations (**Table 2** and **Figure 3**). Thus, the highest population frequencies of F508del mutation were observed in the north-west of Western Europe, reaching in Scotland and Denmark – 0.015 and 0.013, respectively (Brandt et al., 1994; Brock et al., 1998); in the Mediterranean countries, the frequency of F508del mutation was lower: for example, in Italy – 0.010 (Gasparini et al., 1999), and in Israel (among Ashkenazi Jews) – 0.0089 (Kalman et al., 1994; Quint et al., 2005; World Health Organization [WHO], 2021). In Estonia, the F508del variant frequency was 0.0059 (Teder et al., 2000), which is not significantly different from the one obtained for Russians in the European part of Russia. The relative frequency of F508del mutation in CF patients decreases from northwestern to southeastern Europe (Bobadilla et al., 2002; Farrell et al., 2018; World Health Organization [WHO], 2021). Apparently, the population frequency of the F508del mutation also changes. This is also consistent with the low frequency of F508del mutation observed in the indigenous population of India (0.00209) (Kapoor et al., 2006).

The Volga-Ural region of Russia is situated at the border of Europe and Asia and during historical times was a place of interaction of many ethnic groups (Alekseev, 1974; Kuzeev, 1985).

Three Turkic-speaking groups (Tatars, Chuvash, and Bashkirs) and two Finno-Ugric groups (Mari and Udmurts) were studied. In two of studied Turkic-speaking populations

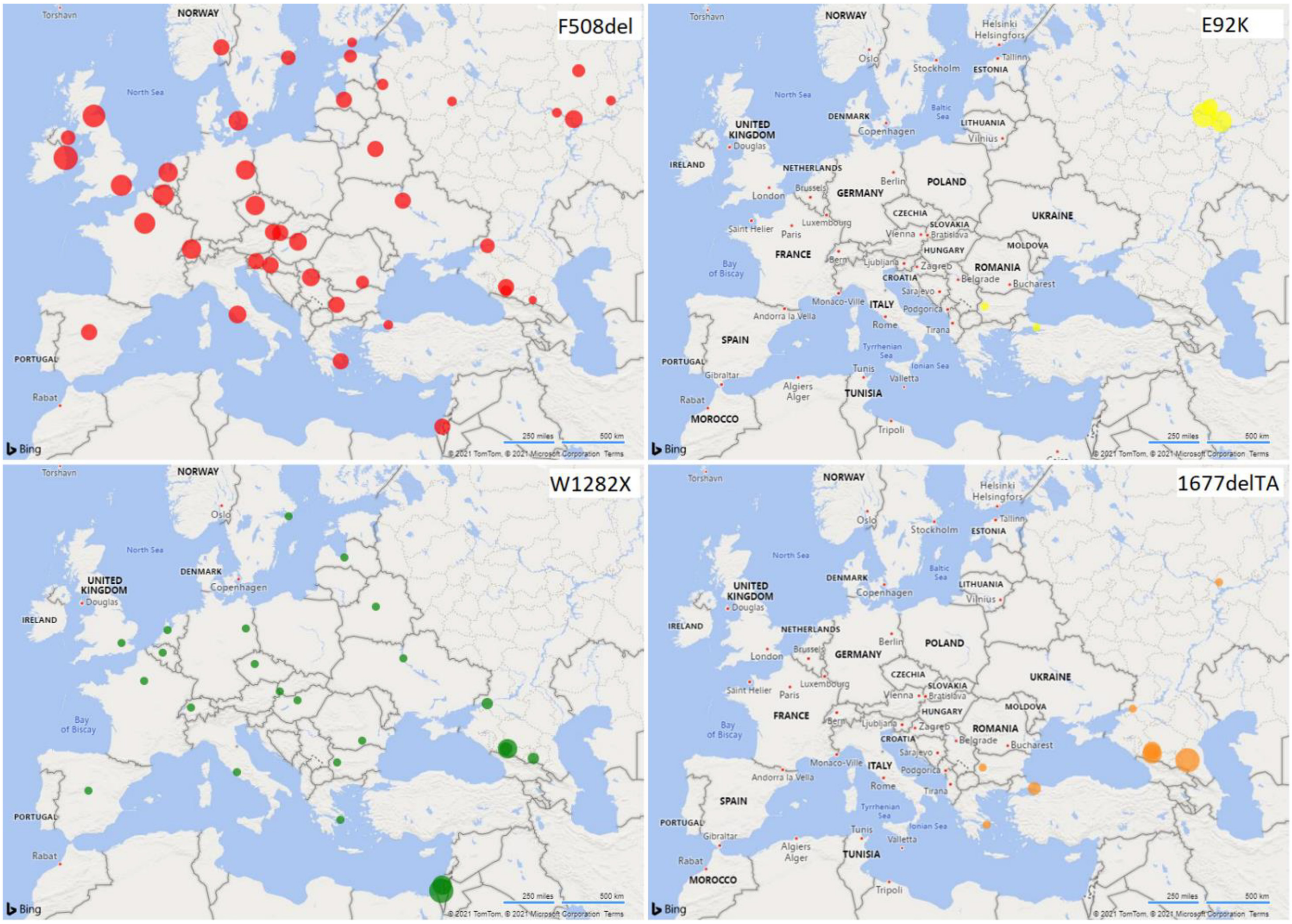


FIGURE 3 | Distribution of *CFTR* variant frequencies in European populations. The radii of the circles reflect the frequency values.

of the Volga-Ural region, F508del mutation was found with relatively high frequencies of 0.0099, 0.0019 in Tatars and Chuvashes, respectively, but was not revealed in Bashkirs. Finno-Ugric populations of the Volga-Ural region demonstrated a low frequency of F508del mutation in Udmurts (0.0016) and its absence in Maries (**Table 1**). When comparing the frequency of F508del with Russians, the differences were significant only for the Mari and Bashkirs (p = 0.0351 and 0.0326) (**Supplementary Table 3**).

Stepanova et al. (2012) showed that in Chuvash CF patients, the predominant cause of the disease was E92K and F508del variants, the carrier frequency of the E92K mutation is 1: 68 (5/343 persons), and the F508del mutation is 1: 86 (4/343). The differences in the frequencies of these two variants in Stepanova's work and in our work are not significant (p = 1.000 and 0.2630, correspondently). Among the studied Turkic-speaking groups of the Volga-Ural region, variants E92K and F508del were found in the Tatar population, while these variants were not found in the Bashkir population. It should be noted that according to the Russian CF Patients Registry-2018 (RCFPR-2018), the CF incidence in Bashkirs is significantly lower than in Chuvash and Tatars (MEDPRAKTIKA-M, 2019).

The North Caucasus region is characterized by a wide variety of ethnic populations, complicated history of the formation of ethnic groups and high genetic diversity (Alekseev, 1974). The F508del variant was not detected in the Turkic-speaking populations of the North Caucasus (Karachay and Nogais) and in the Chechens: the W1282X variant was predominant in the former, and the 1677delTA variant in the latter (**Figure 3**). A significant difference between the samples of Ossetians and the samples of Abaza ($p < 0.05$), Chechens ($p < 0.05$), and Circassians ($p < 0.1$) in the 1677delTA variant frequency is shown (**Supplementary Table 5**). When comparing ethnic groups of the Volga-Ural region and the North Caucasus region, significant differences in the frequency of F508del were found between Karachays and Tatars ($p < 0.05$), as well as between Circassians and Mari (p = 0.0243) and Bashkirs (0.0256) (**Table 2**, **Supplementary Table 6**, and **Figure 3**).

The W1282X mutation was assumed to occur as a single mutation event in a population of Middle Eastern Jews before their migration to Europe (World Health Organization). Further distribution of this mutation in various regions was connected with the resettlement of Ashkenazi Jews. The W1282X mutation was found in different regions of the world (**Figure 3**). The highest frequency of the mutation was found in the population of Ashkenazi Jews (up to 50% of the mutant alleles among CF patients, carrier frequency – 1: 54 and population frequency – 0.0092) (Kalman et al., 1994; Quint et al., 2005). The high population frequency of W1282X mutation was found in Turkic-speaking North Caucasus groups (Karachay and Nogais, 0.0092 and 0.0132), in Abaza (0.0039) and in Ossetians (0.0032). The 1677delTA mutation was previously found to be common in populations neighboring or with historic links to the greater Black Sea region (e.g., Bulgaria, Romania, Greece, Cyprus, and Turkey [Estivill et al., 1997; Bobadilla et al., 2002; Atag et al., 2019; Petrova G. et al., 2019; World Health Organization [WHO], 2021)], including Northern Iran and Georgia (Ivashchenko and Baranov, 2002). We found the high population frequencies of 1677delTA variant in such autochthonous populations of the North Caucasus as Abkhazian-Adyghe [Abaza (0.0171) and Circassians (0.0098)] and Nakh [Chechens (0.0150)] groups, but not in Ossetians and Nogais (**Figure 3**). Significant differences in the 1677delTA variant frequency were shown in Abkhazian-Adyghe (Circassians, $p < 0.05$; and Abazins, $p < 0.01$) and Nakh groups (Chechens, $p < 0.01$) compared to all studied ethnic groups of Volga-Ural region (**Table 2** and **Supplementary Table 6**).The data obtained in this study allow, to a certain extent, to fill the gap in information on the prevalence of the F508del, E92K, 1677delTA, and W1282X variants of the *CFTR* gene in some indigenous ethnic groups living on the territory of European Russia, and to get an entire picture of the prevalence lapse rate in the considered region. Further studies are necessary to consider the importance of extensive study of the CF pathogenic variants in the populations of the European and North Caucasian part of the Russian Federation, by direct gene sequencing to determine the molecular basis of CF in Russian Federation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethical Committee of Research Centre for Medical Genetics (Research Centre for Medical Genetics, 115522, Moscow, Moskvorechie St., 1, Russian Federation, Protocol No 17/2006 of 02.02.2006). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NP: conceptualization and writing – original draft preparation. RZ and VG: resources. NP and NB: investigation. TV: validation. NP and AM: formal analysis. NK: data curation. AM, TV, NK, and EG: writing – review and editing. SK and RZ: project administration and funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was partially funded by the RSF grant (project 17-15-01051) and was carried out as part of the State Assignment for the Research Centre of Medical Genetics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.678374/full#supplementary-material

REFERENCES

Abramov, D. D., Kadochnikova, V. V., Yakimova, E. G., Belousova, M. V., Maerle, A. V., Sergeev, I. V., et al. (2015). High carrier frequency of CFTR gene mutations associated with cystic fibrosis, and PAH gene mutations associated with phenylketonuria in Russian population. *VESTNIK RGMU.RU* 4, 32–35.

Alekseev, V. (1974). *The Geography of Human Races*, 1st Edn. Moscow: Nauka.

Atag, E., Ikizoglu, N. B., Ergenekon, A. P., Gokdemir, Y., Eralp, E. E., Ata, P., et al. (2019). Novel mutations and deletions in cystic fibrosis in a tertiary cystic fibrosis center in Istanbul. *Pediatr. Pulmonol.* 54, 743–750. doi: 10.1002/ppul.24299

Bobadilla, J. L., Macek, M. Jr., Fine, J. P., and Farrell, P. M. (2002). Cystic fibrosis: a worldwide analysis of CFTR mutations - correlation with incidence data and application to screening. *Hum. Mutaṯ.* 19, 575–606. doi: 10.1002/humu.10041

Brandt, N. J., Schwartz, M., and Skovby, F. (1994). Screening for carriers of cystic fibrosis. result of a pilot study among pregnant women. *Ugeskr. Laeger* 156, 3751–3757.

Brock, D. J., Gilfillan, A., and Holloway, S. (1998). The incidence of cystic fibrosis in Scotland calculated from heterozygote frequencies. *Clin. Genet.* 53, 47–49. doi: 10.1034/j.1399-0004.1998.531530109.x

Clopper, C. J., and Pearson, S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404–413.

Cystic Fibrosis Mutation Database (2011). Available online at: http://www.genet.sickkids.on.ca (Accessed March 1, 2021)

Estivill, X., Bancells, C., and Ramos, C. (1997). Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. the biomed CF mutation analysis consortium. *Hum. Mutat.* 10, 135–154. doi: 10.1002/(SICI)1098-1004199710:2<135::AID-HUMU6<3.0.CO;2-J

Farrell, P., Férec, C., Macek, M., Frischer, T., Renner, S., Riss, K., et al. (2018). Estimating the age of p.(Phe508del) with family studies of geographically distinct European populations and the early spread of cystic fibrosis. *Eur. J. Hum. Genet.* 26, 1832–1839. doi: 10.1038/s41431-018-0234-z

Gasparini, P., Arbustini, E., Restagno, G., Zelante, L., Stanziale, P., Gatta, L., et al. (1999). Analysis of 31 CFTR mutations by polymerase chain reaction/oligonucleotide ligation assay in a pilot screening of 4476 newborns for cystic fibrosis. *J. Med. Screen.* 6, 67–69. doi: 10.1136/jms.6.2.67

Ivashchenko, T. E., and Baranov, V. S. (2002). *Biohemical and Molecular Genetic Basics of Cystic Fibrosis Pathogenesis*. Intermedika: Saint-Petersburg.

Kalman, Y. M., Kerem, E., Darvasi, A., DeMarchi, J., and Kerem, B. (1994). Difference in frequencies of the cystic fibrosis alleles, delta F508 and W1282X, between carriers and patients. *Eur. J. Hum. Genet.* 2, 77–82. doi: 10.1159/000472347

Kapoor, V., Shastri, S. S., Kabra, M., Kabra, S. K., Ramachandran, V., Arora, S., et al. (2006). Carrier frequency of F508del mutation of cystic fibrosis in Indian population. *J. Cyst. Fibros* 5, 43–46. doi: 10.1016/j.jcf.2005.10.002

Kuzeev, R. (1985). *The Peoples of the Volga and Ural Regions*, 1st Edn. Moscow: Nauka.

Lao, O., Andres, A. M., Mateu, E., Bertranpetit, J., and Calafell, F. (2003). Spatial patterns of cystic fibrosis mutations spectra in European populations. *Eur. J. Hum. Genet.* 11, 385–394. doi: 10.1038/sj.ejhg.5200970

Lucotte, G., and Hazout, S. (1995). Geographic and ethnic distribution of the more frequent cystic fibrosis mutations in Europe show that a founder effect is apparent for several mutant alleles. *Hum. Biol.* 67, 561–576.

MEDPRAKTIKA-M (2019). *Register of Patients with Cystic Fibrosis in the Russian Federation*. 2017. eds A. Y. Voronkova, E. A. Amelina, N. Y. Kashirskaya, E. I. Kondratieva, S. A. Krasovsky, M. A. Starinova, et al. (Moscow: Medpraktika-M).

Petrova, G., Yaneva, N., Hrbková, J., Libik, M., Savov, A., and Macek, M. Jr. (2019). Identification of 99% of CFTR gene mutations in Bulgarian-, Bulgarian Turk-, and Roma cystic fibrosis patients. *Mol. Genet. Genom. Med.* 7:e696. doi: 10.1002/mgg3.696

Petrova, N. V., Kashirskaya, N. Y., Kondratyeva, E. I., Getoeva, Z. K., Vasilieva, T. A., Voronkova, A. Y., et al. (2020a). The features of spectrum and frequency of CFTR gene mutations in populations of Southern Russia and Northern Caucasus. *Med. Newse North Caucasus* 2, 174–178. doi: 10.14300/mnnc.2020.15042

Petrova, N. V., Kashirskaya, N. Y., Saydaeva, D. K., Polyakov, A. V., Adyan, T. A., Simonova, O. I., et al. (2019). Spectrum of CFTR mutations in Chechen cystic fibrosis patients: high frequency of c.1545_1546delTA (p.Tyr515X; 1677delTA) and c.274G>A (p.Glu92Lys, E92K) mutations in North Caucasus. *BMC Med. Genet.* 20:44. doi: 10.1186/s12881-019-0785-z

Petrova, N. V., Kashirskaya, N. Y., Vasilyeva, T. A., Kondratyeva, E. I., Zhekaite, E. K., Voronkova, A. Y., et al. (2020b). Analysis of CFTR mutation spectrum in ethnic Russian cystic fibrosis patients. *Genes* 11:554. doi: 10.3390/genes11050554

Petrova, N. V., Kashirskaya, N. Y., Vasilyeva, T. A., Timkovskaya, E. E., Voronkova, A. Y., Shabalova, L. A., et al. (2016). High proportion of W1282X mutation in CF patients from Karachai-Cherkessia. *J. Cyst. Fibros* 15, e28–e32. doi: 10.1016/j.jcf.2016.02.003

Petrova, N. V., Timkovskaya, E. E., and Zinchenko, R. A. (2008). The analysis of CFTR mutations frequencies in different populations of Russia. *Eur. J. Hum. Genet.* 16:38.

Quint, A., Lerer, I., Sagi, M., and Abeliovich, D. (2005). Mutation spectrum in Jewish cystic fibrosis patients in Israel: implication to carrier screening. *Am. J. Med. Genet. A.* 136, 246–248. doi: 10.1002/ajmg.a.30823

Schrijver, I. (2011). Mutation distribution in expanded screening for cystic fibrosis: making up the balance in a context of ethnic diversity. *Clin. Chem.* 57, 799–801. doi: 10.1373/clinchem.2011.164673

Stepanova, A. A., Abrukova, A. V., Savaskina, E. N., and Polyakov, A. V. (2012). Mutation p.E92K is the primary cause of cystic fibrosis in Chuvashes. *Russ. J. Genet.* 48, 731–737. doi: 10.1134/S1022795412060166

Teder, M., Klaassen, T., Oitmaa, E., Kaasik, K., and Metspalu, A. (2000). Distribution of CFTR gene mutations in cystic fibrosis patients from Estonia. *J. Med. Genet.* 37:E16. doi: 10.1136/jmg.37.8.e16

World Health Organization [WHO] (2021). *The Molecular Genetic Epidemiology of Cystic Fibrosis*. Geneva: WHO.

Zhivotovsky, L. A. (1991). *Population-Based Biometrics*. Moscow: Nauka.

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 August 2021
doi: 10.3389/fgene.2021.701373

Heterogeneity of Genetic Admixture Determines SLE Susceptibility in Mexican

Susana Hernández-Doño[1], Juan Jakez-Ocampo[2], José Eduardo Márquez-García[3], Daniela Ruiz[4], Víctor Acuña-Alonzo[5], Guadalupe Lima[2], Luis Llorente[2], Víctor Hugo Tovar-Méndez[6], Rafael García-Silva[7], Julio Granados[1]*, Joaquín Zúñiga[8,9] and Gilberto Vargas-Alarcón[10]

[1] Immunogenetics Division, Department of Transplant, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico, [2] Department of Immunology and Rheumatology, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico, [3] Molecular Biology Core Facility, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico, [4] Department of Dermatology, Hospital General Dr. Manuel Gea González, Mexico City, Mexico, [5] Laboratory of Physiology, Biochemistry, and Genetics, Escuela Nacional de Antropología e Historia, Mexico City, Mexico, [6] Department of Endocrinology, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico, [7] Department of Internal Medicine, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico, [8] Laboratory of Immunobiology and Genetics, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico, [9] Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud, Mexico City, Mexico, [10] Research Direction, Instituto Nacional de Cardiología Ignacio Chavez, Mexico City, Mexico

OPEN ACCESS

Edited by:
Elvira Galieva,
Novosibirsk State University, Russia

Reviewed by:
Cesar Fortes-Lima,
Uppsala University, Sweden
Francesca Lantieri,
University of Genoa, Italy

***Correspondence:**
Julio Granados
julgrate@yahoo.com
orcid.org/0000-0001-8553-1027

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *27 April 2021*
Accepted: *12 July 2021*
Published: *03 August 2021*

Citation:
Hernández-Doño S, Jakez-Ocampo J, Márquez-García JE, Ruiz D, Acuña-Alonzo V, Lima G, Llorente L, Tovar-Méndez VH, García-Silva R, Granados J, Zúñiga J and Vargas-Alarcón G (2021) Heterogeneity of Genetic Admixture Determines SLE Susceptibility in Mexican. Front. Genet. 12:701373. doi: 10.3389/fgene.2021.701373

Systemic Lupus Erythematosus (SLE) is an autoimmune inflammatory disorder for which Major Histocompatibility Complex (MHC) genes are well identified as risk factors. SLE patients present different clinical phenotypes, which are partly explained by admixture patterns variation among Mexicans. Population genetic has insight into the high genetic variability of Mexicans, mainly described through HLA gene studies with anthropological and biomedical importance. A prospective, case-control study was performed. In this study, we recruited 146 SLE patients, and 234 healthy individuals were included as a control group; both groups were admixed Mexicans from Mexico City. The HLA typing methods were based on Next Generation Sequencing and Sequence-Based Typing (SBT). The data analysis was performed with population genetic programs and statistical packages. The admixture estimations based on HLA-B and -DRB1 revealed that SLE patients have a higher Southwestern European ancestry proportion (48 ± 8%) than healthy individuals (30 ± 7%). In contrast, Mexican Native American components are diminished in SLE patients (44 ± 1%) and augmented in Healthy individuals (63 ± 4%). HLA alleles and haplotypes' frequency analysis found variants previously described in SLE patients from Mexico City. Moreover, a conserved extended haplotype that confers risk to develop SLE was found, the HLA-A*29:02∼C*16:01∼B*44:03∼DRB1*07:01∼DQB1*02:02, $pC = 0.02$, OR = 1.41. Consistent with the admixture estimations, the origin of all risk alleles and haplotypes found in this study are European, while the protection alleles are Mexican Native American. The analysis of genetic distances supported that the SLE patient group is closer to the Southwestern European parental populace and farthest from Mexican Native Americans than healthy individuals. Heterogeneity of genetic admixture determines SLE susceptibility and protection in Mexicans. HLA sequencing is helpful to determine susceptibility alleles and haplotypes restricted to some populations.

Keywords: HLA, Mexican, population genetics, heterogeneity, HLA-DRB1*03:01, admixture, conserved extended haplotype

INTRODUCTION

Systemic Lupus Erythematosus (SLE) is a chronic autoimmune disease characterized by the loss of tolerance to self-antigens and interferon responses dysregulation. SLE manifestations are diverse; the condition can affect almost any organ in the body (Ghodke-Puranik and Niewold, 2015). Patients with Hispanic, African, and Asian ancestry develop SLE earlier than European populations. These patients also have more acute disease onset, more severe clinical manifestations, higher disease activity, chronic organ damage, and higher mortality (Carter et al., 2016). Hispanic SLE prevalence is ∼138/100000 per inhabitants per year, which is higher than in Asian and European populations (Atisha-Fregoso et al., 2011). Many differences in the disease presentation across the ethnic barrier have been explained throughout genetic predisposition. One of the most studied systems is the Major Histocompatibility Complex (MHC) class I and class II genes, known as Human Leukocyte Antigen (HLA) Class I and II. The HLA variability among populations is helpful to explain many characteristics of SLE (Tsokos et al., 2016), and it continues to give more information about the genetic predisposition and pathophysiological mechanism of the disease.

As mentioned before, HLA haplotypes confer susceptibility or protection in an ethnic-dependent manner (Vargas-Alarcón et al., 2001; Vasconcelos et al., 2009; Furukawa et al., 2014; Alarcón-Riquelme et al., 2016; Molineros et al., 2019). In Mexico, susceptibility varies as the ethnic admixture does, the admixture in Mexico is very heterogeneous across the Country, contributing to the disease variability (Moreno-Estrada et al., 2014; Barquera et al., 2020b). Therefore, the admixture diversity has contributed to the enrichment of susceptibility markers and an ample SLE phenotypes specter in Mexicans (Salgado-Galicia et al., 2020). For instance, the most common susceptibility allele, HLA-DRB1*03:01, previously identified in Mexico City, was not found in Tapachula Chiapas SLE patients. Otherwise, HLA-DR2 Chiapas patients showed a higher risk of developing SLE once infected with Zika or Chikungunya viruses common in that region but absent in Mexico City (Sepúlveda Delgado et al., 2018). Besides, population genetic studies revealed significantly different admixture estimates for Mexico City and Tapachula, Chiapas (Barquera et al., 2020e).

Equally important, conserved extended haplotypes (CEHs) help to defined susceptibility, and the alleles help MHC genetic diversity measurements in autoimmune conditions. CEHs are DNA blocks defined as combinations of HLA-B, -DR, complement, and other immune-related genes. CEHs are known as DNA stretches with fixed alleles, including those at loci not tested (Wescott et al., 1987). Thus, CEHs have a high genetic load, so determining HLA CEHs gives information about the aggregated risk confer by non-classical and non-HLA genes linked to HLA. The frequency and allele combination of CEHs varies between major ethnic groups. Hence, the determination of CEH and the ethnic admixture print allows knowing immunogenetic variants and blocks of anthropological origin and evolution but biomedical importance in Mexican mestizo patients.

Thus, this study aimed to describe SLE patients' ethnic admixture proportions compared with healthy Mexican Mestizo individuals. We searched the distribution of HLA class I and class II blocks and CEH and their most likely ancestral origins using high-resolution HLA typing in a Mexican SLE group of admixed-ancestry.

SUBJECTS, MATERIALS, AND METHODS

Subjects

We recruited 146 consecutive SLE patients between 2015 and 2018 from the Rheumatology Outpatient Clinic at the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán (INCMNSZ) in Mexico City. As a control group, 234 unrelated healthy Mexican admixed individuals were included. Eligible SLE patients included those born in Mexico, whose parents and grandparents were also born in Mexico. Specifically, those who were born in Mexico City and its borders. The same criterion was applied for controls.

The general health status was evaluated or investigated for both SLE and controls. The SLE patients' diagnosis and classification were based on clinical manifestations, laboratory tests, and the American College of Rheumatology (ACR) criteria (Hochberg, 1997). The general health status in SLE patients was assessed with the SLE Activity Index (SLEDAI) (Bombardier et al., 1992; Guzmán et al., 1992). The patients were classified with punctuations between 0 and 4 it means inactive disease state or mild activity. The SLE classification punctuation was a criterion to establish the activity of the disease at the time of the interview.

Additionally, Systemic Lupus International Collaborating Clinics/American College of Rheumatology (SLICC/ACR) Damage Index (SDI) (Gladman et al., 2000) was registered for patients. This index was considered part of the health status evaluations, and it was thought as applicable to associate chronic evolution to genetic predisposition. The patients' quality of life was also evaluated with the questionnaire Lupus Quality of Life (Lupus Qol), ranging in punctuations from 0 to 100

(Devilliers et al., 2012). Additionally, it was verified that all patients were consecutive in their medical controls and complete compliance with pharmacological treatment. All the above, to rule out that a severe phenotype, evolution at the study moment, or chronic damage was due to lack of accessibility to medical service, lack of patient adherence, additional stresses concerning the quality of life, and more likely associated with an immunogenetic predisposition.

In controls, the general state of health was recorded as self-perceived health and ruling out autoimmune, metabolic, cancer, or any repetitive disease or chronically treated. To establish the similarity in the exposure to possible triggers of the disease, both control individuals and patients belong to the same geographical area, which is expressed as latitude and longitude in the sociodemographic information **Table 1**.

Finally, the socioeconomic status was verified. The social worker's department verified the patients' socioeconomic strata using a validated instrument applied to all the institutions belonging to the Coordinating Commission of National Institutes and High Specialty Regional Hospitals (CCIHSHAE). This instrument includes variables with a numerical value to assign the socioeconomic level: monthly family income, occupation of the primary economic provider, monthly family expenses, housing, and family health status. The sum of variables gives an approximation of the socioeconomic level.

Human Leukocyte Antigen Typing

Sanger Sequencing-Based Typing

Genomic DNA was obtained from whole blood using the QIAamp DNA mini kit (Qiagen, Valencia, CA, United States). DNA quality was assessed using a NanoDrop 2000 (Thermo Fisher Scientific, MA, United States) and Qubit Fluorometric Quantification (Invitrogen). The DNA integrity was evaluated by gel electrophoresis. Samples were stored at −20°C until analysis. The HLA typing was performed using a sequence-based method (SBM) described previously (Zúñiga et al., 2013).

Briefly, HLA class I typing was done by generic amplification of exons 2, 3, and 4 of each gene. For HLA class II, exon 2 and 3 of the HLA-DRB1 and -DQB1 genes were amplified using allele group-specific primer pairs. Polymerase chain reactions (PCRs) utilized 1.5 mm KCl, 1.5 mM MgCl2, 10 mM Tris-HCl (pH 8.3), 200 mM dNTPs, 10 pM of each primer, 30 ng of DNA, and 0.5 U of Taq DNA polymerase in a final volume of 25 μl. Amplifications were performed on a PE9700 thermal cycler (Applied Biosystems, Foster City, CA, United States) under the following cycling conditions: 95°C for 30 s, 65°C for 30 s, 72°C for 1 min, preceded by 5 min at 95°C and followed by a final elongation step at 72°C for 5 min. The amplified products were sequenced independently in both directions using BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems^{TM}) on the ABI PRISM® 3500 Genetic Analyzer (Applied Biosystems®). Sequencing products were purified with the BigDye XTerminator® Purification Kit (Applied Biosystems) to remove unincorporated BigDye^{TM} terminators and salts.

We analyzed data with matching allele assignment software (Applied Biosystems) using the IMGT/HLA sequence database alignment tool http://www.ebi.ac.uk/imgt/hla/align.html (Robinson, 2001). We solved ambiguities using group-specific sequencing primers (GSSPs) that had been previously validated (Lebedeva et al., 2011).

High-Resolution Typing by Next-Generation Sequencing

Next-generation sequencing Illumina® TruSight® HLA v2 Sequencing Panel (Illumina, San Diego, CA, United States) was also used to confirm HLA allele-level typing. We performed the process as the manufacturer recommends. Briefly, genomic DNA samples were adjusted to a working concentration of 10 ng/μL using Qubit equipment (Thermo Fisher Scientific, Waltham, MA, United States).

TABLE 1 | Clinical and sociodemographic information from SLE patients and controls.

Variable	SLE	Controls
Diagnosis Age, years, Mean ± SD	24. 5 ± 11.2	–
Age, years, Mean ± SD	39.7 ± 14.5	38 ± 15.3
Female	104 (89%)	120 (51%)
Male	13 (11%)	114 (49%)
Family	Grandparents and parents born at the same location	Grandparents and parents born at the same location
Latitude (Living locality)	19° 25′ N	19° 26′ N
Longitude (Living locality)	99° 7′ W	99° 8′ W
Urban/rural (Living locality)	Urban	Urban
Socioeconomic status	Low to middle	Low to middle
Health status		
Chronic diseases (Diabetes, hypertension)	6%	0%
SLEDAI % of individuals (score)	64% Inactive disease (0–2) 36% mild activity (3–4)	NA
SDI % of individuals (score)	84.6% (0–1) 13.67% (2–4) 1.73% (5–8)	NA
LupusQol % of individuals (score)	68% (>60)	NA
Lupus phenotype and clinical manifestations		
Articular	79.8%	
Serositis	29.7%	
Renal	65.6%	
Neurologic	16.4%	
Hematologic	64.8%	
Antiphospholipid syndrome	30.4%	
Other autoimmune diseases	34.5%	
ANAs	86%	
Relatives' information		
Relative with SLE	18.6%	
Relative with autoimmune disease	29.4%	

NA, not applicable; SLEDAI, SLE disease activity index; SDI, Systemic Lupus International Collaborating Clinics/American College of Rheumatology (SLICC/ACR) Damage Index; ANAs, antinuclear antibodies.

Generation of Long-range PCR templates. HLA-A,-B,-C,-DRB1, and -DQB1 loci were prepared using specific primers included in the TruSight HLA Pre 24 sample kit (Illumina) and MasterAmpTM Extra-Long DNA Polymerase (Lucien Corporation, Middleton, WI, United States).

Polymerase chain reactions were performed in a 96-well plate on the 9700 PE thermal cycler (Applied Biosystems/Thermo Fisher Scientific) using the following reagents proportions: 25 μl of HPM (HLA-PCR Mix), 2 μl of MasterAmpTM Extra-Long DNA Polymerase, 13 μl of water, and 5 μl of gDNA (10 ng/μl). Two PCR programs were performed for the fragment amplification of HLA loci. The first one for amplification of HLA-A, -B, -C, and -DRB1 loci, under the following conditions: initial denaturation at 94°C for 3 min, 30 cycles at 94°C for 30 s, 60°C for 2 min, 68°C for 15 min, 68°C for 10 min, and a final hold at 10°C.

The second PCR program for locus HLA-DQB1 was performed under the following conditions: 94°C for 3 min; followed by 10 cycles at 94°C for 30 s, 55°C for min, 72°C for 15 min; 20 cycles of 94°C for 30 s, 60°C for 2 min, 72°C for 15 min, 72°C for 10 min; and a final hold at 10°C. PCR products were confirmed by 1% agarose gel electrophoresis. The PCR clean-up was performed too.

Normalization and tagmentation. All loci PCR products' concentrations were normalized using magnetic beads (LNA1, LNB1, TruSight HLA, Illumina). This process is accomplished for multiplex library preparation and sequencing. After normalization, 40 μl of each PCR product was used for fragmentation (800 and 1200 pb), and fragmentation buffers HTM and HTB (TruSight HLA Pre-PCR 24, Illumina) were added to the reaction (10 μl each) and incubated at 58°C for 12 min in the presence of sequencing primers. The purified fragmented PCR products were pooled, and adaptor addition was performed using the Nextera XT DNA sample preparation kit (Illumina). Amplification was performed based on the following PCR program: denaturation at 72°C for 3 min and 98°C for 30 s, followed by 10 cycles at 98°C for 10 s, 60°C for 30 s, 72°C for 5 min, and a final hold at 10°C. Appropriate clean-up was performed.

Sequencing. Seven microliters of the PCR sequencing products were denatured with 10 μl of 0.1N NaOH and sequenced on a MiSeq instrument using the paired-end 300 cycle (2 × 150 bp paired-end) MiSeq Reagent Kit (Illumina) following the manufacturer instructions.

Next-Generation Sequencing Data Analysis

After the sequencing, MiSeq reporter analysis software-generated FASTQ sequence files and BAM alignment files. Next, we generated allele calls using the Assign 2.0^{TM} TruSight HLA Analysis software. The software used reference sequences from the IMGT/ HLA database (release 3.23.0.0).

Statistical Analysis

Clinical and Sociodemographic Characteristics

We analyzed clinical and demographic variables with the IBM SPSS Statistics 26 program.

Human Leukocyte Antigen Class I and Class II Alleles and CEH Frequencies

Differences in HLA class I and II alleles and haplotypes frequencies between patients and controls were analyzed using X2, and *p*-Values less than 0.05 were considered statistically different. *P*-values were also corrected using the Bonferroni method (for allele frequencies, multiplying the original *p*-Value by the number of alleles). Odds ratios (OR) and 95% confidence intervals (95%CI) were calculated to measure association strength with the program Epi Info^{TM} 7.2 version (Centers for Disease Control and Prevention, 2011). We generated a Forest plot of HLA alleles and haplotypes, which showed statistical significance using R programming version 4.0.3 (R Core Team, 2020).

Hardy-Weinberg equilibrium (HWE) at a locus-by-locus level was calculated. HLA alleles data was faced with the hypothesis that the observed diploid genotypes are the product of a random union of gametes. To detect significant departure from HWE was followed an analogous Fisher's exact test on a two-by-two contingency table but extended to a triangular contingency table of arbitrary size. The test was done using a modified version of the Markov-chain with the populations' genetic computer program Arlequin version 3.5.2.2 (Guo and Thompson, 1992; Excoffier and Lischer, 2010).

Furthermore, we calculated the diversity values: observed heterozygosity (OH), expected heterozygosity (EH), and the polymorphic information content (PIC) for each locus (HLA-A, -C, -B, -DRB1, and -DQB1). CEH of known Mexican Native American, European, African, and Asian origin were assigned based on previously reported frequencies investigated as Most Probably Ancestry (MPA) in previous studies. MPA is based on the frequency occurrence of haplotypes (Cao et al., 2001; Yunis et al., 2003, 2005; Zúñiga et al., 2013). Individual alleles and haplotypes frequencies and locality of occurrence searching HLA tool can be found in Allele Frequency Net Database, the gold-standard data classification (Gonzalez-Galarza et al., 2020).

Linkage disequilibrium (LD) between HLA loci pairs was calculated based on delta (Δ), a LD coefficient, which measures the deviation from a random association between alleles of different loci. The results have been reported as relative delta (Δ′) values. Δ′ oscillates among values from -1 to 1. 1 represents the highest probability that a pair of alleles or DNA segments segregates as a block. In contrast, -1 represents the probability of a total random pairing (Excoffier and Lischer, 2010).

Admixture Estimation

The maximum likelihood method was used to estimate SLE patients' and healthy individuals' admixture proportions using the population genetics LEADMIX software (Wang, 2003). Four major parental populations (k = 4) were evaluated, including a population per continental location according to the settlement history in Mexico. According to the availability of HLA data, the populations included were Mexican Native American, Southwestern European, Sub Saharan African, and Eastern Asian. HLA-B and -DRB1 were used as genetic estimators in the included populations. We estimated Mexican Native

American contribution from Oaxaca Mixtecs data, a populace from southeastern Mexico (Arnaiz-Villena et al., 2014), and Chihuahua Tarahumaras, a northern Mexico population (García-Ortiz et al., 2006). Similarly, we estimated non-autochthonous parental populations' contribution with data obtained from Southwestern European components from a representative sample from Spain (Catalunya, Navarra, Extremadura, Aragon, and Cantabria) (Enrich et al., 2019), Sub-Saharan African components from Zimbabwe Harare Shona Inhabitants (Louie, 2006b), and Eastern Asian components from the China Han populace (Trachtenberg et al., 2007).

Genetic Distance and Principal Component Analysis

Genetic distances were assessed for the parental populations evaluated in the admixture estimation ($k = 4$) (Mexican Native American, Southwestern European, Sub-Saharan African, and Eastern Asian). This analysis was based on HLA-B. Besides, we included other Mexican Native American groups to deepen the comparison; the groups added were: Lacandon Mayans (Barquera et al., 2020f), Oaxaca Mixe (Hollenbach et al., 2001; Single et al., 2020), Oaxaca Zapotecans (Hollenbach et al., 2001), and Seri (Gorodezky, 2006). The genetic distance analysis included Nei's distance performed with the Arlequin genetics population program. We generated graphics with an R programming extension added to Arlequin.

Principal Components Analysis (PCA) for fifty-six populations with HLA-B data available was performed using the BioVinci Software 2.0 to analyze the distribution of HLA-B alleles. There were included populations from three continental locations European, African, and Latin-American. The study groups SLE and controls HLA-B were adjusted to low-resolution. PCA included population data of African populations (7 populations): Burkina Faso Mossi (Modiano et al., 2001), Cameroon Yaounde (Pimtanothai et al., 2001), Ghana Ga-Adangbe (Norman et al., 2013), Kenya (Mack et al., 2006), Kenya Nandi (Cao et al., 2004), Uganda Kampala (Kijak et al., 2009), and Zimbabwe Harare Shona (Louie, 2006a). European populations (8 populations): England*, France Lyon*, Germany Essen*, Ireland Northern (Williams et al., 1999; Middleton et al., 2000), Italy*, Italy Sardinia (Grimaldi et al., 2001), Netherlands* and Spain (Enrich et al., 2019). Mexican Mestizo populations from Northern Mexico (7 populations): Baja California (Escobedo-Ruíz et al., 2020), Chihuahua (Pacheco-Ubaldo et al., 2020), Colima (Barquera et al., 2020b), Durango (González-Medina et al., 2020), Nuevo León (Barquera et al., 2020a), Sinaloa (Clayton et al., 2020), and Sonora (Uribe-Duarte et al., 2020). Mexican Mestizo populations from the Center of Mexico (14 populations): Aguascalientes (Bravo-Acevedo et al., 2020), Guanajuato (Pantoja-Torres et al., 2020), Guerrero (Juárez-Nicolás et al., 2020), Jalisco (Bravo-Acevedo et al., 2020), Mexico City Center, Mexico City Western, Mexico City Eastern, Mexico City Southern, Mexico City Northern (Barquera et al., 2020e), Michoacán (Ballesteros-Romero et al., 2020), Morelos (Ortega-Yáñez et al., 2020), Nayarit (Goné-Vázquez et al., 2020), Querétaro (Martínez-Álvarez et al., 2020), and San Luis Potosí (Hernández-Zaragoza et al., 2020). Mexican Mestizo populations from Southern Mexico (4 populations): Chiapas (Barquera et al., 2020d), Oaxaca (Hernández-Hernández et al., 2020), Quintana Roo (Medina-Escobedo et al., 2020) and Tabasco (Solís-Martínez et al., 2020). Native American Mexican populations (8 populations): Amerindian pooled population 1 (Tarahumaras, Mixtecs, and Zapotecan) (Hollenbach et al., 2001; García-Ortiz et al., 2006) and Amerindian pooled population 2 (Mixtecs and Tarahumaras), Mixe (Hollenbach et al., 2001), Mixtecs (Arnaiz-Villena et al., 2014), Nahuas (Vargas-Alarcón et al., 2003), Lacandon (Barquera et al., 2020f), Seri (Gorodezky, 2006), and Zapotecan (Hollenbach et al., 2001). Non-Mexican Latin-American populations (8 populations): Argentina Amerindian (Cerna et al., 1993), Bolivia Amerindian Aymara*, Colombia*, Costa Rica African-Caribbean (Arrieta-Bolaños et al., 2018), Costa Rica Amerindians (Arrieta-Bolaños et al., 2018), Ecuador*, Nicaragua Managua*, and Panama*.

*Populations from bone marrow or transplantation registry or the dataset is not associated with a publication or has not been yet published. The accessions are detailed in Data Availability Statement.

To establish the populations' similarities and differences in the PCA plot, we generate a discrimination tree that shows high variance features (HLA alleles which frequency, distribution, and representativeness in populations are established as a criterion of discrimination, separation, or conjunction between the evaluated populations). The alleles which differentiate the populations better are showed in the discrimination tree. This is evaluated with the software algorithm base on the frequencies of each allele.

RESULTS

Clinical and Sociodemographic Characteristics

The clinical and sociodemographic characteristics of the Mestizo Mexican SLE patients are summarized in **Table 1**. The SLE patients were 89% female, and they had a mean age of 39.7 ± 14.5 years. While controls were 51% female, and they had a mean age of 38.0 ± 15.0 years. Both groups were considered Mexican Mestizos. The latitude and longitude where the people live were very similar for both groups, took from The Allele Frequency Net Database (AFND). This data was important to validate that both groups are possible exposed to the same disease triggers in the same area. Both SLE and controls were cataloged as urban, which reports a similar utility. The lifestyle and habits could be another factor to restrict the exposition to the same triggers.

The socioeconomic status was defined with an instrument that allowed classifying SLE and Controls as low and middle socioeconomic strata individuals in similar proportions.

The general healthy state in the patients was corroborated by ACR classification, SLEDAI, and SDI questionaries with the clinical and history evaluation of the Rheumatologist. Different SLE phenotypes were found and are detailed in **Table 1**. Most of the patients showed over one clinical phenotype. The most common were arthritis (79.8%), renal (65.6%), and hematological (64.8%). The mean age of disease onset was 24.5 ± 11.2 years, and 34.5% of the patients presented a concomitant autoimmune disease. The most frequent were Graves' disease and hypothyroidism. Other conditions less frequent were Sjogren's syndrome, autoimmune acquired hemophilia, neuromyelitis optica, vitiligo, and scleroderma.

The healthy state of controls was verified by questionnaire. All participants expressed good self-perceived health: no diseases or chronic treatments and no impediment to daily activities.

Genetic Diversity and Admixture

As expected, the HLA-B and HLA-DRB1 loci were the most polymorphic in both groups, whereas HLA-DQB1 was the least polymorphic locus. Diversity parameters, polymorphic information content, expected and observed heterozygosity values are shown in **Table 2**. Hardy-Weinberg Equilibrium analysis is displayed with the p corrected for each locus. HLA-DRB1 locus tends to have a marginal deviation from Hardy-Weinberg equilibrium after Bonferroni correction, which is being studied more deeply.

Systemic lupus erythematosus patients have a higher non-autochthonous HLA gene load, while controls have a higher Mexican Native American HLA gene load. We performed the admixture estimations with the Maximum Likelihood method based on HLA-B and -DRB1 of four digits for each included population. HLA-B is the most polymorphic locus and, most of the time, is selected for admixture estimations. Besides, HLA-DRB1 is the locus most associated with autoimmunity. Both loci analyses conduct to the same conclusion: SLE patients have a higher proportion of Southwestern European ancestry, 48 ± 8% (Mean proportion between HLA-B and HLA-DRB1 ± Standard deviation) than healthy individuals, 30 ± 7%. Instead, SLE patients have a minor proportion of Mexican Native American ancestry, 44 ± 1% than healthy individuals 63 ± 4% and the Sub-Saharan African component appears to have a more similar distribution between SLE and controls 7 ± 7% and 6 ± 7%, respectively. The Asian admixture proportions are not showed; it was minimal. Detailed data for HLA-B and HLA-DRB1 are shown in **Figure 1**.

Genetic Distances

As expected, because the admixed populations (SLE and healthy individuals) belong to the same region, the variations in genetic distances are low. However, the variations are consistent with the admixture analysis and showed statistical significance ($pC < 0.05$). Nei's distances (d) have shown that SLE patients are closer to non-autochthonous populations (Southwestern European, Sub-Saharan African, and Eastern Asian) than healthy individuals, while healthy individuals are closer to autochthonous populations than the SLE group (**Figure 2**). Also, calculating the corrected average pairwise distance between populations gave the same information obtained through Nei's distance. Additionally, the average number of pairwise differences within populations was considered to validate each populations' genetic structure. As it is shown, the most isolated Mexican Native American groups (Seri, Lacandon, and Mixe, which are Mexican Native American groups that are geographically and culturally independent from other Mexican populations, implies a reduced or null gene flow from and to these populations, and consequently reduced diversity) showed fewer intra-group differences. These results are shown in a color matrix. The values corresponding to admixed groups (SLE and healthy) have been included in the matrix. The entire matrix values are shown in **Supplementary Information 1** as population average pairwise differences (π). The same conclusions were reached with other genetic structure analysis as FST value, coancestry coefficients, and Slatkin linearized FSTs (**Supplementary Figures 1–3**).

Further evidence of the differentiated distribution of immunogenetic diversity in Mexican mestizo groups (SLE and healthy individuals) can be seen in the principal component analysis (PCA) graph. Populations from three continental locations with HLA-B data were considered to construct the PCA. The objective was to visualize the differences between LES and Controls better. We opted to generate a visual discrimination tree, which calls the strongest elements that differentiate the included populations. The PCA point distribution is consistent with ethnic differences among the included populations. The separation of SLE and controls is consistent with the analysis of genetic distances. But it is observed that both the SLE groups and the controls are in the area of the plot where other groups of Mexican mestizos of Mexico City are found (MxCC, MxCW, MxCE, MxCS, MxCN, and MxCS; Centro Ciudad de México, West, East, South, and North, respectively). The distribution of the groups of Mexican mestizos in the PCA shows previous knowledge about Mexican mestizos. The northern states have higher European ancestry, while the southeast has more similarities among Mexican Native Americans (**Figure 3**).

Additionally, it can be observed how the Native American groups of Mexico present important differences with the mestizo populations. The genetics of Mexico recapitulates the substructure of the Native Americans and affects the biomedical traits. Therefore, although they seem distant and little related, these autochthonous populations confer complexity to the Mexican admixture. Thus, the immunogenetic diversity of the HLA system in Mexico correlates with the genetic structure of the underlying population (**Figure 3**).

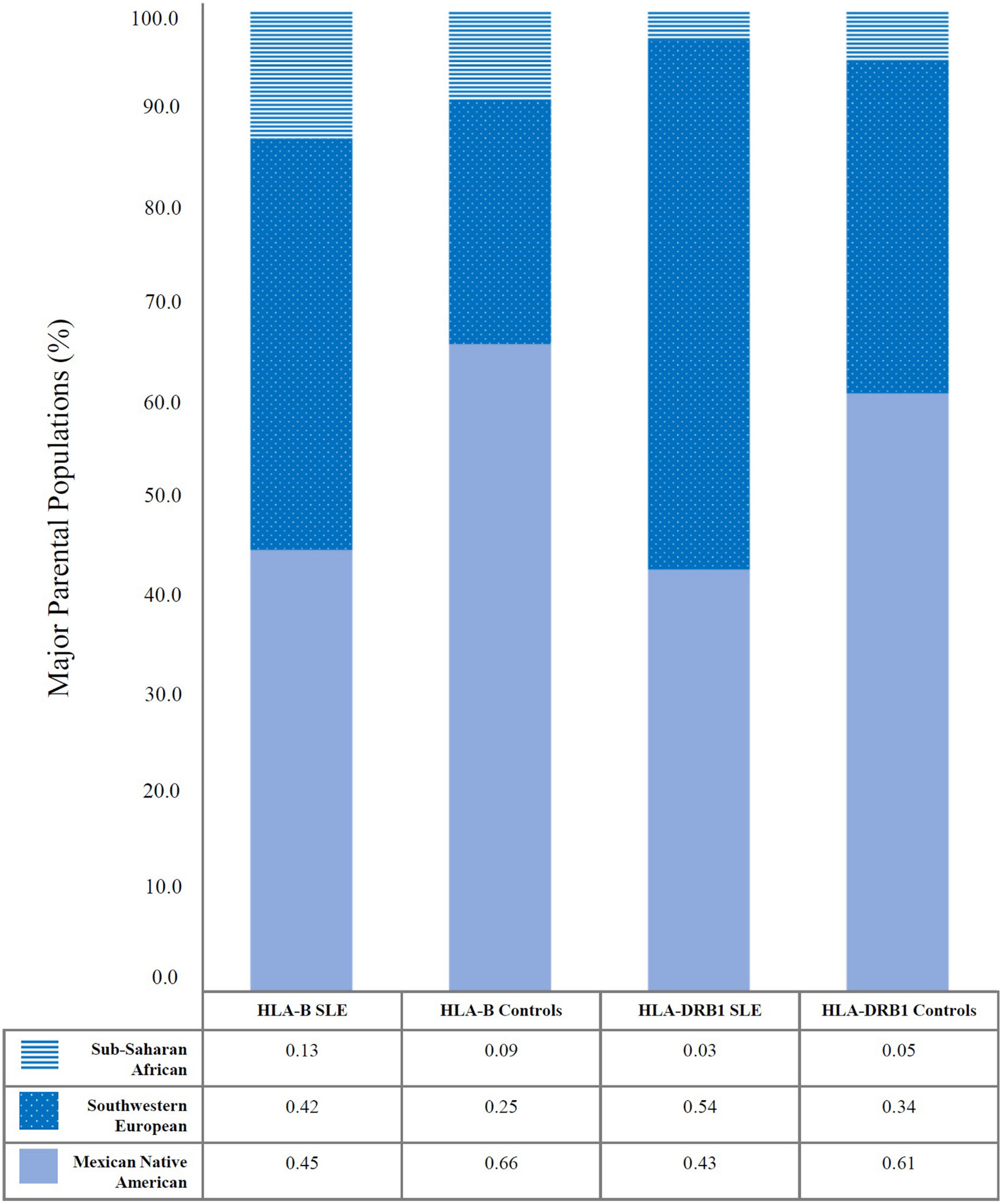
HLA-A*29:02∼C*16:01∼B*44*03∼DRB1*07:01∼DQB1*02:02 as a novel CEH of susceptibility identified in Mexican SLE patients.

The determination of allele and haplotype frequencies and the risk or protection conferred by the HLA alleles and haplotypes

TABLE 2 | Estimations of genetic diversity of HLA class I and class II loci in SLE patients and healthy Individuals.

HLA-locus	SLE				Healthy individuals			
	EH	OH	*p*C	PIC	EH	OH	*p*C	PIC
A	0.9060	0.9697	0.201	0.9002	0.8919	0.8718	0.531	0.9761
C	0.8922	0.9571	0.352	0.8879	0.8947	0.9009	0.196	0.9907
B	0.9654	0.9929	0.833	0.9608	0.9668	0.9487	0.216	0.9767
DRB1	0.9249	0.9489	0.080	0.9190	0.9193	0.9013	0.010	0.8350
DQB1	0.8610	0.9203	0.696	0.8600	0.8256	0.8205	0.269	0.9447

***pC**, p corrected value after Bonferroni correction; **OH**, observed heterozygosity; **EH**, expected heterozygosity; **PIC**, polymorphic information content.*



	HLA-B SLE	HLA-B Controls	HLA-DRB1 SLE	HLA-DRB1 Controls
▤ Sub-Saharan African	0.13	0.09	0.03	0.05
▨ Southwestern European	0.42	0.25	0.54	0.34
■ Mexican Native American	0.45	0.66	0.43	0.61

FIGURE 1 | Ethnic admixture estimations revealed differences in Mexican Native American and Southwestern European components between Mexican SLE patients and healthy individuals. It was analyzed by Maximum likelihood approximation using HLA-B and HLA-DRB1. The values presented in Y-axis are the percentages of the main parental populations included (Mexican Native American, Southwestern European, and Sub-Saharan African). The table shows the relative frequencies from 0 to 1. Sub-Saharan African component is shown in blue horizontal lines, Southwestern European component is shown in blue with white points, and Mexican Native American component is shown in solid light blue. Eastern Asian ancestry was minimal, and it is not represented in the graphic.

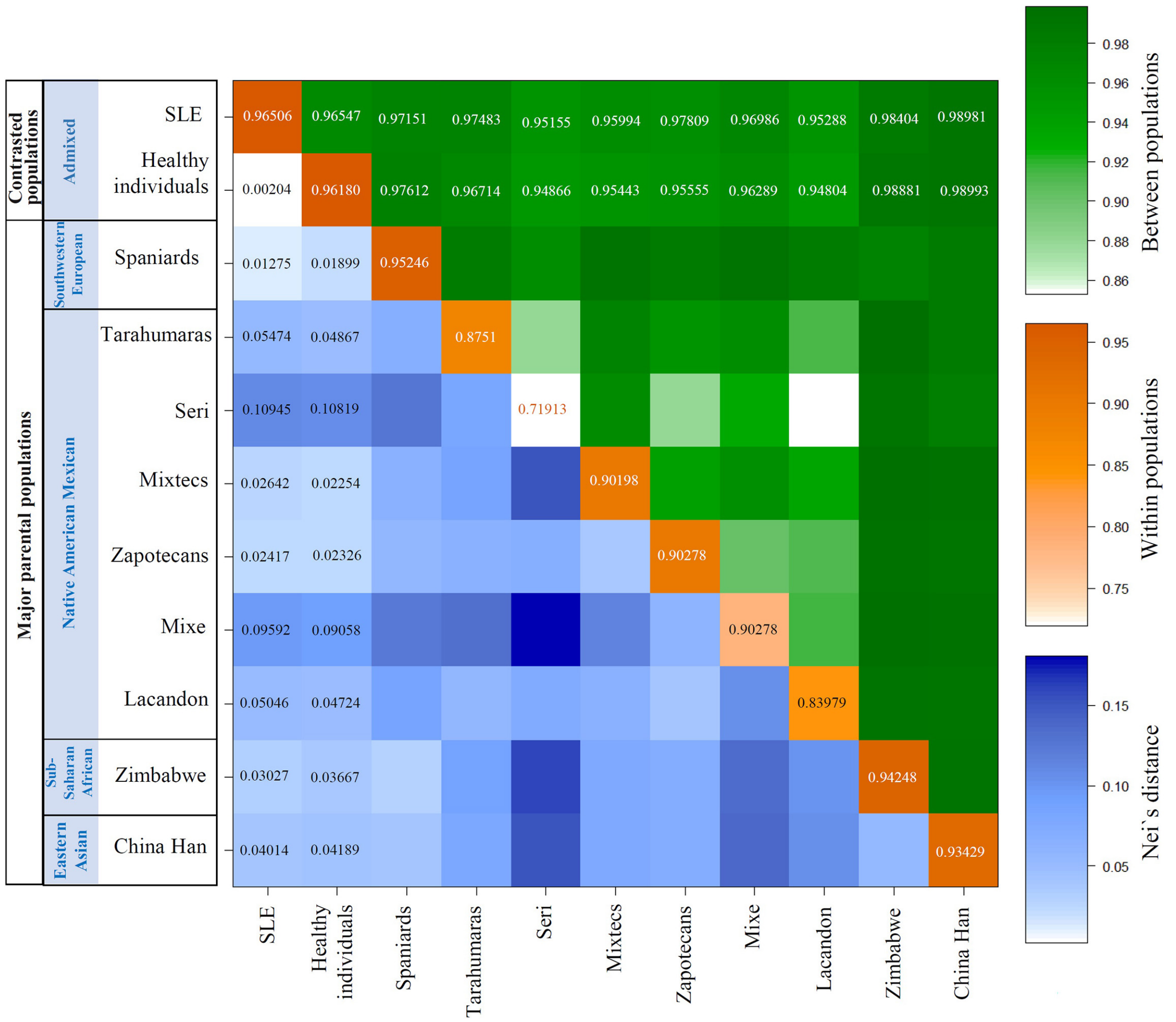


FIGURE 2 | Genetic distances revealed differences in Mexican Native American and Southwestern European ancestry between the SLE and healthy individuals. In this matrix, we represent three different color scales, the average number of pairwise differences (π) between populations. Orange on diagonal: π within populations; Green above diagonal: πxy between pairs of populations Blue below diagonal: net number of nucleotide differences between populations (Nei's distance). The interpretation is the same for the three scales; it is shown in the right color bars. To > value > genetic distance. To > intensity color > genetic distance.

are shown in **Tables 3–7** and **Supplementary Tables 1–4**. We found the frequency of a novel susceptibility haplotype incremented in SLE patients. The European block HLA-A*29:02~C*16:01~B*44*03~DRB1*07:01~DQB1*02:02 with *p*C = 0.02, OR = 6.7 (**Table 7**). None of the individual alleles that make up this CEH showed statistical significance in this study.

The CEH HLA-A*01:01~C*07:01~B*08:01~DRB1*03:01~DQB1*02:01 is the highest risk factor to develop SLE in Mexicans mestizo patients from Mexico City. As shown in previous studies in low-resolution HLA typing (Granados et al., 1996; Vargas-Alarcón et al., 2001; Graham et al., 2007) and the current one in high-resolution HLA typing, the European conserved extended haplotype HLA-A*01:01~C*07:01~B*08:01~DRB1*03:01~DQB1*02:01 (*p*C = 0.0004, OR = 18.7) has been found as the high-risk factor to develop SLE in Mexicans from Mexico City (**Table 7**). The individual loci statistical analysis has identified relative risks given by each of the alleles that compound this CEH (**Figure 4**). The relative risk observed is highly influenced by linkage disequilibrium with neighboring risk genes, both HLA and non-HLA (**Supplementary Figure 4**). Additionally, the allele HLA-A*11:01 was found as a risk factor with no haplotype linkage, with *p*C = 0.035, OR = 2.5 (**Figure 4** and **Supplementary Table 1**). This allele has previously identified in Malays and Chinese SLE patients (Mohd-Yusuf et al., 2011).

HLA-DRB1*14:06~DQB1*03:01 and -DRB1*16:02~DQB1*03:01 the protective Mexican Native American haplotypes which frequency is diminished in SLE patients from Mexico City, as it is shown in previous studies (Salgado-Galicia et al., 2020) and the current one. The HLA-DRB1*14 has been previously found in Asians as -DRB1*14:03 (Furukawa et al., 2014). The frequency of HLA class II protective alleles HLA-DRB1*14:06

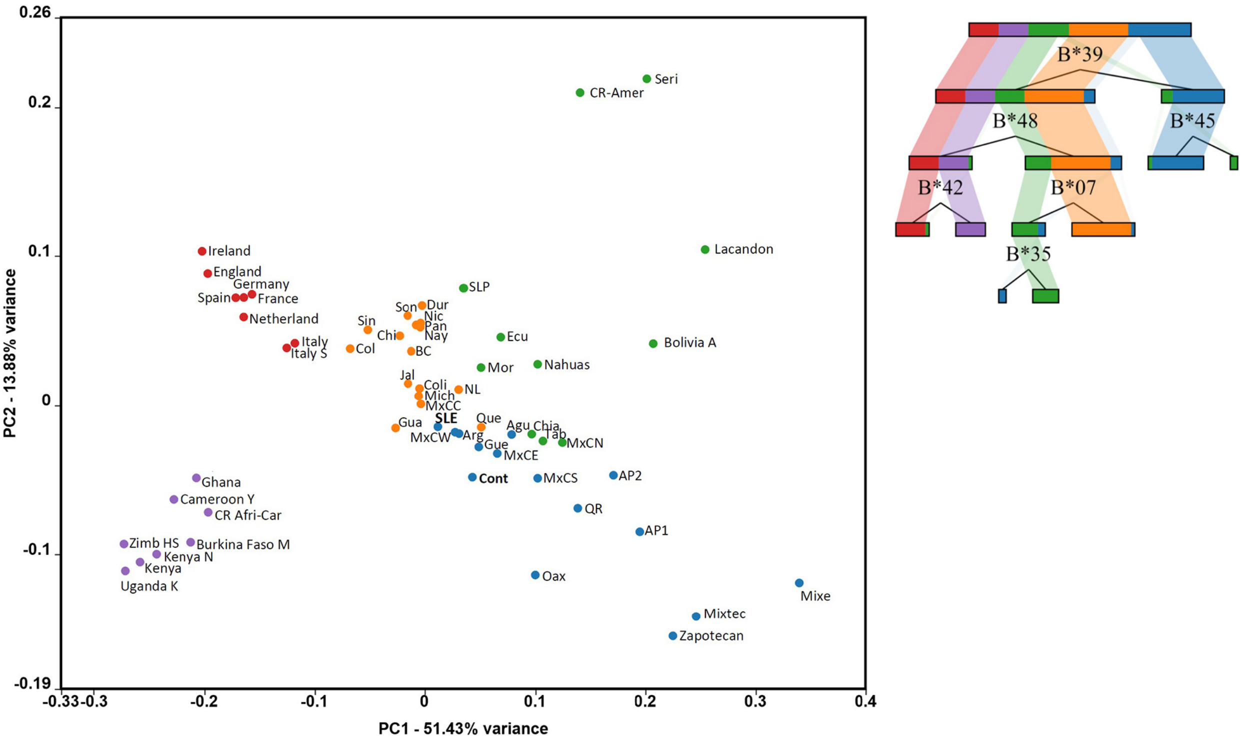


FIGURE 3 | The principal component analysis (PCA) plot shows the SLE Mexico City particularities. Populations are colored by high variance features represented by the alleles in the decision tree shown on the right. Some populations match in color and continental location with the HLA allele frequency discrimination. Purple dots represent African, and red represent European. Orange, blue, and green dots represent Mestizo and Native populations from Mexico states and Central and South America, properly described in the "Materials and Methods" section. Burkina Faso M: Burkina Faso Mossi, Uganda K: Uganda Kampala, Cameroon Y: Cameroon Yaounde, Kenya N: Kenya Nandi, Ghana: Ghana Ga-Adangbe, Italy S: Sardinia, BC: Baja California, Coli: Colima, Sin: Sinaloa, Dur: Durango, Son: Sonora, and NL: Nuevo León, Nic: Nicaragua, Pan: Panama, Col: Colombia, Que: Queretaro, Gua: Guanajuato, Mich: Michoacán, MxCC: Mexico City Center. Agu: Aguascalientes, Gua: Guanajuato, Que: Queretaro, Gue: Guerrero, Mich: Michoacán, SLP: San Luis Potosí, Nay: Nayarit, Jal: Jalisco, MxCC: Mexico City Center, MxCW: Mexico City Western, MxCE: Mexico City Eastern, MxCS: Mexico City Southern, MxCN: Mexico City Northern. Chi: Chiapas, Tab: Tabasco, Oax: Oaxaca, QR: Quintana Roo. AP1: Amerindian pooled 1, AP2: Amerindian pooled 2, CR-Amer: Costa Rica Amerindian, CR Afri-Car: Costa Rica African-Caribbean. Ecu: Ecuador, Mor: Morelos, SLP: San Luis Potosí, Bolivia A: Bolivia Aymaras.

(*p*C = 0.006, OR = 0.4) and HLA-DRB1*16:02 (*p*C = 0.006, OR = 0.3) were found diminished in the SLE group (**Table 4**). The Mexican Native American haplotypes, showed protection, the HLA-DRB1*14:06~DQB1*03:01 (*p*C = 0.007, OR = 0.4) and HLA-DRB1*16:02~DQB1*03:01 (*p*C = 0.006, OR = 0.3) (**Tables 5**, **6**).

DISCUSSION

The HLA sequencing has led both to identify a new susceptibility block HLA-A*29:02~C*16:01~B*44*03~DRB1*07:01~DQB1*02:02 and to determine the admixture print which distinguishes patients and healthy individuals in Mexico City. Therefore, alleles and haplotypes of susceptibility and protection found, both the previously described and the new one, will be analyzed, emphasizing discussing the influence of ethnic admixture and the genetic load of parental populations in the development of lupus.

The CEH associated for the first time with SLE development in Mexicans was the HLA-A*29:02~C*16:01~B*44*03~DRB1*07:01~DQB1*02:02. This CEH has been cataloged in previous studies as European (Szilágyi et al., 2010). Only the isolated allele HLA-DRB1*07 has been previously associated with antiphospholipid syndrome in SLE patients from Mexico City (Granados et al., 1997). However, this CEH has been a risk factor in Mexican patients diagnosed with Achalasia (Furuzawa-Carballeda et al., 2018). This condition is a motility disorder of the esophagus with abnormalities in the neurons that controls peristaltic movements, whose underlying cause is unknown. Notably, previous studies conducted in other populations suggested that achalasia patients have increased frequency of HLA-DRB1*15 and -DQB1(eight-amino-acid insertion in the cytoplasmic tail of HLA-DQβ1) alleles in an ethnicity-specific manner. This fact proposes an immunogenetic mechanism (Verne et al., 1999; Gockel et al., 2014; Becker et al., 2016). However, in this study, no SLE patient who carries this CEH showed some achalasia symptomatology. Probably, the immunosuppressive treatment might mask achalasia symptoms or prevents the development in predisposed individuals. A congruent reason could be that the mean age of achalasia onset in Mexicans (42.3 ± 15.8 years) is slightly superior to SLE (26.5 ± 12.2 years). The data needs further research because the shared susceptibility conferred by this CEH could mean a

TABLE 3 | Human leukocyte antigen-B allele frequencies in SLE patients and healthy individuals.

HLA-B alleles	**SLE**		**Healthy individuals**		***pC***	**OR**	**95%IC**	
	***N* = 143 (286 alleles)**		***N* = 234 (468 alleles)**					
	n	**AF**	***n***	**AF**				
B*08:01	**20**	**0.0699**	**3**	**0.0064**	**0.000003**	**11.7**	**3.43**	**39.59**
B*39:05	24	0.0839	37	0.0791	ns			
B*35:01	19	0.0664	27	0.0577	ns			
B*51:01	16	0.0559	28	0.0598	ns			
B*35:17	15	0.0524	18	0.0385	ns			
B*07:02	13	0.0455	19	0.0406	ns			
B*44:03	13	0.0455	13	0.0278	ns			
B*35:12	13	0.0455	18	0.0385	ns			
B*40:02	10	0.0350	25	0.0534	ns			
B*18:01	10	0.0350	8	0.0171	ns			
B*39:06	**8**	**0.0280**	**32**	**0.0684**	**0.03**	**0.4**	**0.18**	**0.86**
B*52:01	8	0.0280	10	0.0214	ns			
B*48:01	9	0.0315	20	0.0427	ns			
B*14:02	7	0.0245	15	0.0321	ns			
B*15:15	6	0.0210	15	0.0321	ns			
B*15:03	6	0.0210	2	0.0043	ns			
B*49:01	6	0.0210	9	0.0192	ns			
B*35:02	5	0.0175	2	0.0043	ns			
B*15:01	5	0.0175	10	0.0214	ns			
B*37:01	4	0.0140	4	0.0085	ns			
B*14:01	3	0.0105	4	0.0085	ns			
B*38:01	4	0.0140	6	0.0128	ns			
B*35:14	3	0.0105	7	0.0150	ns			
B*39:01	3	0.0105	5	0.0107	ns			
B*57:03	3	0.0105	1	0.0021	ns			
B*35:16	2	0.0070	3	0.0064	ns			
B*15:16	2	0.0070	1	0.0021	ns			
B*15:17	2	0.0070	3	0.0064	ns			
B*15:30	2	0.0070	8	0.0171	ns			
B*13:02	2	0.0070	6	0.0128	ns			
B*35:43	3	0.0105	9	0.0192	ns			
B*39:02	2	0.0070	10	0.0214	ns			
B*41:01	2	0.0070	5	0.0107	ns			
B*44:02	2	0.0070	5	0.0107	ns			
B*50:01	2	0.0070	4	0.0085	ns			
B*57:01	2	0.0070	7	0.0150	ns			
B*58:01	2	0.0070	3	0.0064	ns			
B*39:08	1	0.0035	3	0.0064	ns			
Other alleles	27							

AF, *allele frequency;* ***ns***, *not significant;* ***pC***, *p corrected value using Bonferroni method;* ***OR***, *odds ratio;* ***95%CI***, *95% confidence interval;* ***N***, *total number of individuals;* ***n***, *absolute frequency for each allele. Bold text highlights the results with statistical significance.*

TABLE 4 | Human leukocyte antigen-DRB1 allele frequencies in SLE patients and healthy individuals.

HLA-DRB1 alleles	**SLE**		**Healthy individuals**		***pC***	**OR**	**95%IC**	
	***N* = 143 (286 alleles)**		***N* = 234 (468 alleles)**					
	n	**AF**	***n***	**AF**				
DRB1*08:02	52	0.1818	91	0.1944	ns			
DRB1*04:07	27	0.0944	55	0.1175	ns			
DRB1*03:01	**29**	0.1014	**15**	0.0321	**0.0002**	**3.4**	**1.79**	**6.47**
DRB1*07:01	24	0.0839	33	0.0705	ns			
DRB1*15:01	20	0.0699	17	0.0363	ns			
DRB1*04:04	17	0.0594	31	0.0662	ns			
DRB1*11:04	10	0.0350	8	0.0171	ns			
DRB1*13:01	11	0.0385	12	0.0256	ns			
DRB1*14:06	**12**	0.0420	**47**	0.1004	**0.006**	**0.4**	**0.20**	**0.75**
DRB1*01:01	8	0.0280	9	0.0192	ns			
DRB1*04:11	7	0.0245	9	0.0192	ns			
DRB1*16:02	**5**	0.0175	**30**	0.0641	**0.006**	**0.3**	**0.10**	**0.68**
DRB1*04:01	3	0.0105	3	0.0064	ns			
DRB1*11:02	4	0.0140	4	0.0085	ns			
DRB1*13:03	5	0.0175	3	0.0064	ns			
DRB1*14:02	5	0.0175	11	0.0235	ns			
DRB1*04:05	3	0.0105	1	0.0021	ns			
DRB1*08:04	3	0.0105	2	0.0043	ns			
DRB1*15:03	4	0.0140	1	0.0021	ns			
DRB1*01:02	4	0.0140	11	0.0235	ns			
DRB1*01:03	3	0.0105	3	0.0064	ns			
DRB1*04:02	4	0.0140	10	0.0214	ns			
DRB1*08:01	2	0.0070	1	0.0021	ns			
DRB1*09:01	2	0.0070	1	0.0021	ns			
DRB1*12:01	1	0.0035	2	0.0043	ns			
DRB1*13:04	2	0.0070	1	0.0021	ns			
DRB1*04:03	1	0.0035	10	0.0214	ns			
DRB1*04:08	2	0.0070	1	0.0021	ns			
DRB1*04:10	1	0.0035	2	0.0043	ns			
DRB1*11:01	1	0.0035	6	0.0128	ns			
DRB1*13:02	3	0.0105	10	0.0214	ns			
DRB1*13:05	1	0.0035	1	0.0021	ns			
DRB1*16:01	1	0.0035	2	0.0043	ns			
DRB1*12:02	1	0.0035	3	0.0064	ns			
DRB1*14:01	1	0.0035	8	0.0171	ns			
Other alleles	7							

AF, *allele frequency;* ***ns***, *not significant;* ***pC***, *p corrected value using Bonferroni method;* ***OR***, *odds ratio;* ***95%CI***, *95% confidence interval****;*** ***N***, *total number of individuals;* ***n,*** *absolute frequency for each allele. Bold text highlights the results with statistical significance.*

share immunopathological mechanism depending on ethnic background. The onset of either lupus or achalasia could rely on the triggers, which might differ for each condition.

Achalasia onset has been associated with varicella herpes zoster virus previous infection (Becker et al., 2016). In contrast, SLE development has been associated with other viruses' infections as Epstein-Barr virus (EBV), parvovirus B19 (B19V), and human endogenous retroviruses (HERVs) (Quaglia et al., 2021). Interestingly, the novel haplotype associated with SLE in Mestizo Mexicans from Mexico City is in 2.8% of these individuals, while only in 0.43% of healthy Mexicans, 3.42%

TABLE 5 | Frequencies of HLA-DRB1~DQB1 haplotypes in SLE patients and healthy individuals.

HLA-DRB1~DQB1 Haplotypes	SLE *N* = 143 (286 alleles)			Healthy individuals (468 alleles)			*p*C	OR	95%IC	
	n	HF	Δ′	*n*	HF	Δ′				
African										
DRB1*13:01~DQB1*05:01	5	0.0175	0.410	1	0.0021	0.0159	ns			
DRB1*08:04~DQB1*03:01	1	0.0035	0.216	2	0.0043	1.0000	ns			
DRB1*12:01~DQB1*05:01	1	0.0035	1.000	1	0.0021	0.4632	ns			
Amerindian										
DRB1*08:02~DRB1*04:02	52	0.1818	1.000	89	0.1902	0.9723	ns			
DRB1*04:07~DQB1*03:02	27	0.0944	1.000	53	0.1132	0.9518	ns			
DRB1*14:06~DQB1*03:01	**12**	**0.0420**	**1.000**	**46**	**0.0983**	**0.9717**	**0.007**	**0.4**	**0.21**	**0.77**
DRB1*16:02~DQB1*03:01	**5**	**0.0175**	**1.000**	**30**	**0.0641**	**1.0000**	**0.006**	**0.3**	**0.10**	**0.68**
DRB1*14:02~DQB1*03:01	4	0.0140	0.765	11	0.0235	1.0000	ns			
DRB1*04:11~DQB1*04:02	1	0.0035	0.818	1	0.0021	–0.4595	ns			
Asian										
DRB1*11:02~DQB1*03:01	2	0.0070	0.412	3	0.0064	0.6674	ns			
DRB1*09:01~DQB1*03:03	1	0.0035	0.216	1	0.0021	1.0000	ns			
DRB1*13:02~DQB1*05:01	1	0.0035	0.279	1	0.0021	0.0338	ns			
Caucasian										
DRB1*03:01~DQB1*02:01	**29**	**0.1014**	**1.000**	**15**	**0.0321**	**1.0000**	**0.0002**	**3.4**	**1.79**	**6.47**
DRB1*15:01~DQB1*06:02	18	0.0629	0.891	15	0.0321	0.8779	ns			
DRB1*11:04~DQB1*03:01	8	0.0280	0.765	8	0.0171	0.0171	ns			
DRB1*13:01~DQB1*06:03	6	0.0210	0.740	6	0.0128	1.0000	ns			
DRB1*04:01~DQB1*03:02	2	0.0070	0.576	3	0.0064	1.0000	ns			
DRB1*07:01~DQB1*03:03	2	0.0070	0.635	5	0.0107	0.4620	ns			
DRB1*11:01~DQB1*03:01	1	0.0035	1.000	2	0.0043	0.1130	ns			
DRB1*04:02~DQB1*03:02	4	0.0140	1.000	10	0.0214	1.0000	ns			
Caucasian shared with other populations										
DRB1*07:01~DQB1*02:02	22	0.0769	0.908	28	0.0598	1.0000	ns			
DRB1*04:04~DQB1*03:02	17	0.0594	1.000	29	0.0620	0.9144	ns			
DRB1*01:02~DQB1*05:01	12	0.0420	1.000	11	0.0235	1.0000	ns			
DRB1*13:03~DQB1*03:01	3	0.0105	0.529	3	0.0064	1.0000	ns			
DRB1*01:03~DQB1*05:01	2	0.0070	0.640	3	0.0064	1.0000	ns			
DRB1*04:05~DQB1*03:02	2	0.0070	0.576	1	0.0021	1.0000	ns			
DRB1*08:01~DQB1*04:02	2	0.0070	1.000	1	0.0021	1.0000	ns			
DRB1*11:02~DQB1*03:19	2	0.0070	0.493	1	0.0021	0.2419	ns			
DRB1*04:03~DQB1*03:02	1	0.0035	1.000	10	0.0214	1.0000	ns			
DRB1*12:02~DQB1*03:01	1	0.0035	1.000	3	0.0064	1.0000	ns			
DRB1*13:02~DQB1*06:04	1	0.0035	1.000	9	0.0192	0.8978	ns			
DRB1*13:05~DQB1*03:01	1	0.0035	1.000	1	0.0021	1.0000	ns			
DRB1*14:01~DQB1*05:03	1	0.0035	1.000	8	0.0171	1.0000	ns			
DRB1*16:01~DQB1*05:02	1	0.0035	1.000	1	0.0021	0.4968	ns			

__HF__, haplotype frequency; ns, not significant; __pC__, p corrected value using Bonferroni method; __OR__, odds ratio; __95%CI__, 95% confidence interval. __N__, total number of individuals__,__ __n__, absolute frequency for each allele; Δ′, linkage disequilibrium value. Bold text highlights the results with statistical significance.

in Spaniards (Enrich et al., 2019), and surprisingly in 3.84% of Achalasia Mexican patients (Furuzawa-Carballeda et al., 2018). Therefore, the increased frequency of this new haplotype in Mexican Mestizo patients may represent the convenience of preserving this haplotype as an immunological advantage against pathogens.

On the other hand, it is observable that the percentages of this haplotype in SLE and Achalasia Mexicans are similar to healthy Spaniards. So, we would expect scenarios with similar prevalence and disease phenotypes if it only depended on HLA susceptibility alleles. Unfortunately, there are no official statistics about the prevalence and incidence of SLE in Mexico. Still, reports of SLE in Hispanics show high SLE prevalence in Hispanics than Spaniards (138/100000 and 17.5 to 34.1/100000 per inhabitants per year, respectively) (Atisha-Fregoso et al., 2011; Izmirly et al., 2017; Rees et al., 2017). Furthermore, some severe manifestations such as kidney damage are also more evident in Hispanics (62%) than Caucasian patients (25%).

TABLE 6 | Frequencies of HLA-C∼B∼DRB1∼DQB1 haplotypes in SLE patients and healthy individuals.

HLA-C∼B∼DRB1∼DQB1 Haplotypes	**SLE *N* = 143 (286 alleles)**			**Healthy controls *N* = 234 (468 alleles)**			***pC***	**OR**	**%95IC**
	n	**HF**	**Δ′**	***n***	**HF**	**Δ′**			
Amerindian									
C*07:02∼B*39:05∼DRB1*04:07∼DQB1*03:02	16	0.0559	0.663	19	0.0406	0.502	ns		
C*04:01∼B*35:12∼DRB1*08:02∼DQB1*04:02	10	0.0350	0.717	7	0.0150	0.305	ns		
C*04:01∼B*35:17∼DRB1*08:02∼DQB1*04:02	10	0.0350	0.591	14	0.0299	0.726	ns		
C*01:02∼B*15:15∼DRB1*08:02∼DQB1*04:02	5	0.0175	0.795	8	0.0171	0.525	ns		
C*03:03∼B*52:01∼DRB1*14:06∼DQB1*03:01	5	0.0175	0.826	3	0.0064	0.446	ns		
C*07:02∼B*39:06∼DRB1*14:06∼DQB1*03:01	4	0.0140	0.478	16	0.0342	0.548	ns		
C*03:04∼B*40:02∼DRB1*08:02∼DQB1*04:02	3	0.0105	0.181	2	0.0043	–0.044	ns		
C*01:02∼B*15:30∼DRB1*08:02∼DQB1*04:02	2	0.0070	1.000	4	0.0085	0.383	ns		
C*08:01∼B*48:01∼DRB1*08:02∼DQB1*04:02	2	0.0070	1.000	8	0.0171	1.000	ns		
C*01:02∼B*15:01∼DRB1*16:02∼DQB1*03:01	1	0.0035	0.321	2	0.0043	0.237	ns		
C*01:02∼B*35:43∼DRB1*08:02∼DQB1*04:02	1	0.0035	0.181	2	0.0043	0.040	ns		
C*07:02∼B*39:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	0.181	2	0.0043	0.383	ns		
C*15:02∼B*51:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	1.000	4	0.0085	0.314	ns		
Caucasian									
C*07:01∼B*08:01∼DRB1*03:01∼DQB1*02:01	**17**	**0.0594**	**0.883**	**2**	**0.0043**	**1.000**	**0.00001**	**14.7**	**3.38 64.23**
C*16:01∼B*44:03∼DRB1*07:01∼DQB1*02:02	9	0.0315	0.891	6	0.0128	0.734	ns		
C*07:02∼B*07:02∼DRB1*15:01∼DQB1*06:02	6	0.0210	0.514	7	0.0150	0.448	ns		
C*05:01∼B*18:01∼DRB1*03:01∼DQB1*02:01	4	0.0140	1.000	3	0.0064	0.587	ns		
C*08:02∼B*14:02∼DRB1*01:02∼DQB1*05:01	4	0.0140	1.000	5	0.0107	0.441	ns		
C*05:01∼B*44:02∼DRB1*04:02∼DQB1*03:02	2	0.0070	1.000	2	0.0043	0.489	ns		
C*06:02∼B*13:02∼DRB1*07:01∼DQB1*02:02	2	0.0070	1.000	4	0.0085	0.787	ns		
C*07:01∼B*57:01∼DRB1*07:01∼DQB1*03:03	1	0.0035	1.000	3	0.0064	1.000	ns		
Caucasian shared with other populations									
C*04:01∼B*35:01∼DRB1*04:04∼DQB1*03:02	2	0.0070	0.077	3	0.0064	0.147	ns		
C*03:03∼B*52:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	∼0.103	2	0.0043	0.176	ns		
C*06:02∼B*50:01∼DRB1*03:01∼DQB1*02:01	1	0.0035	0.442	2	0.0043	0.483	ns		
C*08:02∼B*14:01∼DRB1*07:01∼DQB1*02:02	1	0.0035	0.276	3	0.0064	0.734	ns		
C*08:02∼B*14:02∼DRB1*03:01∼DQB1*02:01	1	0.0035	0.044	2	0.0043	0.155	ns		
Unknown									
C*04:01∼B*35:01∼DRB1*08:02∼DQB1*04:02	4	0.0140	0.099	3	0.0064	0.012	ns		
C*07:02∼B*39:05∼DRB1*08:02∼DQB1*04:02	4	0.0140	∼0.064	5	0.0107	–0.227	ns		
C*08:01∼B*48:01∼DRB1*04:04∼DQB1*03:02	3	0.0105	0.335	3	0.0064	0.147	ns		
C*04:01∼B*35:14∼DRB1*16:02∼DQB1*03:01	2	0.0070	0.661	4	0.0085	0.644	ns		
C*07:02∼B*39:06∼DRB1*04:07∼DQB1*03:02	2	0.0070	0.170	4	0.0085	0.039	ns		
C*01:02∼B*15:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	0.181	3	0.0064	0.294	ns		
C*06:02∼B*37:01∼DRB1*01:03∼DQB1*05:01	1	0.0035	0.493	2	0.0043	0.665	ns		

HF, haplotype frequency; ns, not significant; pC, p corrected value using Bonferroni method; OR, odds ratio; 95%CI, 95% confidence interval. N, total number of individuals, n, absolute frequency for each allele; Δ′, Linkage disequilibrium value. Bold text highlights the results with statistical significance.

In addition, activity disease score as Systemic Lupus Activity Measure revised (SLAM-R) has been found highest in Hispanic than Caucasian patients (Fernández et al., 2007). Thus, all the above shows that the severity and clinical manifestations differ in an ethnic-dependent manner even when susceptibility haplotypes like the one found in this study are shared between populations. Therefore, the importance of population genetic studies and admixture estimations on Countries with high variability as Mexico is. No less important, there is a conjunction of variables as the availability and access to medical services and treatments that modify the severity phenotype in different populations. This factor always is important to consider before concluding about ethnicity and admixture disparities among populations.

The ethnic background and admixture influence the development of SLE and autoimmunity in Mexicans. This assumption was introduced because the haplotype HLA-B8-DR3 is found in high frequencies in the Caucasian population and has a considerable prevalence of SLE and other autoimmune diseases (Awdeh et al., 1983; Alper et al., 1986; Szilágyi et al., 2010). In contrast, the Mexican Native American populations lack both the HLA-B8-DR3 haplotype and SLE cases. Therefore, the high

TABLE 7 | Frequencies of HLA∼A/∼C/∼B/∼DRB1/∼DQB1 haplotypes in SLE patients and healthy individuals.

HLA∼A/∼C/∼B/∼DRB1/∼DQB1 Haplotypes	**SLE** *N* = 286			**Healthy Individuals** *N* = 468			***pC***	**OR**	**95%IC**	
	n	**HF**	**Δ'**	***n***	**HF**	**Δ'**				
Amerindian										
A*02:01∼C*04:01∼B*35:12∼DRB1*08:02∼DQB1*04:02	6	0.0210	0.493	4	0.0085	0.444	ns			
A*68:03∼C*07:02∼B*39:05∼DRB1*04:07∼DQB1*03:02	5	0.0175	0.380	5	0.0107	0.283	ns			
A*02:06∼C*07:02∼B*39:05∼DRB1*04:07∼DQB1*03:02	4	0.0140	0.203	5	0.0107	0.185	ns			
A*02:01∼C*01:02∼B*15:15∼DRB1*08:02∼DQB1*04:02	3	0.0105	0.239	3	0.0064	0.190	ns			
A*02:01∼C*04:01∼B*35:17∼DRB1*08:02∼DQB1*04:02	2	0.0070	0.013	7	0.0150	0.352	ns			
A*02:06∼C*04:01∼B*35:17∼DRB1*08:02∼DQB1*04:02	2	0.0070	0.174	2	0.0043	0.517	ns			
A*24:02∼C*04:01∼B*35:12∼DRB1*08:02∼DQB1*04:02	2	0.0070	0.059	2	0.0043	0.141	ns			
A*24:02∼C*07:02∼B*39:06∼DRB1*14:06∼DQB1*03:01	2	0.0070	0.412	12	0.0256	0.699	ns			
A*31:01∼C*04:01∼B*35:17∼DRB1*08:02∼DQB1*04:02	2	0.0070	0.161	2	0.0043	0.069	ns			
A*68:01∼C*01:02∼B*15:15∼DRB1*08:02∼DQB1*04:02	2	0.0070	0.368	3	0.0064	0.321	ns			
A*02:01∼C*08:01∼B*48:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	0.366	3	0.0064	0.190	ns			
Caucasian										
A*01:01∼C*07:01∼B*08:01∼DRB1*03:01∼DQB1*02:01	**11**	**0.0385**	**0.656**	**1**	**0.0021**	**0.481**	**0.0004**	**18.7**	**2.40**	**145.48**
A*29:02∼C*16:01∼B*44:03∼DRB1*07:01∼DQB1*02:02	**8**	**0.0280**	**0.884**	**2**	**0.0043**	**0.316**	**0.02**	**6.7**	**1.41**	**31.80**
A*02:01∼C*07:02∼B*07:02∼DRB1*15:01∼DQB1*06:02	3	0.0105	0.366	4	0.0085	0.444	ns			
A*30:02∼C*05:01∼B*18:01∼DRB1*03:01∼DQB1*02:01	3	0.0105	0.742	3	0.0064	1.000	ns			
A*02:01∼C*16:01∼B*44:03∼DRB1*07:01∼DQB1*02:02	1	0.0035	–0.475	2	0.0043	0.135	ns			
Unknown										
A*02:01∼C*03:03∼B*52:01∼DRB1*14:06∼DQB1*03:01	4	0.0140	0.746	2	0.0043	0.568	ns			
A*02:01∼C*07:02∼B*39:05∼DRB1*04:07∼DQB1*03:02	3	0.0105	–0.114	6	0.0128	0.113	ns			
A*02:01∼C*01:02∼B*15:30∼DRB1*08:02∼DQB1*04:02	2	0.0070	1.000	3	0.0064	0.676	ns			
A*02:01∼C*01:02∼B*35:43∼DRB1*08:02∼DQB1*04:02	1	0.0035	1.000	2	0.0043	1.000	ns			
A*02:01∼C*04:01∼B*35:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	0.049	2	0.0043	0.568	ns			
A*24:02∼C*07:02∼B*39:06∼DRB1*04:07∼DQB1*03:02	1	0.0035	0.412	2	0.0043	0.399	ns			
A*30:01∼C*06:02∼B*13:02∼DRB1*07:01∼DQB1*02:02	1	0.0035	0.494	4	0.0085	1.000	ns			
A*68:01∼C*07:02∼B*39:01∼DRB1*08:02∼DQB1*04:02	1	0.0035	1.000	2	0.0043	1.000	ns			

HF, haplotype frequency; ns, no significant; pC, p corrected value using Bonferroni method; OR, odds ratio; 95%IC, 95% interval confidence; N, total number of individuals; n, absolute frequency for each allele; Δ', Linkage disequilibrium value. Bold text highlights the results with statistical significance.

frequency of HLA-B8-DR3 in SLE Mexican patients from Mexico City has been explained as the product of the introduced alleles and haplotypes during the Spaniard's arrival to the Americas. Surprisingly, SLE is more prevalent and aggressive in Mexicans than Europeans. Since the origin of the susceptibility alleles is from Europeans, it would be expected that the severity of the disease and the phenotypes would be similar. However, many elements can intervene to generate a unique scenario in Mexico. For example, the specific interaction of Amerindian, European, and other populations genes influenced by the architecture and tridimensional arrangement of the genes in each mixed population. Not least, the pressure exerted by local triggers and the environment itself.

The complex scenario of autoimmunity in Mexico is framed in that the susceptibility varies as the ethnic admixture pattern does. It has been corroborated in SLE studies. For example, SLE Mestizo patients in Guadalajara (northwestern Mexico) have the haplotype HLA-DRB1*15-DQA1*01:02-DQB1*06:02 as a susceptibility factor (Cortes et al., 2004) while SLE Mestizo patients from Tapachula, Chiapas (southeast of Mexico) have the susceptibility alleles HLA-DRB1*15 and -DRB1*16 (Sepúlveda Delgado et al., 2018). But neither in Guadalajara nor in Tapachula, the HLA-DRB1*03:01 is a susceptibility factor. However, HLA-DRB1*03:01 is one of the main susceptibility alleles found in this study and other previously conducted in Mexico City (Granados et al., 1996; Vargas-Alarcón et al., 2001). Parallel to these susceptibility differences, independent studies of population genetics have demonstrated de admixture variability among Guadalajara, Tapachula, and Mexico City (current study). The Native American Mexican load in Guadalajara, Tapachula, and Mexico City is ∼44%, ∼72%, ∼63%, respectively, while the European is ∼48%, ∼26%, ∼30%, and the African is ∼8%, ∼2%, ∼6% (Barquera et al., 2020d; Bravo-Acevedo et al., 2020). Therefore, it looks like admixture proportions in each Mexican State or region could drive toward specific susceptibilities for the same disease. However, it is not only the variation of admixture across Mexico. The presence of different triggers in each region could shape a more comprehensive scenery.

Therefore, the influence of admixture variations in SLE susceptibility could be briefly explained as follows: having a more European HLA load increases the chances of carrying a risk haplotype of this origin. Likewise, there are risk

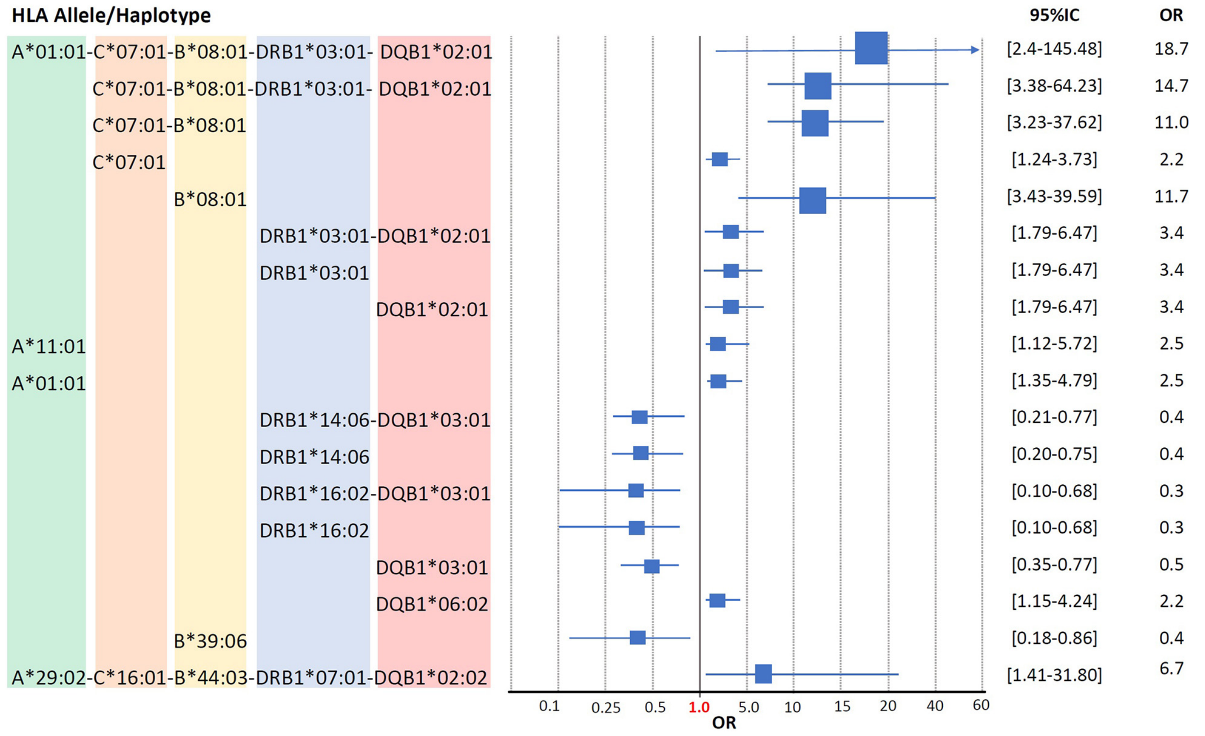


FIGURE 4 | Human leukocyte antigen susceptibility and protection alleles and haplotypes in Mestizo Mexican SLE patients. The alleles and haplotypes, which showed statistically significant differences between SLE patients and healthy individuals ($p < 0.05$), are shown in the Forest plot. The Odds ratios and 95% interval confidence are included in the graphic. Blue boxes are not at scale.

alleles and haplotypes of African and Asian origin; however, because the percentage of these ancestries is low in Mexican Mestizos, they are less likely to carry them. HLA alleles have been conserved in the Mexican Mestizo, probably because they represent an immunological advantage against infections by common pathogens in Mexico. However, the cost of efficiency in pathogen clearance has predisposed to immune hyperresponsiveness, prolonging inflammatory pathways, and leading to autoimmunity. Also, the advance in medicine and better treatments have augmented the survival, life expectancy, and quality of life, which allows the preservation of susceptibility alleles. All the above might be part of the eventual increase in statistics of autoimmune diseases in Mexico, a recent phenomenon caused mainly due to better diagnosis, but influenced by the recently acquired HLAs that have contributed to the fitness of Mexican Mestizos against local pathogens, but with less tolerance to the presence of triggers. This means the carriers of susceptibility alleles or haplotypes are "selected" to manifest the disease based on the presence of the triggers. Therefore, this selection includes individuals with susceptibility alleles and, since those with higher European ancestry are those that most likely have susceptibility alleles; these are the individuals that enrich the autoimmune group.

In truth, there are many assumptions for which further investigation should be carried out. However, the above assumptions could explain why the SLE group differs in admixture proportions compared to control individuals. In studies like the current, it is necessary to ensure the comparability of the groups of patients and controls since useful and accurate results depend on that data quality. Recently published characterization of individuals from Mexico City has been an additional resource to validate the admixture estimations in our control group. In this study were studied 1217 individuals northern (N = 751), southern (N = 52), eastern (N = 79), western (N = 33), and central (N = 152) Mexico City and surroundings. Admixture estimates are very similar to the calculations performed by Maximum likelihood in our study, being Native American Mexican 63.85%, European 28.53%, and African 7.61% (Barquera et al., 2020e).

Further evidence is shown in the PCA plot; our Mexican admixed samples (SLE and Healthy individuals) are separated from the non-autochthonous populations, which was expected. But, notably, the "Mestizo" sample of SLE patients from Mexico City is not overlapped with Controls.

Systemic lupus erythematosus is closer to non-autochthonous populations, while Healthy individuals are closer to autochthonous populations, congruent with admixture estimations and genetic distances. However, SLE patients retain their identity with the States of Central Mexico, which is expected because SLE patients belong to Mexico City; but, the percentages of ancestry differ, having a greater European load than expected for individuals from Mexico City.

This fact is particularly striking because we are talking about Mestizo Mexican Individuals who share the locality of born and the demographic history of settlement (colonial period) but have notable differences in ancestry proportions. This ambivalence in the SLE group could be explained by the fact that the local Native American alleles that confer identity to the Central zone of Mexico are conserved (which is detected in the PCA). At the same time, the advantageous foreign alleles have also been conserved, slightly modifying the proportions of the ancestry of the patients (admixture estimations and genetic distances) and augmenting the possibilities of carrying susceptibility foreign HLA alleles.

The ancestry variation in individuals with SLE, as said before, could mean that some non-autochthonous HLA alleles have been conserved because these alleles represent an immunogenetic advantage, as demonstrated for HLA-DRB1*03. For instance, some naturally processed HLA-DR3-restricted Human herpesvirus 6B (HHV-6B) peptides are recognized broadly with polyfunctional and cytotoxic CD4 T-cell responses (Becerra-Artiles et al., 2019). However, this convenient immunogenetic mechanism for clearing infections might be the detonator for immunological hyperreactivity, which marks the onset of the disease. Viruses infections associated with SLE have been described previously and are ethnically associated with SLE pathogenesis, specific HLA class II alleles, and the development of antinuclear antibodies, the hallmark of SLE have been associated in the same studies (Perl et al., 1995; Magistrelli et al., 1999; Quaglia et al., 2021).

Likewise, Nei's distance data is congruent with the information given by both the susceptibility and protection haplotypes and the admixture analysis. It was found that the SLE group is closer to foreign parental populations (Southwestern European, Sub-Saharan African, and Eastern Asian) than the healthy individuals. In contrast, healthy individuals are nearer to Mexican Native American populations evaluated (Mixtecs, Zapotecans, Tarahumaras, Lacandon, Mixe, and Seri). As expected for intra-specie analysis for the same geographic area, the genetic distance variability is low, but the differences are consistent and significant. Corrected average pairwise differences gave the same information.

The matrix of genetic distances reflects the differences between the Mestizos of Mexico City (SLE and controls), **Figure 2**. As mentioned before about PCA, SLE and controls are Mestizos from the same geographical area, so it would be expected that there would be no variation in admixture proportions. But indeed, there is. The matrix shows consistent variation indicating SLE patients are genetically more similar to non-autochthonous populations. In comparison, healthy individuals are genetically more similar to Mexican Native American populations. However, some aspects about populations included are worth exploring to substantiate the validity of the comparison.

Regarding the notable genetic variation of the Mexican indigenous groups included, it was expected. The presented Fst matrix and values (**Supplementary Figure 1**), and the other calculations as coancestry and Slatkin's Fst (**Supplementary Figures 2, 3**), reflect a bit of what has been described for the native populations that inhabited Mexico. It has been corroborated a striking genetic stratification among indigenous populations within Mexico at varying degrees of geographic isolation. Some groups are differentiated as Europeans are from East Asians. Seris and Lacandons are good examples (Infante et al., 1999; Barquera et al., 2020c); these groups are exposed to high levels of genetic drift and isolation. However, the value of including them as Mexican Native American parental populations is that present-day Mexicans' genetic composition recapitulates ancient Native American substructure, despite the potential homogenizing effect of postcolonial admixture. Fine-scale population structure going back centuries is not merely a property of isolated or rural indigenous communities. Cosmopolitan populations still reflect the underlying genetic ancestry of local native populations, arguing for a strong relationship between the indigenous and the Mexican mestizo population, albeit without the extreme drift exhibited in some current indigenous groups (Moreno-Estrada et al., 2014).

Another notable feature is the genetic distances between the selected non-indigenous populations. Mainly, the Fst indicates an unexpected but explainable closeness between Zimbabwe and the Spanish. The characteristics of the Zimbabwe Harare Shona sample and previous periods of European colonization may have had a specific influence on this population that we cannot trace but hypothesize. Non-local HLA alleles may have been conserved as an immunological advantage, which has already been mentioned for Mexican Mestizos, additional to the MHC Class III genes with immunological functions conserved due to the linkage disequilibrium with the possibly acquired HLA Class I and Class II variants. Therefore, it should be noted that although the European load is likely to be low in this group of sub-Saharan Africans, the variants that contribute to fitness can be conserved and detected in the genetic distance calculations.

Furthermore, estimates of admixture by HLA are robust. In other studies, in Mexico City individuals, the admixture estimates are similar to the current study. It should be noted that these studies used HLA parental populations data different from that used in this study. Estimates with other techniques such as STRs have also generated similar results (Juárez-Cedillo et al., 2008). Thus, if the selected parental populations agree with the settlement background in Mexico, they can bring us the correct information about admixture and the genetic distances.

Finally, the CEH described as a susceptibility factor for Europeans the HLA-A*01:01∼C*07:01∼B*08:01∼DRB1*02:01 ∼DQB1*02:01 has been found as a high-risk factor for developing SLE. This CEH has a powerful influence on SLE development and other autoimmune diseases in mestizo Mexican individuals from Mexico City (Granados et al., 1996) and other populations. At first glance, this haplotype's relative risk appears to have a summative effect by the alleles that make up it; as more HLA alleles are added to the haplotype, the risk increases. However, the relative risk conferred by HLA-B*08:01, HLA-DRB1*03:01, and haplotypes containing them confirm that the risk is not summative but conferred by the genes in linkage disequilibrium with these variants. Besides, this linkage has been turned more complicated during the HLA research course. Early in the study of HLA was described a strong linkage disequilibrium not only among HLA genes of the haplotype HLA-B8-DR3 but between HLA and complement

(C2, C4a, C4b, and Factor B) (Black et al., 1982; Wescott et al., 1987; Picceli et al., 2016). Currently, it is known that a considerable number of variants, mainly of MHC class three, are in linkage disequilibrium with the variants of B and DR indicated above, which increases the genetic risk load of the individual with lupus.

Models have been developed to confirm the susceptibility conferred by genes in the MHC III region (non-classical HLA and non-HLA genes) linked to HLA-B and -DRB1. The most important linked genes using HLA-DRB1*03:01 as covariant are: the proto-oncogene Notch homologue 4 (NOTCH4), the MHC class I polypeptide–related sequence A and B (MIC-A and MIC-B), the steroid 21-hydroxylase (CYP21A2), (Partanen et al., 1988; Morris et al., 2012) the three 70 kDa heat-shock proteins (HSPA1A, HSPA1B, HSPA1L), (Mišunová et al., 2017), natural cytotoxicity triggering receptor 3 (NCR3); nuclear factor kappa light chain gene enhancer in B cells inhibitor-like 1 (NFKBIL1), allograft inflammatory factor 1 (AIF1) (Dorak et al., 2006); and the three Tumor Necrosis Factor genes (Lymphotoxin-beta, Lymphotoxin-alpha, and TNF-α) (Zúñiga et al., 2001; Ramírez-Bello et al., 2018). Important variants of the genes mentioned above have been deeply studied in Caucasians, and linkage disequilibrium with HLA-B*08:01 and -DRB1*03:01 has been confirmed (Dorak et al., 2006). This type of study is a prospect to be carried out in the Mexican population to determine the integrity of the MHC blocks conserved in Mexican Mestizos and how the variants could have been fixed or discarded, giving the specific characteristics of SLE in Mexicans.

CONCLUSION

In conclusion, the genetic background and the Mexican admixture heterogeneity define part of the dynamic of SLE in Mestizo Mexican patients. Considering the regional distribution of genetic susceptibility and the ethnic barriers can be a useful biomedical tool. Increasingly, the born region and the ethnic admixture (the apparent ethnic phenotypes based on physical characteristics) are considered an element that helps the diagnosis of diseases. Such is the case of neuromyelitis optica and multiple sclerosis, which may be clinically similar but have well-established genetic susceptibility and ethnic differences. In Mexico, the susceptibility to neuromyelitis optica is given by HLA-DRB1*16:02, and it usually occurs in individuals with a high Mexican Native American ancestry (Gorodezky et al., 1986; Romero-Hidalgo et al., 2020) (and usually, the physical characteristics indicate it). But in multiple sclerosis, the susceptibility is mainly given by HLA-DRB1*15, and it occurs in individuals with higher Caucasian ancestry (Ordoñez et al., 2015). Another good example includes Multifocal Epithelial Hyperplasia, which depends on genetic susceptibility and is expressed in specific populations with high Mexican Native American components.

On the other hand, the technical aspect of this type of study is important. HLA sequencing study has helped to describe the ethnic admixture load in the patient and control populations. It deepens susceptibility and protection alleles and haplotypes due to high-fidelity performance. Finally, non-HLA variants in the MHC area contribute to the risk through linkage disequilibrium with HLA genes. Therefore, the CEHs help to explain better the susceptibility and protection in groups with known ethnic composition. However, the linkage of these variants with HLA requires a more in-depth analysis to understand how they add risk to CEH and the specific variants conserved in the Mexican population.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the online repository: Allele frequency net database (AFND) with the accession Mexico Mexico City Mestizo population (*n* = 143) (SLE). African: Burkina Faso Mossi (*n* = 53), Cameroon Yaounde (*n* = 92), Ghana Ga-Adangbe (*n* = 141), Kenya (*n* = 144), Uganda Kampala pop 2 (*n* = 175), Kenya Nandi (*n* = 240), Zimbabwe Harare Shona (*n* = 230). European: England Blood Donors of Mixed Ethnicity (*n* = 519), France Lyon (*n* = 4813), Germany Essen (*n* = 174), Ireland Northern (*n* = 1000), Italy Lombardy (*n* = 674), Italy Sardinia pop 3 (*n* = 100), Netherlands UMCU (*n* = 64), Spain Catalunya, Navarra, Extremadura, Aragon, Cantabria (*n* = 4335). Mestizo Mexican populations from Northern Mexico: Mexico Baja California, Mexicali (*n* = 100), Mexico Chihuahua Chihuahua City (*n* = 119), Mexico Colima, Colima city (*n* = 61), Mexico Durango, Durango city (*n* = 153), Mexico Nuevo Leon, Monterrey city (*n* = 266), Mexico Sinaloa Rural (*n* = 183), Mexico Sonora Rural (*n* = 197). Mestizo Mexican populations from the Center of Mexico: Mexico Aguascalientes State (*n* = 95), Mexico Guanajuato Rural (*n* = 162), Mexico Guerrero state (*n* = 144), Mexico Jalisco, Guadalajara city (*n* = 1189), Mexico Mexico City Center (*n* = 152), Mexico Mexico City West (*n* = 33), Mexico Mexico City East (*n* = 79), Mexico Mexico City South (*n* = 52), Mexico Mexico City North (*n* = 751), Mexico Michoacan Rural (*n* = 348), Mexico Nayarit, Tepic (*n* = 97), Mexico Queretaro, Queretaro city (*n* = 45), Mexico San Luis Potosi Rural (*n* = 87), Mexico Mexico City Mestizo pop 2 (*n* = 234) (Controls). Mestizo Mexican populations from Southern Mexico: Mexico Chiapas Rural (*n* = 121), Mexico Oaxaca, Oaxaca city (*n* = 151), Mexico Quintana Roo Rural (*n* = 50), Mexico Tabasco Rural (*n* = 142), Native American Mexican populations: Mexico Nahuas (*n* = 72), Mexico Oaxaca Mixe (*n* = 55), Mexico Mixtec (*n* = 97), Mexico Chihuahua Tarahumara (*n* = 97), Mexico Oaxaca Mixe (*n* = 55), Mexico Oaxaca Mixtecs (*n* = 103), Mexico Oaxaca Zapotec (*n* = 90), Mexico Sonora Seri (*n* = 34), Mexico Chiapas Lacandon Mayans (*n* = 218). Non-Mexican Latin-American populations: Argentina Gran Chaco Eastern Toba (*n* = 125), Bolivia La Paz Aymaras (*n* = 88), Colombia Barranquilla (*n* = 188), Costa Rica African-Caribbean (*n* = 102), Costa Rica Amerindians (*n* = 125), Ecuador Andes Mixed Ancestry (*n* = 824), Nicaragua Managua (*n* = 339), Panama (*n* = 462). Asian: China Han (*n* = 314) http://www.allelefrequencies.net/ (Gonzalez-Galarza et al., 2020).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics Committee from the Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán. Reference No. 1738. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SH-D, JG, GV-A, and JZ performed the conceptualization of the study. SH-D and JM-G performed the methodology. SH-D, JJ-O, LL, and GL investigated the study. JZ, JG, DR, VT-M, LL, and GL did the contribution with resources. SH-D, JM-G, and VA-A carried out the data curation. SH-D wrote the original draft and visualized the data. SH-D, JG, GV-A, and LL wrote, reviewed, and edited the draft of the manuscript. JZ, GV-A, and JG supervised the project. JG carried out the project administration. JG and JZ carried out the funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

SH-D was a doctoral student from the Biomedical Sciences Doctoral Program, Universidad Nacional Autónoma de México (UNAM) and has received CONACYT fellowship 619959, Support numbers 463857 and CVU 689031. This work was supported by grant 000 000 000273175 from the CONACYT (Consejo Nacional de Ciencia y Tecnología de México).

ACKNOWLEDGMENTS

The authors thank all individuals with SLE and healthy controls who have participated in this genetic study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.701373/full#supplementary-material

REFERENCES

Alarcón-Riquelme, M. E., Ziegler, J. T., Molineros, J., Howard, T. D., Moreno-Estrada, A., Sánchez-Rodríguez, E., et al. (2016). Genome-wide association study in an Amerindian ancestry population reveals novel systemic lupus erythematosus risk loci and the role of European admixture. *Arthritis Rheumatol.* 68, 932–943. doi: 10.1002/art.39504

Alper, C. A., Awdeh, Z. L., and Yunis, E. J. (1986). Complotypes, extended haplotypes, male segregation distortion, and disease markers. *Hum. Immunol.* 15, 366–373. doi: 10.1016/0198-8859(86)90013-3

Arnaiz-Villena, A., Vargas-Alarcón, G., Areces, C., Enríquez-de-Salamanca, M., Abd-El-Fatah-Khalil, S., Fernández-Honrado, M., et al. (2014). Mixtec Mexican Amerindians: an HLA alleles study for America peopling, pharmacogenomics and transplantation. *Immunol. Invest.* 43, 738–755. doi: 10.3109/08820139.2014.926369

Arrieta-Bolaños, E., Madrigal-Sánchez, J. J., Stein, J. E., Órlich-Pérez, P., Moreira-Espinoza, M. J., Paredes-Carias, E., et al. (2018). High-resolution HLA allele and haplotype frequencies in majority and minority populations of Costa Rica and Nicaragua: Differential admixture proportions in neighboring countries. *Hla* 91, 514–529. doi: 10.1111/tan.13280

Atisha-Fregoso, Y., Jakez-Ocampo, J., and Llorente, L. (2011). Systemic lupus erythematosus in Hispanics. *Autoimmunity* 44, 555–561. doi: 10.3109/08916934.2011.592882

Awdeh, Z. L., Raum, D., Yunis, E. J., and Alper, C. A. (1983). Extended HLA/complement allele haplotypes: evidence for T/t-like complex in man. *Proc. Natl. Acad. Sci. U.S.A.* 80, 259–263. doi: 10.1073/pnas.80.1.259

Ballesteros-Romero, M., Barquera, R., Rodríguez-López, M. E., Hernández-Zaragoza, D. I., Goné-Vázquez, I., Clayton, S., et al. (2020). Genetic diversity of HLA system in two populations from Michoacán, Mexico: morelia and rural Michoacán. *Hum. Immunol.* 81, 506–509. doi: 10.1016/j.humimm.2019.05.017

Barquera, R., Bravo-Acevedo, A., Clayton, S., Munguía, T. J. R., Hernández-Zaragoza, D. I., Adalid-Sáinz, C., et al. (2020a). Genetic diversity of HLA system in two populations from Nuevo León, Mexico: monterrey and rural Nuevo León. *Hum. Immunol.* 81, 516–518. doi: 10.1016/j.humimm.2019.06.003

Barquera, R., Hernández-Zaragoza, D. I., Arellano-Prado, F. P., Goné-Vázquez, I., Clayton, S., Arrieta-Bolaños, E., et al. (2020b). Genetic diversity of HLA system in two populations from Colima, Mexico: colima city and rural Colima. *Hum. Immunol.* 81, 513–515. doi: 10.1016/j.humimm.2019.06.004

Barquera, R., Hernández-Zaragoza, D. I., Bravo-Acevedo, A., Arrieta-Bolaños, E., Clayton, S., Acuña-Alonzo, V., et al. (2020c). The immunogenetic diversity of the HLA system in Mexico correlates with underlying population genetic structure. *Hum. Immunol.* 81, 461–474. doi: 10.1016/j.humimm.2020.06.008

Barquera, R., Juárez-Nicolás, F., Martínez-Álvarez, J. C., Ponnandai-Shanmugavel, K. S., Hernández-Zaragoza, D. I., Vázquez-Castillo, T. V., et al. (2020d). Genetic diversity of HLA system in two populations from Chiapas, Mexico: tuxtla Gutiérrez and rural Chiapas. *Hum. Immunol.* 81, 563–565. doi: 10.1016/j.humimm.2019.07.285

Barquera, R., Martínez-Álvarez, J. C., Hernández-Zaragoza, D. I., Bravo-Acevedo, A., Juárez-Nicolás, F., Arriaga-Perea, A. J., et al. (2020e). Genetic diversity of HLA system in six populations from Mexico City Metropolitan Area, Mexico: Mexico City North, Mexico City South, Mexico City East, Mexico City West, Mexico City Center and rural Mexico City. *Hum. Immunol.* 81, 539–543. doi: 10.1016/j.humimm.2019.07.297

Barquera, R., Zuniga, J., Flores-Rivera, J., Corona, T., Penman, B. S., Hernández-Zaragoza, D. I., et al. (2020f). Diversity of HLA Class I and Class II blocks and conserved extended haplotypes in Lacandon Mayans. *Sci. Rep.* 10:3248. doi: 10.1038/s41598-020-58897-5

Becerra-Artiles, A., Cruz, J., Leszyk, J. D., Sidney, J., Sette, A., Shaffer, S. A., et al. (2019). Naturally processed HLA-DR3-restricted HHV-6B peptides are recognized broadly with polyfunctional and cytotoxic CD4 T-cell responses. *Eur. J. Immunol.* 49, 1167–1185. doi: 10.1002/eji.201948126

Becker, J., Haas, S. L., Mokrowiecka, A., Wasielica-Berger, J., Ateeb, Z., Bister, J., et al. (2016). The HLA-DQβ1 insertion is a strong achalasia risk factor and displays a geospatial north-south gradient among Europeans. *Eur. J. Hum. Genet.* 24, 1228–1231. doi: 10.1038/ejhg.2015.262

Black, C. M., Welsh, K. I., Fielder, A., Hughes, G. R., and Batchelor, J. R. (1982). HLA antigens and Bf allotypes in SLE: evidence for the association being with specific haplotypes. *Tissue Antigens* 19, 115–120. doi: 10.1111/j.1399-0039.1982.tb01426.x

Bombardier, C., Gladman, D. D., Urowitz, M. B., Caron, D., Chang, C. H., Austin, A., et al. (1992). Derivation of the sledai. *A disease activity index for lupus patients. Arthritis Rheumatism* 35, 630–640. doi: 10.1002/art.1780350606

Bravo-Acevedo, A., Escobedo-Ruíz, A., Barquera, R., Clayton, S., García-Arias, V. E., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in six populations from Jalisco, Mexico: Guadalajara city, Tlajomulco, Tlaquepaque, Tonalá, Zapopan and rural Jalisco. *Hum. Immunol.* 81, 502–505. doi: 10.1016/j.humimm.2019.05.012

Cao, K., Hollenbach, J., Shi, X., Shi, W., Chopek, M., and Fernández-Viña, M. A. (2001). Analysis of the frequencies of HLA-A, B, and C alleles and haplotypes in the five major ethnic groups of the United States reveals high levels of diversity in these loci and contrasting distribution patterns in these populations. *Hum. Immunol.* 62, 1009–1030. doi: 10.1016/s0198-8859(01)00298-1

Cao, K., Moormann, A. M., Lyke, K. E., Masaberg, C., Sumba, O. P., Doumbo, O. K., et al. (2004). Differentiation between African populations is evidenced by the diversity of alleles and haplotypes of HLA class I loci. *Tissue Antigens* 63, 293–325.

Carter, E. E., Barr, S. G., and Clarke, A. E. (2016). The global burden of SLE: prevalence, health disparities and socioeconomic impact. *Nat. Rev. Rheumatol.* 12, 605–620. doi: 10.1038/nrrheum.2016.137

Centers for Disease Control and Prevention (2011). *Epi Info^{TM}, a Database and Statistics Program for Public Health Professionals*. Atlanta, GA: Centers for Disease Control and Prevention.

Cerna, M., Falco, M., Friedman, H., Raimondi, E., Maccagno, A., Fernandez-Viña, M., et al. (1993). Differences in HLA class II alleles of isolated South American Indian populations from Brazil and Argentina. *Hum. Immunol.* 37, 213–220. doi: 10.1016/0198-8859(93)90504-T

Clayton, S., Barquera, R., Uribe-Duarte, M. G., Goné Vázquez, I., Zúñiga, J., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Sinaloa, Mexico: culiacán and rural Sinaloa. *Hum. Immunol.* 81, 482–484. doi: 10.1016/j.humimm.2019.06.006

Cortes, L. M., Baltazar, L. M., Lopez-Cardona, M. G., Olivares, N., Ramos, C., Salazar, M., et al. (2004). HLA class II haplotypes in Mexican systemic lupus erythematosus patients. *Hum. Immunol.* 65, 1469–1476. doi: 10.1016/j.humimm.2004.09.008

Devilliers, H., Amoura, Z., Besancenot, J. F., Bonnotte, B., Pasquali, J. L., Wahl, D., et al. (2012). LupusQoL-FR is valid to assess quality of life in patients with systemic lupus erythematosus. *Rheumatology (Oxford)* 51, 1906–1915. doi: 10.1093/rheumatology/kes165

Dorak, M. T., Shao, W., Machulla, H. K., Lobashevsky, E. S., Tang, J., Park, M. H., et al. (2006). Conserved extended haplotypes of the major histocompatibility complex: further characterization. *Genes Immun.* 7, 450–467. doi: 10.1038/sj.gene.6364315

Enrich, E., Campos, E., Martorell, L., Herrero, M. J., Vidal, F., Querol, S., et al. (2019). HLA-A, -B, -C, -DRB1, and -DQB1 allele and haplotype frequencies: an analysis of umbilical cord blood units at the Barcelona Cord Blood Bank. *Hla* 94, 347–359. doi: 10.1111/tan.13644

Escobedo-Ruíz, A., Barquera, R., González-Martín, A., Argüelles-San Millán, J. M., Uribe-Duarte, M. G., Hernández-Zaragoza, D. I., et al. (2020). Genetic diversity of HLA system in four populations from Baja California, Mexico: Mexicali, La Paz, Tijuana and rural Baja California. *Hum. Immunol.* 81, 475–477. doi: 10.1016/j.humimm.2019.06.007

Excoffier, L., and Lischer, H. E. (2010). Arlequin suite ver 3.5*: a new series of programs to perform population genetics analyses under Linux and Windows*. *Mol. Ecol. Resour.* 10, 564–567. doi: 10.1111/j.1755-0998.2010.02847.x

Fernández, M., Alarcón, G. S., Calvo-Alén, J., Andrade, R., McGwin, G. Jr., Vilá, L. M., et al. (2007). A multiethnic, multicenter cohort of patients with systemic lupus erythematosus (SLE) as a model for the study of ethnic disparities in SLE. *Arthritis Rheum.* 57, 576–584. doi: 10.1002/art.22672

Furukawa, H., Kawasaki, A., Oka, S., Ito, I., Shimada, K., Sugii, S., et al. (2014). Human leukocyte antigens and systemic lupus erythematosus: a protective role for the HLA-DR6 alleles DRB1*13:02 and *14:03. *PLoS One* 9:e87792. doi: 10.1371/journal.pone.008779

Furuzawa-Carballeda, J., Zuñiga, J., Hernández-Zaragoza, D. I., Barquera, R., Marques-García, E., Jiménez-Alvarez, L., et al. (2018). An original Eurasian haplotype, HLA-DRB1*14:54-DQB1*05:03, influences the susceptibility to idiopathic achalasia. *PLoS One* 13:e0201676. doi: 10.1371/journal.pone.0201676

García-Ortiz, J. E., Sandoval-Ramírez, L., Rangel-Villalobos, H., Maldonado-Torres, H., Cox, S., García-Sepúlveda, C. A., et al. (2006). High-resolution molecular characterization of the HLA class I and class II in the *Tarahumara* Amerindian population. *Tissue Antigens* 68, 135–146. doi: 10.1111/j.1399-0039.2006.00636.x

Ghodke-Puranik, Y., and Niewold, T. B. (2015). Immunogenetics of systemic lupus erythematosus: a comprehensive review. *J. Autoimmun.* 64, 125–136. doi: 10.1016/j.jaut.2015.08.004

Gladman, D. D., Goldsmith, C. H., Urowitz, M. B., Bacon, P., Fortin, P., Ginzler, E., et al. (2000). The systemic lupus international collaborating clinics/American College of Rheumatology (SLICC/ACR) Damage index for systemic lupus erythematosus international comparison. *J. Rheumatol.* 27, 373–376.

Gockel, I., Becker, J., Wouters, M. M., Niebisch, S., Gockel, H. R., Hess, T., et al. (2014). Common variants in the HLA-DQ region confer susceptibility to idiopathic achalasia. *Nat. Genet.* 46, 901–904. doi: 10.1038/ng.3029

Goné-Vázquez, I., Barquera, R., Arellano-Prado, F. P., Hernández-Zaragoza, D. I., Escobedo-Ruíz, A., Clayton, S., et al. (2020). Genetic diversity of HLA system in two populations from Nayarit, Mexico: tepic and rural Nayarit. *Hum. Immunol.* 81, 499–501. doi: 10.1016/j.humimm.2019.06.008

Gonzalez-Galarza, F. F., McCabe, A., Santos, E., Jones, J., Takeshita, L., Ortega-Rivera, N. D., et al. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* 48, D783–D788. doi: 10.1093/nar/gkz1029

González-Medina, L., Barquera, R., Delgado-Aguirre, H., Clayton, S., Adalid-Sáinz, C., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Durango, Mexico: Durango city and rural Durango. *Hum. Immunol.* 81, 489–491. doi: 10.1016/j.humimm.2019.06.005

Gorodezky, C. (2006). "Immunobiology of the Human MHC: vol. 1. International Histocompatibility Workshop and Conference," in *Proceedings of the 13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report: introduction and overview*, ed. J. A. Hansen (Victoria, Ca; Seattle: Int. Histocompatibility Working Group Press).

Gorodezky, C., Najera, R., Rangel, B. E., Castro, L. E., Flores, J., Velázquez, G., et al. (1986). Immunogenetic profile of multiple sclerosis in Mexicans. *Hum. Immunol.* 16, 364–374. doi: 10.1016/0198-8859(86)90063-7

Graham, R. R., Ortmann, W., Rodine, P., Espe, K., Langefeld, C., Lange, E., et al. (2007). Specific combinations of HLA-DR2, and DR3 class II haplotypes contribute graded risk for disease susceptibility, and autoantibodies in human SLE. *Eur. J. Hum. Genet.* 15, 823–830.

Granados, J., Vargas-Alarcón, G., Andrade, F., Melín-Aldana, H., Alcocer-Varela, J., and Alarcón-Segovia, D. (1996). The role of HLA-DR alleles and complotypes through the ethnic barrier in systemic lupus erythematosus in Mexicans. *Lupus* 5, 184–189.

Granados, J., Vargas-Alarcón, G., Drenkard, C., Andrade, F., Melín-Aldana, H., Alcocer-Varela, J., et al. (1997). Relationship of anticardiolipin antibodies and antiphospholipid syndrome to HLA-DR7 in Mexican patients with systemic lupus erythematosus (SLE). *Lupus* 6, 57–62. doi: 10.1177/096120339700600108

Grimaldi, M. C., Crouau-Roy, B., Amoros, J. P., Cambon-Thomsen, A., Carcassi, C., Orru, S., et al. (2001). West Mediterranean islands (Corsica, Balearic islands, Sardinia) and the Basque population: contribution of HLA class I molecular markers to their evolutionary history. *Tissue Antigens* 58, 281–292. doi: 10.1034/j.1399-0039.2001.580501.x

Guo, S. W., and Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* 48, 361–372.

Guzmán, J., Cardiel, M. H., Arce-Salinas, A., Sánchez-Guerrero, J., and Alarcón-Segovia, D. (1992). Measurement of disease activity in systemic lupus erythematosus. *Prospective validation of* 3 clinical indices. *J. Rheumatol.* 19, 1551–1558.

Hernández-Hernández, O., Hernández-Zaragoza, D. I., Barquera, R., Warinner, C., López-Gil, C., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Oaxaca, Mexico: Oaxaca city and rural Oaxaca. *Hum. Immunol.* 81, 553–556. doi: 10.1016/j.humimm.2019.07.278

Hernández-Zaragoza, D. I., Rodríguez-Munguía, T. J., Barquera, R., Adalid-Sáinz, C., Arrieta-Bolaños, E., Clayton, S., et al. (2020). Genetic diversity of HLA system in two populations from San Luis Potosí, Mexico: San Luis Potosí City and rural San Luis Potosí. *Hum. Immunol.* 81, 528–530. doi: 10.1016/j.humimm.2019.07.291

Hochberg, M. C. (1997). Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* 40:1725. doi: 10.1002/art.1780400928

Hollenbach, J. A., Thomson, G., Cao, K., Fernandez-Vina, M., Erlich, H. A., Bugawan, T. L., et al. (2001). HLA diversity, differentiation, and haplotype evolution in Mesoamerican Natives. *Hum. Immunol.* 62, 378–390. doi: 10.1016/s0198-8859(01)00212-9

Infante, E., Olivo, A., Alaez, C., Williams, F., Middleton, D., de la Rosa, G., (1999). Molecular analysis of HLA class I alleles in the Mexican Seri Indians: implications for their origin. *Tissue Antigens.* 54, 35–42. doi: 10.1034/j.1399-0039.1999.540104.x

Izmirly, P. M., Wan, I., Sahl, S., Buyon, J. P., Belmont, H. M., Salmon, J. E., et al. (2017). The incidence and prevalence of systemic lupus erythematosus in New York County (Manhattan), New York: the manhattan lupus surveillance program. *Arthritis Rheumatol.* 69, 2006–2017. doi: 10.1002/art.40192

Juárez-Cedillo, T., Zuñiga, J., Acuña-Alonzo, V., Pérez-Hernández, N., Rodríguez-Pérez, J. M., Barquera, R., et al. (2008). Genetic admixture and diversity estimations in the Mexican Mestizo population from Mexico City using 15 STR polymorphic markers. *Forensic. Sci. Int. Genet.* 2, e37–e39. doi: 10.1016/j.fsigen.2007.08.017

Juárez-Nicolás, F., Barquera, R., Martínez-Álvarez, J. C., Hernández-Zaragoza, D. I., Ortega-Yáñez, A., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in a population from Guerrero, Mexico. *Hum. Immunol.* 81, 550–552. doi: 10.1016/j.humimm.2019.05.015

Kijak, G. H., Walsh, A. M., Koehler, R. N., Moqueet, N., Eller, L. A., Eller, M., et al. (2009). HLA class I allele and haplotype diversity in Ugandans supports the presence of a major east African genetic cluster. *Tissue Antigens* 73, 262–269. doi: 10.1111/j.1399-0039.2008.01192.x

Lebedeva, T. V., Mastromarino, S. A., Lee, E., Ohashi, M., Alosco, S. M., and Yu, N. (2011). Resolution of HLA class I sequence-based typing ambiguities by group-specific sequencing primers. *Tissue Antigens* 77, 247–250. doi: 10.1111/j.1399-0039.2010.01616.x

Louie, L. (2006a). "Immunobiology of the Human MHC: vol. 1. International Histocompatibility Workshop and Conference," in *Proceedings of the 13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report: introduction and overview*, ed. J. A. Hansen (Victoria, Ca; Seattle: Int. Histocompatibility Working Group Press).

Louie L. (2006b). *13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report: HLA genetic differentiation of the 13th IHWC Population Data Relative to Worldwide Linguistic Families*. Victoria, Ca; Seattle: IHWG Press.

Mack, S. J., Meyer, D., Single, R. M., Sanchez-Mazas, A., Thomson, G., and Erlich, H. A. (2006). "13th International Histocompatibility Workshop Anthropology/Human Genetic Diversity Joint Report: introduction and overview," in *Immunobiology of the Human MHC: vol. 1. International Histocompatibility Workshop and Conference*, ed. J. A. Hansen (Seattle: Int. Histocompatibility Working Group Press), 560–563.

Magistrelli, C., Samoilova, E., Agarwal, R. K., Banki, K., Ferrante, P., Vladutiu, A., et al. (1999). Polymorphic genotypes of the HRES-1 human endogenous retrovirus locus correlate with systemic lupus erythematosus and autoreactivity. *Immunogenetics* 49, 829–834. doi: 10.1007/s002510050561

Martínez-Álvarez, J. C., Barquera, R., Hernández-Zaragoza, D. I., Bravo-Acevedo, A., Clayton, S., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Querétaro, Mexico: querétaro city and rural Querétaro. *Hum. Immunol.* 81, 522–524. doi: 10.1016/j.humimm.2019.07.296

Medina-Escobedo, C. E., Barquera, R., Ponnandai-Shanmugavel, K. S., Lara-Riegos, J., Bravo-Acevedo, A., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Quintana Roo, Mexico: cancún and rural Quintana Roo. *Hum. Immunol.* 81, 573–575.

Middleton, D., Williams, F., Hamill, M. A., and Meenagh, A. (2000). Frequency of HLA-B alleles in a Caucasoid population determined by a two-stage PCR-SSOP typing strategy. *Hum. Immunol.* 61, 1285–1297. doi: 10.1016/S0198-8859(00)00186-5

Mišunová, M., Svitálková, T., Pleštilová, L., Kryštufková, O., Tegzová, D., Svobodová, R., et al. (2017). Molecular markers of systemic autoimmune disorders: the expression of MHC-located HSP70 genes is significantly associated with autoimmunity development. *Clin. Exp. Rheumatol.* 35, 33–42.

Modiano, D., Luoni, G., Petrarca, V., Sodiomon Sirima, B., De Luca, M., Simporé, J., et al. (2001). HLA class I in three West African ethnic groups: genetic distances from sub-Saharan and Caucasoid populations. *Tissue Antigens* 57, 128–137. doi: 10.1034/j.1399-0039.2001.057002128.x

Mohd-Yusuf, Y., Phipps, M. E., Chow, S. K., and Yeap, S. S. (2011). HLA-A*11, and novel associations in Malays, and Chinese with systemic lupus erythematosus. *Immunol. Lett.* 139, 68–72.

Molineros, J. E., Looger, L. L., Kim, K., Okada, Y., Terao, C., Sun, C., et al. (2019). Amino acid signatures of HLA Class-I and II molecules are strongly associated with SLE susceptibility and autoantibody production in Eastern Asians. *PLoS Genet.* 15:e1008092. doi: 10.1371/journal.pgen.1008092

Moreno-Estrada, A., Gignoux, C. R., Fernández-López, J. C., Zakharia, F., Sikora, M., Contreras, A. V., et al. (2014). Human genetics. *The genetics of Mexico recapitulates Native American substructure and affects biomedical traits*. *Science* 344, 1280–1285. doi: 10.1126/science.1251688

Morris, D. L., Taylor, K. E., Fernando, M. M., Nititham, J., Alarcón-Riquelme, M. E., Barcellos, L. F., et al. (2012). Unraveling multiple MHC gene associations with systemic lupus erythematosus: model choice indicates a role for HLA alleles and non-HLA genes in Europeans. *Am. J. Hum. Genet.* 91, 778–793. doi: 10.1016/j.ajhg.2012.08.026

Norman, P. J., Hollenbach, J. A., Nemat-Gorgani, N., Guethlein, L. A., Hilton, H. G., Pando, M. J., et al. (2013). Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. *PLoS Genet.* 9:e1003938. doi: 10.1371/journal.pgen.1003938

Ordoñez, G., Romero, S., Orozco, L., Pineda, B., Jiménez-Morales, S., Nieto, A., et al. (2015). Genomewide admixture study in Mexican Mestizos with multiple sclerosis. *Clin. Neurol. Neurosurg.* 130, 55–60. doi: 10.1016/j.clineuro.2014.11.026

Ortega-Yáñez, A., Barquera, R., Curiel-Giles, L., Martínez-Álvarez, J. C., Macías-Medrano, R. M., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Morelos, Mexico: cuernavaca and rural Morelos. *Hum. Immunol.* 81, 557–559. doi: 10.1016/j.humimm.2019.07.289

Pacheco-Ubaldo, H., Adalid-Sáinz, C., Barquera, R., Clayton, S., Arrieta-Bolaños, E., Delgado-Aguirre, H., et al. (2020). Genetic diversity of HLA system in three populations from Chihuahua, Mexico: Chihuahua City, Ciudad Juárez and rural Chihuahua. *Hum. Immunol.* 81, 485–488. doi: 10.1016/j.humimm.2019.05.014

Pantoja-Torres, J. A., Barquera, R., Ballesteros-Romero, M., Bravo-Acevedo, A., Arrieta-Bolaños, E., Montiel-Hernández, G. D., et al. (2020). Genetic diversity of HLA system in three populations from Guanajuato, Mexico: Guanajuato City, León and rural Guanajuato. *Hum. Immunol.* 81, 510–512. doi: 10.1016/j.humimm.2019.06.002

Partanen, J., Koskimies, S., and Johansson, E. (1988). C4 null phenotypes among lupus erythematosus patients are predominantly the result of deletions covering C4 and closely linked 21-hydroxylase A genes. *J. Med. Genet.* 25, 387–391. doi: 10.1136/jmg.25.6.387

Perl, A., Colombo, E., Dai, H., Agarwal, R., Mark, K. A., Banki, K., et al. (1995). Antibody reactivity to the HRES-1 endogenous retroviral element identifies a subset of patients with systemic lupus erythematosus and overlap syndromes. *Correlation with antinuclear antibodies and HLA class II alleles*. *Arthritis Rheum.* 38, 1660–1671. doi: 10.1002/art.1780381119

Picceli, V. F., Skare, T. L., Nisihara, R. M., Nass, F. R., Messias-Reason, I. T., and Utiyama, S. R. (2016). BF*F allotype of the alternative pathway of complement: A marker of protection against the development of antiphospholipid antibodies in patients with systemic lupus erythematosus. *Lupus* 25, 412–417. doi: 10.1177/0961203315615222

Pimtanothai, N., Hurley, C. K., Leke, R., Klitz, W., and Johnson, A. H. (2001). HLA-DR and -DQ polymorphism in Cameroon. *Tissue Antigens* 58, 1–8. doi: 10.1034/j.1399-0039.2001.580101.x

Quaglia, M., Merlotti, G., De Andrea, M., Borgogna, C., and Cantaluppi, V. (2021). Viral Infections and Systemic Lupus Erythematosus: new players in an old story. *Viruses* 13:277. doi: 10.3390/v13020277

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramírez-Bello, J., Cadena-Sandoval, D., Mendoza-Rincón, J. F., Barbosa-Cobos, R. E., Sánchez-Muñoz, F., Amezcua-Guerra, L. M., et al. (2018). Tumor necrosis factor gene polymorphisms are associated with systemic lupus erythematosus susceptibility or lupus nephritis in Mexican patients. *Immunol. Res.* 66, 348–354. doi: 10.1007/s12026-018-8993-8

Rees, F., Doherty, M., Grainge, M. J., Lanyon, P., and Zhang, W. (2017). The worldwide incidence and prevalence of systemic lupus erythematosus: a systematic review of epidemiological studies. *Rheumatology (Oxford)* 56, 1945–1961. doi: 10.1093/rheumatology/kex260

Robinson, J. (2001). IMGT/HLA Database a sequence database for the human major histocompatibility complex. *Nucleic Acids Res.* 29, 210–213. doi: 10.1093/nar/29.1.210

Romero-Hidalgo, S., Flores-Rivera, J., Rivas-Alonso, V., Barquera, R., Villarreal-Molina, M. T., Antuna-Puente, B., et al. (2020). Native American ancestry significantly contributes to neuromyelitis optica susceptibility in the admixed Mexican population. *Sci Rep.* 10:13706. doi: 10.1038/s41598-020-69224-3

Salgado-Galicia, N. A., Hernández-Doño, S., Ruiz-Gómez, D., Jakez-Ocampo, J., Zúñiga, J., Vargas-Alarcón, G., et al. (2020). The role of socioeconomic status in the susceptibility to develop systemic lupus erythematosus in Mexican patients. *Clin. Rheumatol.* 39, 2151–2161. doi: 10.1007/s10067-020-04928-5

Sepúlveda Delgado, J., Lozano Dannis, R., Ocaa-Sibilla, M. J., Ramirez Valdespino, J., Cetina Díaz, J., Bulos Rodríguez, P., et al. (2018). Role of the HLA-DRB1*15 and HLA-DRB1*16 alleles in the genetic susceptibility to develop systemic lupus erythematosus (SLE) after Chikungunya and Zika viruses infection in Mexico. *Blood Genomics* 2, 233–236. doi: 10.46701/APJBG.2018042018127

Single, R. M., Meyer, D., Nunes, K., Francisco, R. S., Hünemeier, T., Maiers, M., et al. (2020). Demographic history and selection at HLA loci in Native Americans. *PLoS One* 15:e0241282. doi: 10.1371/journal.pone.0241282

Solís-Martínez, R., Barquera, R., Ponnandai-Shanmugavel, K. S., Vega-Martínez, M. D. R., Vázquez-Castillo, T. V., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in two populations from Tabasco, Mexico: villahermosa and rural Tabasco. *Hum. Immunol.* 81, 560–562. doi: 10.1016/j.humimm.2019.07.286

Szilágyi, A., Bánlaki, Z., Pozsonyi, E., Yunis, E. J., Awdeh, Z. L., Hossó, A., et al. (2010). Frequent occurrence of conserved extended haplotypes (CEHs) in two Caucasian populations. *Mol. Immunol.* 47, 1899–1904. doi: 10.1016/j.molimm.2010.03.013

Trachtenberg, E., Vinson, M., Hayes, E., Hsu, Y. M., Houtchens, K., Erlich, H., et al. (2007). HLA class I (A, B, C) and class II (DRB1, DQA1, DQB1, DPB1) alleles and haplotypes in the Han from southern China. *Tissue Antigens* 70, 455–463. doi: 10.1111/j.1399-0039.2007.00932.x

Tsokos, G. C., Lo, M. S., Costa Reis, P., and Sullivan, K. E. (2016). New insights into the immunopathogenesis of systemic lupus erythematosus. *Nat. Rev. Rheumatol.* 12, 716–730. doi: 10.1038/nrrheum.2016.186

Uribe-Duarte, M. G., Aguilar-Campos, J. A., Barquera, R., Bravo-Acevedo, A., Clayton, S., Arrieta-Bolaños, E., et al. (2020). Genetic diversity of HLA system in three populations from Sonora, Mexico: ciudad Obregón, Hermosillo and rural Sonora. *Hum. Immunol.* 81, 478–481. doi: 10.1016/j.humimm.2019.05.013

Vargas-Alarcón, G., Hernández-Pacheco, G., Zuñiga, J., Rodríguez-Pérez, J. M., Pérez-Hernández, N., Rangel, C., et al. (2003). Distribution of HLA-B alleles in Mexican Amerindian populations. *Immunogenetics* 54, 756–760. doi: 10.1007/s00251-002-0522-0

Vargas-Alarcón, G., Salgado, N., Granados, J., Gómez-Casado, E., Martinez-Laso, J., Alcocer-Varela, J., et al. (2001). Class II allele and haplotype frequencies in Mexican systemic lupus erythematosus patients: the relevance of considering homologous chromosomes in determining susceptibility. *Hum. Immunol.* 62, 814–820. doi: 10.1016/s0198-8859(01)00267-1

Vasconcelos, C., Carvalho, C., Leal, B., Pereira, C., Bettencourt, A., Costa, P. P., et al. (2009). HLA in Portuguese systemic lupus erythematosus patients and their relation to clinical features. *Ann. N. Y. Acad. Sci.* 1173, 575–580. doi: 10.1111/j.1749-6632.2009.04873.x

Verne, G. N., Hahn, A. B., Pineau, B. C., Hoffman, B. J., Wojciechowski, B. W., and Wu, W. C. (1999). Association of HLA-DR and -DQ alleles with idiopathic achalasia. *Gastroenterology* 117, 26–31. doi: 10.1016/s0016-5085(99)70546-9

Wang, J. (2003). Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 164, 747–765.

Wescott, M. Z., Awdeh, Z. L., Yunis, E. J., and Alper, C. A. (1987). Molecular analysis distinguishes two HLA-DR3-bearing major histocompatibility complex extended haplotypes. *Immunogenetics* 26, 370–374. doi: 10.1007/bf00343707

Williams, F., Meenagh, A., Maxwell, A. P., and Middleton, D. (1999). Allele resolution of HLA-A using oligonucleotide probes in a two-stage typing strategy. *Tissue Antigens* 54, 59–68. doi: 10.1034/j.1399-0039.1999.540107.x

Yunis, E. J., Larsen, C. E., Fernandez-Viña, M., Awdeh, Z. L., Romero, T., Hansen, J. A., et al. (2003). Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks. *Tissue Antigens* 62, 1–20. doi: 10.1034/j.1399-0039.2003.00098.x

Yunis, E., Zuniga, J., Larsen, C., Fernandez-Vina, M., and Granados, J. (2005). *Single Nucleotide Polymorphism blocks and haplotypes: Human MHC blocks diversity. in Encyclopedia of Molecular Cell Biology and Molecular Medicine*, 2nd Edn. Weinheim: Wiley-VCH Verlag GmbH & Co, 191–215.

Zúñiga, J., Vargas-Alarcón, G., Hernández-Pacheco, G., Portal-Celhay, C., Yamamoto-Furusho, J. K., and Granados, J. (2001). Tumor necrosis factor-alpha promoter polymorphisms in Mexican patients with systemic lupus erythematosus (SLE). *Genes Immun.* 2, 363–366. doi: 10.1038/sj.gene.6363793

Zúñiga, J., Yu, N., Barquera, R., Alosco, S., Ohashi, M., Lebedeva, T., et al. (2013). HLA class I and class II conserved extended haplotypes and their fragments or blocks in Mexicans: implications for the study of genetic diversity in admixed populations. *PLoS One* 8:e74442. doi: 10.1371/journal.pone.0074442

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

MINI REVIEW
published: 19 August 2021
doi: 10.3389/fgene.2021.709315

Genetic Susceptibility to Fungal Infections and Links to Human Ancestry

*Bharati Naik, Sumayyah M. Q. Ahmed, Suparna Laha and Shankar Prasad Das**

Yenepoya Research Centre, Yenepoya (Deemed to be University), Mangalore, India

OPEN ACCESS

Edited by:
Elvira Galieva,
Novosibirsk State University, Russia

Reviewed by:
Carina Dehner,
Yale University, United States
Oded Shamriz,
Hadassah-Hebrew University Medical Center, Israel

****Correspondence:***
Shankar Prasad Das
shandas76@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 May 2021*
Accepted: *13 July 2021*
Published: *19 August 2021*

Citation:
Naik B, Ahmed SMQ, Laha S and Das SP (2021) Genetic Susceptibility to Fungal Infections and Links to Human Ancestry. Front. Genet. 12:709315. doi: 10.3389/fgene.2021.709315

Over the ages, fungi have associated with different parts of the human body and established symbiotic associations with their host. They are mostly commensal unless there are certain not so well-defined factors that trigger the conversion to a pathogenic state. Some of the factors that induce such transition can be dependent on the fungal species, environment, immunological status of the individual, and most importantly host genetics. In this review, we discuss the different aspects of how host genetics play a role in fungal infection since mutations in several genes make hosts susceptible to such infections. We evaluate how mutations modulate the key recognition between the pathogen associated molecular patterns (PAMP) and the host pattern recognition receptor (PRR) molecules. We discuss the polymorphisms in the genes of the immune system, the way it contributes toward some common fungal infections, and highlight how the immunological status of the host determines fungal recognition and cross-reactivity of some fungal antigens against human proteins that mimic them. We highlight the importance of single nucleotide polymorphisms (SNPs) that are associated with several of the receptor coding genes and discuss how it affects the signaling cascade post-infection, immune evasion, and autoimmune disorders. As part of personalized medicine, we need the application of next-generation techniques as a feasible option to incorporate an individual's susceptibility toward invasive fungal infections based on predisposing factors. Finally, we discuss the importance of studying genomic ancestry and reveal how genetic differences between the human race are linked to variation in fungal disease susceptibility.

Keywords: genetic predisposition, disease susceptibility, invasive, fungal infection, host genetics, genetic polymorphism, SNP, human ancestry

INTRODUCTION

Fungi are eukaryotic organisms that have a tremendous impact on human health. About 5.1 million fungal species are present on the earth (Hawksworth and Rossman, 1997; Blackwell, 2011). They reproduce asexually by sporulation, budding, and fragmentation. Sexual reproduction involves three phases like plasmogamy, karyogamy, and meiosis. In fungi, hyphae are the main mode of vegetative growth and are collectively called the mycelium. They are usually heterotrophic in nature (Carris et al., 2012) and few are commensal, with the human body acting as a host (Ibrahim and Voelz, 2017). Most of the fungi are adapted to the land environments, and during early episodes of terrestrialization, they had interacted with other organisms having typical parasitic lifestyles (Naranjo-Ortiz and Gabaldón, 2019). Under certain not so well-defined conditions, fungi transform from the non-pathogenic budding yeast to its pathogenic hyphal form, which invades the host tissue (de Pauw, 2011; Underhill and Pearlman, 2015; Kruger et al., 2019; Rai et al., 2021). The fungal species can grow anywhere including plants, animals, and humans. Some enters into our body by inhalation (e.g., *Aspergillus*) and some are commensal (e.g., *Candida*, *Malassezia*) (Underhill and Pearlman, 2015). Commensal like *Malassezia* is more abundant in sebaceous sites of the host. Since they are lipid dependent, they obtain food sources from the host without harming them and colonization starts immediately after birth, when neonatal sebaceous glands are active (Vijaya Chandra et al., 2021). Studies of the microbiome have emerged to be an important area of research, and more importantly, the spotlight is now to understand less studied fungi that have a tremendous influence on the human microbiome especially among immunocompromised individuals. A dysbiotic microbial population is a general characteristic of any fungal infection affecting the mammalian system (Iliev and Leonardi, 2017). Recent reports point toward the role of fungus in pancreatic ductal adenocarcinoma (PDA), a form of human pancreatic cancer caused directly by the presence of budding yeast *Malassezia*, which colonizes the human gut (Aykut et al., 2019). The severity of fungal infection depends on factors such as inoculum load, magnitude of tissue destruction, ability of the fungus to multiply in the tissue, ability to migrate to nearby organs, microenvironment, and immunogenetic status of the host. Resistance to fungi externally is based on cutaneous and mucosal physical barriers and internally by the body's immune molecules and the defensins (Aristizabal and González, 2013; Coates et al., 2018; Salazar and Brown, 2018). Immunosuppression and breakdown of anatomical barriers such as the skin are the major factors behind fungal infections (Kobayashi, 1996). In addition to this, malnutrition, poor hygiene, use of antibiotics, genetic predisposition, environmental factors, and host physiological factors (e.g., oily skin) contribute toward disease progression (**Figure 1**).

Genetic variations play an important role in fungal infection (Pana et al., 2014; Maskarinec et al., 2016; Duxbury et al., 2019; **Table 1**). Recent studies have shown the importance of host genetic variation in influencing the severity and susceptibility to invasive fungal infections (IFIs) (Maskarinec et al., 2016). Increased incidence of opportunistic fungal diseases has been implicated due to gene polymorphism, and genetic errors are frequently observed in immunodeficient phenotypes (Pana et al., 2014). Along with genetic and environmental factors, lifestyle also contributes toward the variation in the genome, as the presence of toxic chemicals and immunosuppressive drugs in an organism's environment leads to altered immune status and inherited deficiencies, which result in susceptibility toward fungal infection (Kumar et al., 2018; **Figure 1**). At the molecular level, epigenetic events like alteration of DNA methylation (a key feature that controls gene expression) (Martin and Fry, 2018), modification in the histones (involved in altered gene expression) (Dolinoy and Jirtle, 2008), and interaction between microbes, genotypes, and environment play a key role in disease progression (Goodrich et al., 2017). Now, challenge for biologists is to identify genetic components that predispose individuals to fungal infection. The study of genes will help to understand the relationship between genetic polymorphism and the cellular phenotype of host and pathogen (Sardinha et al., 2011). Recent research outcomes aided by genomic sequencing point toward an interesting link between genetic predisposition to fungal infections and human ancestry. Single nucleotide polymorphism (SNP) in key immune genes plays an important role in fungal infection affecting particular ancestral populations (Hughes et al., 2008; Dominguez-Andres and Netea, 2019). Thus, with the availability of genetic information, we can study the mechanism behind host defense against the pathogen, susceptibility toward infection (Sardinha et al., 2011), and also have an idea of how the pathogens are evolving and trying to adapt to their host environment through host-pathogen interactions.

GENETIC PREDISPOSITION AND HOST-PATHOGEN INTERACTION

An opportunistic fungus causes diseases mostly in immunocompromised individuals, though normal individuals are also affected (Low and Rotstein, 2011; Eades and Armstrong-James, 2019). Host-pathogen communication initiates through the interactions of the fungal ligands and receptors present on the skin and internal organs (Richmond and Harris, 2014). The better fit of the ligand present on the microbes (against the receptors present on the host), the stronger the interaction (Goyal et al., 2018; Patin et al., 2019). Fungal ligands are a

Abbreviations: PRR, Pattern Recognition Receptor; PAMP, Pathogen Associated Molecular Patterns; TLR, Toll-like Receptor; CLR, C-type Lectin Receptor; NLR, Nod-like Receptor; RLR, Rig-like Receptor; Th cells, Helper T cells; Tc cells, Cytotoxic T cells; Treg cells, Regulatory T cells; ILs, Interleukins; Igs, Immunoglobulins; MBL, Mannose Binding Lectin; *CARD9*, Caspase Recruitment Domain-containing protein 9; CD, Cluster of Differentiation; NET, Neutrophil Extracellular Trap; IFI, Invasive Fungal Infection; *PTX3*, Pentraxin3; *CX3CR1*, C-X3-C Motif Chemokine Receptor 1; Act1, Actin 1; SNPs, Single Nucleotide Polymorphisms; *CYP2C19*, Cytochrome P450 2C19; *ARNT2*, Aryl hydrocarbon Receptor Nuclear Translocator 2; TNFα, Tumor Necrosis Factor-alpha; IFNγ, Interferon-gamma; *MyD88*, Myeloid differentiation primary response 88; *STAT1*, Signal Transducer and Activator of Transcription 1; *STAT3*, Signal Transducer and Activator of Transcription 3; AMP, Anti-Microbial Peptide; APC, Antigen Processing Cell; GWAS, Genome-Wide Association Studies; VNTR, Variable Number Tandem Repeat; Indel, Insertions Deletions; CNV, Copy Number Variation; LOH, Loss of Heterozygosity; MPO, Myeloperoxidase; ROS, Reactive Oxygen Species; CGD, Chronic Granulomatous Disease; CYBB, Cytochrome B beta chain; CYBA, Cytochrome B alpha chain; MASP-2, Mannose-binding lectin-associated serine protease-2; NADPH, Nicotinamide Adenine Dinucleotide Phosphate; NCF, Neutrophil Cytosolic Factor; *CLEC7A*, C-Type Lectin Domain Containing 7A; *NOD2*, Nucleotide-binding Oligomerization Domain containing 2; *RAG*, Recombination Activating Genes; GATA2,GATA-binding factor 2; ZNF341, Zinc Finger Protein 341; IL-12RB1,Interleukin 12 Receptor subunit Beta 1; *AIRE*, Autoimmune Regulator; *RORC*, RAR-related Orphan Receptor C; *DOCK8*, Dedicator of Cytokinesis 8; Tyk2, Tyrosine Kinase 2; CMC, Chronic Mucocutaneous Candidiasis; NLRP3, NOD-, LRR- and Pyrin domain-containing protein 3; PCP, Pneumocystis pneumonia; IPA, Invasive Pulmonary Aspergillosis; HLA-B22, Human Leukocyte Antigen–B22; Nox2, NADPH oxidase 2; PDA, Pancreatic Ductal Adenocarcinoma; IA, Invasive Aspergillosis; HIES, Hyper–Immunoglobulin E Syndrome; RVVC, Recurrent Vulvovaginal Candidiasis; IBD, Inflammatory bowel disease; PIDD, Primary Immunodeficiency disease.

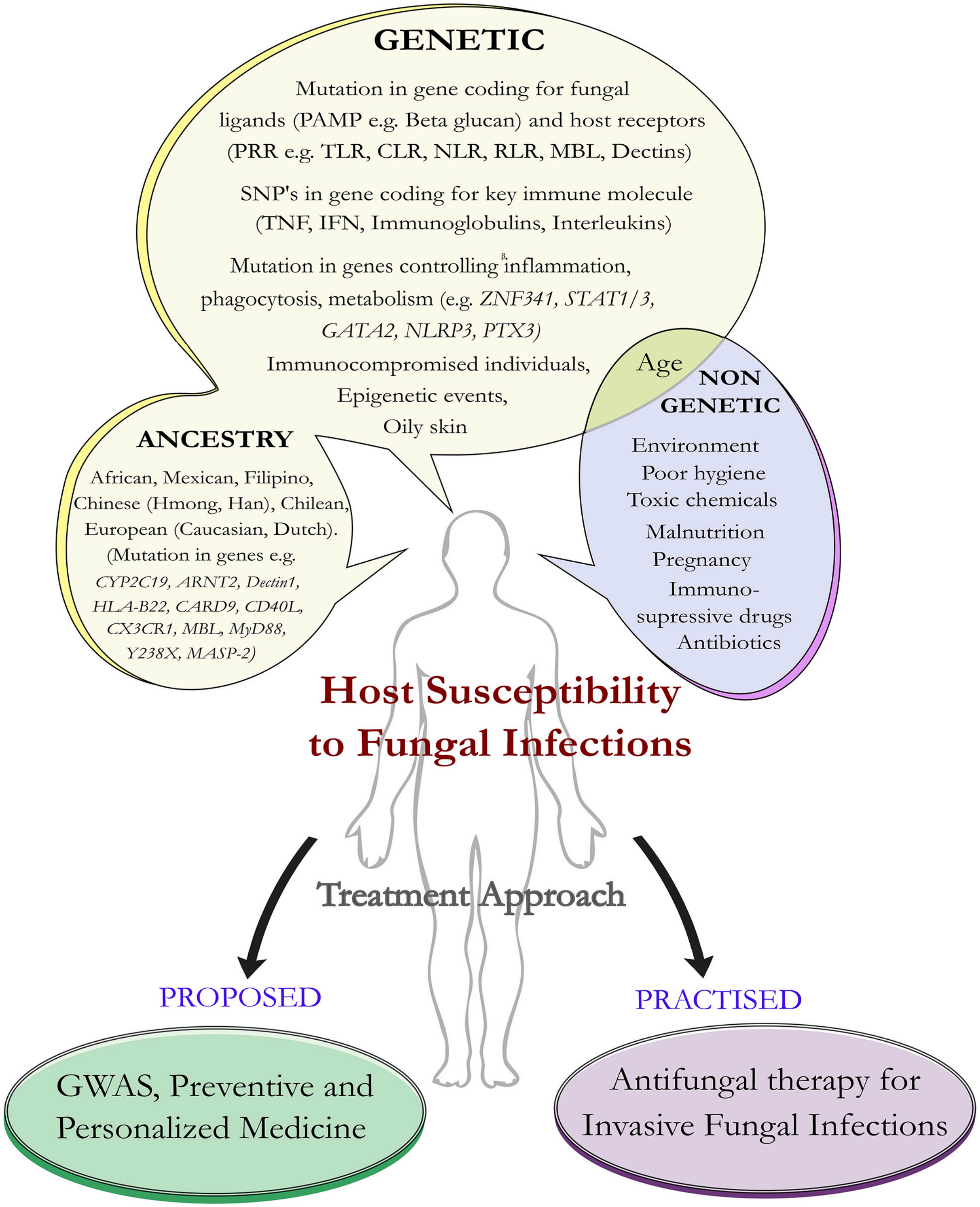

GENETIC
Mutation in gene coding for fungal ligands (PAMP e.g. Beta glucan) and host receptors (PRR e.g. TLR, CLR, NLR, RLR, MBL, Dectins)
SNP's in gene coding for key immune molecule (TNF, IFN, Immunoglobulins, Interleukins)
Mutation in genes controlling inflammation, phagocytosis, metabolism (e.g. ZNF341, STAT1/3, GATA2, NLRP3, PTX3)
Immunocompromised individuals,
Epigenetic events,
Oily skin
Age
ANCESTRY
African, Mexican, Filipino, Chinese (Hmong, Han), Chilean, European (Caucasian, Dutch). (Mutation in genes e.g. CYP2C19, ARNT2, Dectin1, HLA-B22, CARD9, CD40L, CX3CR1, MBL, MyD88, Y238X, MASP-2)
NON GENETIC
Environment
Poor hygiene
Toxic chemicals
Malnutrition
Pregnancy
Immuno-supressive drugs
Antibiotics
Host Susceptibility to Fungal Infections
Treatment Approach
PROPOSED
PRACTISED
GWAS, Preventive and Personalized Medicine
Antifungal therapy for Invasive Fungal Infections

FIGURE 1 | Host susceptibility to invasive fungal infection: predisposing factors and treatment approach. Schematic diagram represents predisposition of the host to certain factors that make them susceptible to fungal infections. Such factors can be genetic as well as non-genetic. Apart from genetic mutations in the host ligand, fungal receptors, and immune genes, human ancestry plays an important role in susceptibility toward invasive fungal infections. The future approaches would be geared toward the investigation (as part of preventive medicine) of the genetic mutations that predispose individuals to fungal infections and offer personalized medicine compared to the more traditional approach that is practiced in the form of antifungal medication.

TABLE 1 | Genetic mutations, human ancestry, and fungal infections.

Immune response		Genes	Ancestry link**	Fungal pathogen	Diseases
Innate immunity	Cell mediated	DOCK8, MyD88**, CARD9**, NCF1, TLR1, MPO, CYBB, CYBA, NADPH oxidase	Chinese (Han) African	*Candida*	Chronic mucocutaneous candidiasis (CMC), chronic granulomatous disease (CGD), candidemia
		PTX3, NCF1, NCF2, NCF4, DOCK8, TLR4, NADPH oxidase, MPO		*Aspergillus*	Invasive aspergillosis (IA)
		CARD9, DOCK8		*Malassezia*	Pityriasis versicolor
		MBL, MASP-2**	Chinese	*Sporothrix*	Sporotrichosis
		MBL**	Chinese	*Pneumocystis jirovecii*	Pneumocystis pneumonia (PCP)
				Candida	Recurrent vulvovaginal candidiasis (RVVC)
		Ferroxidase		Mucorales	Mucormycosis
		HLA-B22**	Mexican	*Histoplasma capsulatum*	Histoplasmosis
	Humoral	IL-17F, Act1, IL-12RB1, IL-17R, IL-17A, IL-17RA, IL-4, IL-12, TyK2, IL-17RC ZNF341, IL-17, IL-22, *Y238X, CLEC7A,* IL-10		*Candida*, *Histoplasma capsulatum*	Chronic mucocutaneous candidiasis (CMC), recurrent vulvovaginal candidiasis (RVVC), histoplasmosis, hyper–immunoglobulin E syndrome (HIES)
		Dectin 1**, IL-10	Chinese (Han) Dutch	*Coccidioides immitis*	Coccidioidomycosis
		Y238X**, IL-10**, TNFα**, IFN-γ**, CLEC7A, CX3CR1**, ARNT2** Asp299Gly, Thr39Ile	European (Dutch, Caucasian)	*Aspergillus*	Invasive pulmonary aspergillosis (IPA)
		rs2243250(IL-4)		*Pneumocystis jirovecii*	Pneumocystis pneumonia (PCP)
		IL-6		*Blastomyces*	Blastomycosis
		IL-2		*Histoplasma*	Histoplasmosis
Adaptive immunity	Cell mediated	*STAT1*		*Histoplasma*	Histoplasmosis
				Coccidiodes	Coccidioidomycosis
		STAT1, STAT3, AIRE, GATA2, RORC, CYP2C19**, RAG1, RAG2	Chinese (Han) Chilean	*Candida*	Candidiasis
		NOD2, STAT3, CYP2C19**	Chinese (Han)	*Aspergillus*	Aspergillosis
		CD40L ** CD50, CD80	Chinese mainland	*Pneumocystis jirovecii, Trichophyton*	Invasive fungal infection (IFI)
	Humoral	IgG, IgA, IgE, IgM, defect in MHC class II molecule		*Pneumocystis jirovecii Candida, Aspergillus, Blastomyces, Coccidioides, Cryptococcus, Histoplasma, Paracoccidioides*	Pneumocystis pneumonia (PCP), candidiasis, aspergillosis, blastomycosis, coccidioidomycosis, cryptococcosis, histoplasmosis, paracoccidioidomycosis.

The symbol ** is used for the genes having the ancestry link.

class of evolutionarily conserved structures called the pathogen associated molecular patterns (PAMPs) and are recognized by receptors present on the host surface called pattern recognition receptors (PRRs). Post internalization, fungi are primarily recognized by the innate cells (e.g., macrophages and dendritic cells) of the immune system (Mogensen, 2009). The main receptors that recognize the fungal-derived PAMPs are Toll-like receptor (TLR like TLR2, TLR4, and TLR9), C-type lectin receptor (CLR like Dectin1 and Dectin2), Nod-like receptor (NLR), Rig-like receptor (RLR), complement receptor, and mannose binding lectin (MBL) (Akira et al., 2006; van de Veerdonk et al., 2008; Hatinguais et al., 2020). These receptors are a crucial component of fungal recognition and trigger an innate immune response.

The host immune response mainly consists of two types, innate and adaptive immunity (Chaplin, 2010; Aristizabal and González, 2013; Netea et al., 2019). Cell-mediated innate immunity is through antigen-presenting cells (APC), which recognize the fungal antigen and process and present it to the T cells. The T cells that participate in antifungal immunity involve

Th (helper T cells) cells, Tc cells (cytotoxic T cells), and Treg (regulatory T cells) cells (Hamad et al., 2018). As soon as the body's immune cells see the foreign fungus, a chain reaction is initiated. Phagocytosis of the fungal pathogen is mediated by neutrophils, macrophages, and dendritic cells, and the oxidative burst kills fungal pathogen by the activity of NADPH oxidase (Rosales and Uribe-Querol, 2017; Warris and Ballou, 2019). The deficiency of this enzyme disrupts the formation of reactive oxygen species (ROS) and makes an individual more susceptible to fungal infection (Hamad et al., 2018). The non-oxidative killing of the fungal pathogen is enhanced by antimicrobial peptides (AMPs) that disrupt the fungal cell wall and also produce neutrophil extracellular traps (NETs) consisting of calprotectin, which induces antifungal activity (Pathakumari et al., 2020; Ulfig and Leichert, 2021). Calprotectin released from NET is an antimicrobial heterodimer that helps in clearing fungus like *Candida*, and its deficiency leads to increased fungal burden (Urban et al., 2009). Innate immune response activates adaptive immunity, which is enhanced by both humoral and cell-mediated immune response, aiding in recognizing fungal antigen, generating inflammation, activating the complement cascade, and further leading to opsonization and neutralization of fungal pathogen (Drummond et al., 2014).

Characterization of single gene defects that predispose individuals to fungal infections needs urgent attention. Monogenic causes for susceptibility of invasive fungal infections have unmasked novel molecules and key signaling pathways in defense against mucosal and systemic antifungal threats (Lionakis et al., 2014; Constantine and Lionakis, 2020). Genetic changes in some key genes play a crucial role in host-pathogen recognition (Kumar et al., 2018; Cunha and Carvalho, 2019; Merkhofer and Klein, 2020). Fungal β-glucan (PAMP) activity can be masked through a change in cell wall components and thus prevent target recognition (Plato et al., 2015). A genetic defect in the different types of PRR families makes the host susceptible to fungal infection (Netea et al., 2012). Defect in the CLR Dectin1, encoded by *CLEC7A* (C-type lectin domain containing 7A) predisposes humans to invasive aspergillosis (IA), chronic mucocutaneous candidiasis (CMC), and recurrent vulvovaginal candidiasis (RVVC) (Reid et al., 2009; Plantinga et al., 2012; Cunha et al., 2018). The *CLEC7A* intronic SNPs rs3901533 and rs7309123 are associated with susceptibility to invasive pulmonary aspergillosis (IPA) in patients with hematologic diseases (Taylor et al., 2007; Sainz et al., 2012). Dectin-1 *Y238X* polymorphism leading to diminished Dectin-1 receptor activity plays a role in RVVC and IA (Plantinga et al., 2009; Cunha et al., 2010; Zahedi et al., 2016). Dectin-1 gene variant also contributes susceptibility to coccidioidomycosis (del Pilar Jiménez-A et al., 2008). Another receptor MBL interacts with pathogens, helps in triggering an immune response, and plays an important role in innate immunity. Deficiency in MBL expression is associated with susceptibility to RVVC (Carvalho et al., 2010) and pneumocystis pneumonia (PCP) (Yanagisawa et al., 2020). Polymorphism in MBL is also associated with chronic cavitary pulmonary aspergillosis and *Candida* infection (Vaid et al., 2007).

SNPs in TLR lead to genetic variation that results in susceptibility to *Candida* and *Aspergillus* infections (Cunha et al., 2010; **Table 1**). Mutation in *TLR1* is associated with candidemia (Ferwerda et al., 2009; Plantinga et al., 2009, 2012). Genetic variation in the PRR TLR4 can also make an individual susceptible to diseases like IPA (Cunha and Carvalho, 2019). Polymorphism in Asp299Gly and Thr399lle present in the *TLR4* impacts hyporesponsiveness to lipopolysaccharide signaling in epithelial cells or alveolar macrophages and results in chronic cavitary pulmonary aspergillosis (Arbour et al., 2000; Carvalho et al., 2008). In addition, polymorphism in immune response *NOD2* (nucleotide binding oligomerization domain containing 2) gene results in IPA. Variation in another receptor type RLR is also associated with *Candida* infection (Gresnigt et al., 2018; Jaeger et al., 2019). Thus, a mutation in the gene coding for a receptor is an important susceptibility factor for CMC and plays a central role in host immune response (Glocker et al., 2009).

GENETIC POLYMORPHISM OF THE IMMUNE SYSTEM LINKED TO FUNGAL INFECTIONS

Genetic variants leading to immunological susceptibility have long been recognized with a few immunodeficiencies characterized by their vulnerability to IFIs (Pana et al., 2014; Maskarinec et al., 2016; Merkhofer and Klein, 2020). Deficiency in *PTX3* (Pentraxin 3), which is involved in innate immunity, leads to susceptibility toward IA (Garlanda et al., 2002). Recently, downregulation of cluster of differentiation molecules CD50 and CD80 has been shown to make an individual susceptible to *Trichophyton* infection (Hamad et al., 2018). Polymorphism in the *CX3CR1* gene (C-X3-C motif chemokine receptor 1, encoding chemokine receptor) is associated with fungal infection in the gut, and it plays an important role in antifungal activity through activation of Th17 cells and IgG antibody response (Kumar et al., 2018). *Candida* infections (ranging from mucosal to bloodstream, including deep-seated infections) are influenced by genetic variants in the human genomes like polymorphism in signal transducer and activator of transcription protein-coding genes *STAT1* and *STAT3* (Plantinga et al., 2012; Smeekens et al., 2013). The important adaptor protein *CARD9* (caspase recruitment domain-containing protein 9) is involved in signal transduction from a variety of receptors, and mutation in this gene not only leads to mucosal infection but also is associated with IFIs, development of autoimmune diseases, inflammatory bowel disease (IBD), and cancer (Glocker et al., 2009; Drummond et al., 2018). *CARD9* plays an important role in Th17 cell differentiation and helps in the release of cytokines (Vautier et al., 2010; Speakman et al., 2020; Vornholz and Ruland, 2020). Recently, defects in *CARD9* and *STAT3* have been found to cause IFI with gastrointestinal manifestations (Vinh, 2019) and mutation in *STAT3* results in reduced Th17 cells causing candidiasis (Engelhardt and Grimbacher, 2012). A heterozygous missense mutation in *STAT1* is associated with coccidioidomycosis and histoplasmosis (Sampaio et al., 2013). Mutation in another transcription factor GATA2 (GATA-binding factor 2) makes patients vulnerable to myeloid malignancy who have a high risk for treatment-associated IFIs involving aspergillosis and candidiasis (Spinner et al., 2014; Donadieu et al., 2018; Vedula et al., 2021). ZNF341 (zinc finger protein 341) is a transcription factor that resides in the nucleus and regulates the activity of *STAT1* and *STAT3* genes. ZNF341-deficient patients lack Th17 cells and have an excess of Th2 cells and low memory B cells. Upon *Candida* infection, individuals with *STAT3* mutation result in hyper–immunoglobulin E syndrome (HIES) associated with defective Th17 cell differentiation and characterized by elevated serum IgE (Béziat et al., 2018; Frey-Jakobs et al., 2018; Egri et al., 2021). Patients with *AIRE* (autoimmune regulator) gene mutations are also susceptible to *Candida albicans* infection and present themselves with autoimmune disorders (Pedroza et al., 2012; de Albuquerque et al., 2018). Genes encoding immune molecules of the adaptive immune system play an important role in controlling fungal invasion (Kawai and Akira, 2007). Immunoglobulins (Igs) IgG, IgA, IgE, and IgM as part of the humoral adaptive immunity mediate protection through direct actions on fungal cells, and classical mechanisms such as phagocytosis and complement activation are affected in case of mutations in genes coding for those Igs (Kaufman, 1985; Lionakis et al., 2014; **Table 1**). MHC class II defects lead to primary immunodeficiency disease (PIDD) and make individuals susceptible to a high rate of fungal infection like Candidiasis and PCP (Lanternier et al., 2013; Abd Elaziz et al., 2020). Mutation in *CARD9* and *DOCK8* (dedicator of cytokinesis 8) among PIDD individuals makes them susceptible to *Malassezia* infection, and deficiency in *STAT3* leads to IPA (Abd Elaziz et al., 2020). Summary of the immune-related genes responsible for susceptibility to fungal infection is highlighted in **Table 1**.

Interleukins (ILs) play a crucial role during fungal infection and help in the maturation of B cells and antibody secretion, which helps fight fungal infections (Antachopoulos and Roilides, 2005; Verma et al., 2015; Sparber and LeibundGut-Landmann, 2019; Griffiths et al., 2021). Mutations in genes encoding for members of the IL-1 family are associated with acute and chronic inflammation and are essential for the innate response to infection (Caffrey et al., 2015; Griffiths et al., 2021). Genetic variation in IL-6 results in blastomycosis (Merkhofer et al., 2019), and defect in IL-10 and IL-6 signaling affects *STAT3*, a key immune response molecule. Genetic variation in IL-10 has also been found to be the underlying cause of susceptibility toward fungal infections like IA (Zaas, 2006). IL-10 mutation makes an individual susceptible to *Candida* and *Coccidiodes immitis* infection (Fierer, 2006), and IL-4 polymorphism resulted in susceptibility toward *Candida* infection (Babula et al., 2005; Choi et al., 2005). SNP in rs2243250, known to influence IL-4 production, is associated with susceptibility to PCP in HIV-positive patients (Wójtowicz et al., 2019). In addition, deficiency of interleukin IL-17 and IL-22 production as a result of genetic mutation has been reported to be the cause of RVVC (Sobel, 2016). IL-2 mutation too predisposes individuals to invasive fungal infection like histoplasmosis by affecting T cell functions (Smeekens et al., 2013; Lionakis et al., 2014; Kumaresan et al., 2017; Pathakumari et al., 2020). The emerging role of the IL-12 family of cytokines in the fight against candidiasis has been reported (Ashman et al., 2011; Thompson and Orr, 2018). IL-12RB1 (interleukin 12 receptor subunit beta 1) impairs the development of human IL-17 producing T cells (Huppler et al., 2012; Johnson et al., 2012; Thompson and Orr, 2018), and mutations inherited might be responsible for histoplasmosis (León-Lara et al., 2020). RAR-related orphan receptor C (*RORC*) encoding transcription factors play an integral role in both IL-17 and IFNγ pathways in CMC (De Luca et al., 2007; Constantine and Lionakis, 2020). Deficiency of tyrosine kinase 2 (Tyk2) that participates in signal transduction for various cytokine receptors leads to impaired helper T cell type 1 (Th1) differentiation and accelerated helper T cell type 2 (Th2) differentiation in candidiasis (Minegishi et al., 2006). Mutation in the main inflammasome NLRP3 (NOD-, LRR-, and pyrin domain-containing protein 3), associated with fungal infection, leads to susceptibility toward RVVC or IPA (Kasper et al., 2018; Wang et al., 2020; Briard et al., 2021). Also, mutations in key recombination activating genes (*RAG1* and *RAG2*) lead to loss of T and B cells, making individuals susceptible to CMC and a broad spectrum of pathogens (Schuetz et al., 2008; Delmonte et al., 2018). Genetic polymorphism in the IL-17 family genes, which encode for the Th17 cellular differentiation, results in an individual's susceptibility toward fungal infection (Hamad et al., 2018). One of the key signaling molecule pathways, the IL-17R signaling is dependent on Act1 (Actin1—a conserved protein that helps in key cellular processes), and mutation in the gene coding for Actin1 leads to defect in IL-17R signaling pathway, which ultimately fails to provide immunity against CMC (Boisson et al., 2013). IL-17RA binds to homo- and heterodimers of IL-17A and IL-17F, and its deficiency or genetic mutation in any of the gene coding for receptors IL-17RA or IL-17RC leads to CMC (Puel et al., 2011; Sawada et al., 2021).

Mutation in *DOCK8* characterized by elevated IgE level is also known to be responsible for recurrent fungal infections like IA and mucocutaneous candidiasis (Biggs et al., 2017; Nahum, 2017). During *Aspergillus* infection, tumor necrosis factor-alpha (TNFα) enhances the phagocytic activity and the polymorphic site in TNF promotor predisposes individuals to IA (Roilides et al., 1998; Sainz et al., 2007). Neutrophil cytosolic factors (NCFs) are part of the group of proteins that form the enzyme complex called NADPH oxidase, and mutation in any of the key genes *NCF1*, *NCF2*, and *NCF4* leads to impaired fungal eradication (as in aspergillosis) due to non-functional NADPH oxidase (Panday et al., 2015; Giardino et al., 2017; Dinauer, 2019; Wu et al., 2019). Decreased myeloperoxidase (MPO) activity (inability to produce hypochlorous acid) in neutrophils leads to delayed killing of pathogen and makes an individual susceptible to invasive *Candida* infection (Aratani et al., 1999; Merkhofer and Klein, 2020). Myeloperoxidase mutants lead to impaired ROS production, making the host susceptible to infection, and thus, both MPO and NADPH oxidase mutants are unable to eradicate fungal threats like chronic granulomatous disease (CGD) and IA (Lehrer and Cline, 1969; Aratani et al., 2004; Segal and Romani, 2009; Ren et al., 2012). Cytochrome b -245 is a primary component of the microbicidal oxidase system of phagocytes encoded by the alpha and beta chains *CYBA* and *CYBB*/Nox2 (NADPH oxidase 2), respectively (Stasia, 2016), and cytochrome b deficiency is also linked to CGD (Clark, 1999; Stasia et al., 2003; Kutukculer et al., 2019). Recently, it has been reported that mutants in the ferroxidase gene make individuals susceptible to mucormycosis (Navarro-Mendoza et al., 2018), an infection that has been affecting COVID-19 patients (Raut and Huy, 2021). Thus, mutations of key genes of the immune system play an important role in fungal resistance, and interestingly, genetic polymorphisms in these genes have revealed some links with human ancestry.

HUMAN ANCESTRY AND GENETIC PREDISPOSITION TO FUNGAL INFECTIONS

There is limited research investigating the link between genetic polymorphism in key immune genes, human ancestry, and susceptibility toward fungal infection (**Figure 1**). But recent research outcomes aided by genomic sequencing point toward an interesting fact. Infection with the fungus *Coccidioides immitis* among Filipino ancestry was found to be common among men and non-white persons causing coccidioidomycosis (van Burik and Magee, 2001). Studies on DNA, which provides genetic information transferred from ancestors to their family members and relatives, indicate that the Hmong ancestry are more susceptible to fungal infection (Xiong et al., 2013). In another report, genetic differentiation among the Hmong ancestry originating from Wisconsin makes them more susceptible to blastomycosis. The Chinese Han population was found to suffer due to poor metabolism as a result of the *CYP2C19* gene (cytochrome P450 2C19) polymorphism involved in the metabolism of xenobiotics. This is one of the direct evidence to prove the role played by genetic polymorphisms in IFIs among a particular human race. Interestingly, polymorphism in the *CYP2C19* allele (because of the presence of variant rs12248560) has been reported to cause aspergillosis among the Chileans (Espinoza et al., 2019). Similarly, deficiency as a result of a mutation in the gene coding for CD40L (binds to CD40 cells and plays role in B cell proliferation) influences susceptibility to PCP among people belonging to the Chinese mainland (Du et al., 2019). It was also reported that genetic variation in *CARD9* led to increased susceptibility toward *Candida* infections in the African population (Rosentul et al., 2012).

SNP plays an important role in fungal infection affecting particular ancestral populations (Hughes et al., 2008; Dominguez-Andres and Netea, 2019). SNPs in genes like *ARNT2* (aryl hydrocarbon receptor nuclear translocator 2) and *CX3CR1* are responsible for cytokine activation, and polymorphism in these genes has been found to play an important role in the invasiveness of aspergillosis infection among European ancestry (Lupiañez et al., 2020). Variations in the PRR MBL and mannose-binding lectin-associated serine protease-2 (MASP-2) proteins were shown to be responsible for sporotrichosis in the Chinese population. It was observed that individuals with elevated levels of the protein are more susceptible to *Sporothrix* infection (Bao et al., 2019). Another importance of SNP is associated with the varying protein expression levels associated with autoimmune diseases (Lionakis, 2012; Jonkers and Wijmenga, 2017). SNPs in cytokine coding genes influence the low production of TNFα, IFNγ, and IL-10, and it was observed that these variations make the Caucasian population susceptible to fungal infections (Larcombe et al., 2005). In a recent study, genetic variant of the key immune adapter MyD88 (myeloid differentiation factor 88) in the Chinese Han population was found to be associated with higher fungal infection and it was shown that the defect in Dectin1 was the primary cause (Chen et al., 2019). Susceptibility to candidiasis and IPA as a result of a defect in Dectin1 was observed in the Dutch family (Ferwerda et al., 2009; Chai et al., 2011). In addition, susceptibility to histoplasmosis as a result of the human leukocyte antigen B22 (*HLA-B22*) variant was reported in the Mexican population (Taylor et al., 1997). The human race thus plays a crucial role in fungal invasion as seen among white transplant recipients who are more susceptible compared to black recipients due to differences in their pharmacogenetics (Boehme et al., 2014). All the above studies show direct links of human ancestry to fungal diseases and indicate how genetic mutations among the human race make them predisposed to certain fungal infections (**Table 1**).

DISCUSSION

Fungi play an important role in the human microbiome (Huseyin et al., 2017; Perez et al., 2021; Tiew et al., 2021). In this review, we have focused on genetic predisposition to human fungal infections and discussed the link that exists between ancestry and susceptibility to IFIs. Among those fungi that are commensal with the warm-blooded host, few turn pathogenic under not so well-defined conditions (Hall and Noverr, 2017; Jacobsen, 2019; Limon et al., 2017). Such conversion to pathogenic forms is aided by external factors like environment, immunological status, and most importantly host genetics (Kobayashi, 1996; Kumar et al., 2018; **Figure 1**). As we learn more about fungal biology, we also understand genetic signatures in the host that make them prone to fungal infections. This is explained by the term genetic predisposition, and external players like the environment also play a role in triggering an autoimmune, inflammatory, or allergic reaction to fungal infections (**Figure 1**). Identification of fungal allergens has become challenging because most of the allergens mimic immune molecules (Pfavayi et al., 2020). We have seen how mutations in key recognition molecules (**Table 1**) play a trigger for several fungal infections. We looked into variations introduced by SNPs that are present in the immune response genes (**Table 1**) critical for fungal infections. The polymorphism in the immune genes (*PTX3, CX3CR1, CARD9*, *STAT3*, and others, **Figure 1**) make the host susceptible (Garlanda et al., 2002; Kumar et al., 2018; Vinh, 2019), and defect in interleukins (e.g., IL-4, IL-10) leads to genetic predisposition toward fungal infection (Babula et al., 2005; Choi et al., 2005; Zaas, 2006; **Table 1**). The study of these genes helps us to understand the relationship between genetic polymorphism and the cellular phenotype of host, pathogen, and associated defense mechanisms (Sardinha et al., 2011). Thus, the composition of both host and pathogen plays important role in disease progression, and the challenge is to identify the genetic components involved in pathogenesis.

A few studies point toward a link between human ancestry and genetic predisposition to fungal infections (van Burik and Magee, 2001; Ferwerda et al., 2009; Xiong et al., 2013; Chen et al., 2019; Du et al., 2019; Espinoza et al., 2019; **Table 1**). Mutations in several components of the immune system make certain human ancestral descendants more prone to fungal infections. Few studies have looked into genetic associations and human ancestry. This aspect is an important and emerging research area in terms of population genetics (Hirschhorn et al., 2002; Gnat et al., 2021). Mutation in key genes relating to the immune system of the host makes certain ancestral descendants susceptible to fungal infections as we observe in the case of certain European, African, and Caucasian individuals (Larcombe et al., 2005; Kwizera et al., 2019; Pfavayi et al., 2020), making them more susceptible to emerging fungal pathogens (**Figure 1**). Such fungi are a threat to global public health and can colonize the skin, spread from person to person, and cause many high-risk diseases (Lamoth and Kontoyiannis, 2018). To deal with such organisms, we require better surveillance methods, rapid and accurate diagnostics, and decolonization protocols that include administration of antimicrobial or antiseptic agents and new antifungal drugs (Jeffery-Smith et al., 2018; Jackson et al., 2019; Chowdhary et al., 2020; Fisher et al., 2020; Steenwyk et al., 2020). Genome-wide association studies (GWAS) would help us to evaluate the difference in the DNA sequences and understand heritability, disease risk, and susceptibility to antifungals (Bloom et al., 2019; Guo et al., 2020; **Figure 1**). From genome sequencing, genomic variations like SNPs, variable number tandem repeats (VNTRs), and insertion/deletions (Indels) can be identified. Structural genome variations like aneuploidy and copy number variations (CNVs) also provide important clues to fungal virulence (Tsai and Nelliat, 2019). During fungal microevolution, many of these events like insertion/deletion of genes, loss of heterozygosity (LOH), and genome plasticity help fungus to adapt against antifungal drugs and harsh host environment (Beekman and Ene, 2020). Thus, as part of preventive medicine, a better understanding of host genetics behind fungal infection will help us to study infectious diseases through modern genomic approaches and offer personalized therapy against invasive fungal diseases.

AUTHOR CONTRIBUTIONS

SD conceptualized, reviewed, and approved the manuscript. BN drafted the manuscript, revised the article critically, and provided critical suggestions. SA contributed toward artwork and reviewed the manuscript. SL provided critical review and revised intellectual content. All authors contributed to the article and approved the submitted version.

FUNDING

SD was supported by the Grant in Aid program of DHR GIA/2019/000620/PRCGIA. BN is a JRF supported by Yenepoya (Deemed to be University). SA was supported by an ICMR grant to SD.

ACKNOWLEDGMENTS

We would like to acknowledge funding from the Department of Health Research (DHR), the Indian Council of Medical Research (ICMR) New Delhi for funding our research program. We also thank Yenepoya Research Centre for providing laboratory facilities.

REFERENCES

Abd Elaziz, D., Abd El-Ghany, M., Meshaal, S., El Hawary, R., Lotfy, S., Galal, N., et al. (2020). Fungal infections in primary immunodeficiency diseases. *Clin. Immunol. Orlando Fla* 219:108553. doi: 10.1016/j.clim.2020.108553

Akira, S., Uematsu, S., and Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell* 124, 783–801. doi: 10.1016/j.cell.2006.02.015

Antachopoulos, C., and Roilides, E. (2005). Cytokines and fungal infections. *Br. J. Haematol.* 129, 583–596. doi: 10.1111/j.1365-2141.2005.05498.x

Aratani, Y., Koyama, H., Nyui, S., Suzuki, K., Kura, F., and Maeda, N. (1999). Severe impairment in early host defense against *Candida albicans* in mice deficient in myeloperoxidase. *Infect. Immun.* 67, 1828–1836. doi: 10.1128/IAI.67.4.1828-1836.1999

Aratani, Y., Kura, F., Watanabe, H., Akagawa, H., Takano, Y., Suzuki, K., et al. (2004). In vivo role of myeloperoxidase for the host defense. *Jpn. J. Infect. Dis.* 57:S15.

Arbour, N. C., Lorenz, E., Schutte, B. C., Zabner, J., Kline, J. N., Jones, M., et al. (2000). TLR4 mutations are associated with endotoxin hyporesponsiveness in humans. *Nat. Genet.* 25, 187–191. doi: 10.1038/76048

Aristizabal, B., and González, Á (2013). "Innate immune system," in *Autoimmunity: From Bench to Bedside*, eds J. M. Anaya, Y. Shoenfeld, A. Rojas-Villarraga, R. A. Levy, and R. Cervera (Bogota: El Rosario University Press).

Ashman, R. B., Vijayan, D., and Wells, C. A. (2011). IL-12 and related cytokines: function and regulatory implications in *Candida albicans* infection. *Clin. Dev. Immunol.* 2011:686597. doi: 10.1155/2011/686597

Aykut, B., Pushalkar, S., Chen, R., Li, Q., Abengozar, R., Kim, J. I., et al. (2019). The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* 574, 264–267. doi: 10.1038/s41586-019-1608-2

Babula, O., Lazdâne, G., Kroica, J., Linhares, I. M., Ledger, W. J., and Witkin, S. S. (2005). Frequency of interleukin-4 (IL-4) -589 gene polymorphism and vaginal concentrations of IL-4, nitric oxide, and mannose-binding lectin in women with recurrent vulvovaginal candidiasis. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* 40, 1258–1262. doi: 10.1086/429246

Bao, F., Fu, X., Yu, G., Wang, Z., Liu, H., and Zhang, F. (2019). Mannose-Binding lectin and mannose-binding lectin-associated serine protease-2 genotypes and serum levels in patients with sporotrichosis. *Am. J. Trop. Med. Hyg.* 101, 1322–1324. doi: 10.4269/ajtmh.19-0470

Beekman, C. N., and Ene, I. V. (2020). Short-term evolution strategies for host adaptation and drug escape in human fungal pathogens. *PLoS Pathog.* 16:e1008519. doi: 10.1371/journal.ppat.1008519

Béziat, V., Li, J., Lin, J.-X., Ma, C. S., Li, P., Bousfiha, A., et al. (2018). A recessive form of hyper-IgE syndrome by disruption of ZNF341-dependent STAT3 transcription and activity. *Sci. Immunol.* 3:eaat4956. doi: 10.1126/sciimmunol.aat4956

Biggs, C. M., Keles, S., and Chatila, T. A. (2017). DOCK8 deficiency: insights into pathophysiology, clinical features and management. *Clin. Immunol. Orlando Fla* 181, 75–82. doi: 10.1016/j.clim.2017.06.003

Blackwell, M. (2011). The fungi: 1, 2, 3 . 5.1 million species? *Am. J. Bot.* 98, 426–438. doi: 10.3732/ajb.1000298

Bloom, J. S., Boocock, J., Treusch, S., Sadhu, M. J., Day, L., Oates-Barker, H., et al. (2019). Rare variants contribute disproportionately to quantitative trait variation in yeast. *ELife* 8:e49212. doi: 10.7554/eLife.49212

Boehme, A. K., McGwin, G., Andes, D. R., Lyon, G. M., Chiller, T., Pappas, P. G., et al. (2014). Race and invasive fungal infection in solid organ transplant recipients. *Ethn. Dis.* 24, 382–385.

Boisson, B., Wang, C., Pedergnana, V., Wu, L., Cypowyj, S., Rybojad, M., et al. (2013). An ACT1 mutation selectively abolishes interleukin-17 responses in humans with chronic mucocutaneous candidiasis. *Immunity* 39, 676–686. doi: 10.1016/j.immuni.2013.09.002

Briard, B., Malireddi, R. K. S., and Kanneganti, T.-D. (2021). Role of inflammasomes/pyroptosis and PANoptosis during fungal infection. *PLoS Pathog.* 17:e1009358. doi: 10.1371/journal.ppat.1009358

Caffrey, A. K., Lehmann, M. M., Zickovich, J. M., Espinosa, V., Shepardson, K. M., Watschke, C. P., et al. (2015). IL-1α signaling is critical for leukocyte recruitment after pulmonary *Aspergillus fumigatus* challenge. *PLoS Pathog.* 11:e1004625. doi: 10.1371/journal.ppat.1004625

Carris, L. M., Little, C. R., and Stiles, C. M. (2012). *Introduction to Fungi. The Plant Health Instructor*. doi: 10.1094/PHI-I-2012-0426-01

Carvalho, A., Cunha, C., Pasqualotto, A. C., Pitzurra, L., Denning, D. W., and Romani, L. (2010). Genetic variability of innate immunity impacts human susceptibility to fungal diseases. *Int. J. Infect. Dis. IJID Off. Publ. Int. Soc. Infect. Dis.* 14, e460–e468. doi: 10.1016/j.ijid.2009.06.028

Carvalho, A., Pasqualotto, A. C., Pitzurra, L., Romani, L., Denning, D. W., and Rodrigues, F. (2008). Polymorphisms in toll-like receptor genes and susceptibility to pulmonary aspergillosis. *J. Infect. Dis.* 197, 618–621. doi: 10.1086/526500

Chai, L. Y. A., de Boer, M. G. J., van der Velden, W. J. F. M., Plantinga, T. S., van Spriel, A. B., Jacobs, C., et al. (2011). The Y238X stop codon polymorphism in the human β-glucan receptor dectin-1 and susceptibility to invasive aspergillosis. *J. Infect. Dis.* 203, 736–743. doi: 10.1093/infdis/jiq102

Chaplin, D. D. (2010). Overview of the immune response. *J. Allergy Clin. Immunol.* 125, S3–S23. doi: 10.1016/j.jaci.2009.12.980

Chen, M.-J., Hu, R., Jiang, X.-Y., Wu, Y., He, Z.-P., Chen, J.-Y., et al. (2019). Dectin-1 rs3901533 and rs7309123 polymorphisms increase susceptibility to pulmonary invasive fungal disease in patients with acute myeloid leukemia from a Chinese Han population. *Curr. Med. Sci.* 39, 906–912. doi: 10.1007/s11596-019-2122-3

Choi, P., Xanthaki, D., Rose, S. J., Haywood, M., Reiser, H., and Morley, B. J. (2005). Linkage analysis of the genetic determinants of T-cell IL-4 secretion, and identification of Flj20274 as a putative candidate gene. *Genes Immun.* 6, 290–297. doi: 10.1038/sj.gene.6364192

Chowdhary, A., Tarai, B., Singh, A., and Sharma, A. (2020). Multidrug-Resistant *Candida auris* infections in critically Ill coronavirus disease patients, India, April-July 2020. *Emerg. Infect. Dis.* 26, 2694–2696. doi: 10.3201/eid2611.203504

Clark, R. A. (1999). Activation of the neutrophil respiratory burst oxidase. *J. Infect. Dis.* 179, S309–S317. doi: 10.1086/513849

Coates, M., Blanchard, S., and MacLeod, A. S. (2018). Innate antimicrobial immunity in the skin: a protective barrier against bacteria, viruses, and fungi. *PLoS Pathog.* 14:e1007353. doi: 10.1371/journal.ppat.1007353

Constantine, G. M., and Lionakis, M. S. (2020). Recent advances in understanding inherited deficiencies in immunity to infections. *F1000Res.* 9:F1000 Faculty Rev-243. doi: 10.12688/f1000research.22036.1

Cunha, C., and Carvalho, A. (2019). Genetic defects in fungal recognition and susceptibility to invasive pulmonary aspergillosis. *Med. Mycol.* 57, S211–S218. doi: 10.1093/mmy/myy057

Cunha, C., Di Ianni, M., Bozza, S., Giovannini, G., Zagarella, S., Zelante, T., et al. (2010). Dectin-1 Y238X polymorphism associates with susceptibility to invasive aspergillosis in hematopoietic transplantation through impairment of both recipient- and donor-dependent mechanisms of antifungal immunity. *Blood* 116, 5394–5402. doi: 10.1182/blood-2010-04-279307

Cunha, D., de, O., Leão-Cordeiro, J. A. B., Paula, H. D. S. C., Ataides, F. S., Saddi, V. A., et al. (2018). Association between polymorphisms in the genes encoding toll-like receptors and dectin-1 and susceptibility to invasive aspergillosis: a systematic review. *Rev. Soc. Bras. Med. Trop.* 51, 725–730. doi: 10.1590/0037-8682-0314-2018

de Albuquerque, J. A. T., Banerjee, P. P., Castoldi, A., Ma, R., Zurro, N. B., Ynoue, L. H., et al. (2018). The role of AIRE in the immunity against *Candida albicans* in a model of human macrophages. *Front. Immunol.* 9:567. doi: 10.3389/fimmu.2018.00567

De Luca, A., Montagnoli, C., Zelante, T., Bonifazi, P., Bozza, S., Moretti, S., et al. (2007). Functional yet balanced reactivity to *Candida albicans* requires TRIF, MyD88, and IDO-dependent inhibition of Rorc. *J. Immunol. Baltim. Md 1950* 179, 5999–6008. doi: 10.4049/jimmunol.179.9.5999

de Pauw, B. E. (2011). What are fungal infections? *Mediterr. J. Hematol. Infect. Dis.* 3:e2011001. doi: 10.4084/MJHID.2011.001

del Pilar Jiménez-A, M., Viriyakosol, S., Walls, L., Datta, S. K., Kirkland, T., Heinsbroek, S. E. M., et al. (2008). Susceptibility to *Coccidioides* species in C57BL/6 mice is associated with expression of a truncated splice variant of Dectin-1 (Clec7a). *Genes Immun.* 9, 338–348. doi: 10.1038/gene.2008.23

Delmonte, O. M., Schuetz, C., and Notarangelo, L. D. (2018). RAG deficiency: two genes, many diseases. *J. Clin. Immunol.* 38, 646–655. doi: 10.1007/s10875-018-0537-4

Dinauer, M. C. (2019). Insights into the NOX NADPH oxidases using heterologous whole cell assays. *Methods Mol. Biol. Clifton NJ* 1982, 139–151. doi: 10.1007/978-1-4939-9424-3_9

Dolinoy, D. C., and Jirtle, R. L. (2008). Environmental epigenomics in human health and disease. *Environ. Mol. Mutagen.* 49, 4–8. doi: 10.1002/em.20366

Dominguez-Andres, J., and Netea, M. G. (2019). Impact of historic migrations and evolutionary processes on human immunity. *Trends Immunol.* 40, 1105–1119. doi: 10.1016/j.it.2019.10.001

Donadieu, J., Lamant, M., Fieschi, C., de Fontbrune, F. S., Caye, A., Ouachee, M., et al. (2018). Natural history of GATA2 deficiency in a survey of 79 French and Belgian patients. *Haematologica* 103, 1278–1287. doi: 10.3324/haematol.2017.181909

Drummond, R. A., Franco, L. M., and Lionakis, M. S. (2018). Human CARD9: a critical molecule of fungal immune surveillance. *Front. Immunol.* 9:1836. doi: 10.3389/fimmu.2018.01836

Drummond, R. A., Gaffen, S. L., Hise, A. G., and Brown, G. D. (2014). Innate defense against fungal pathogens. *Cold Spring Harb. Perspect. Med.* 5:a019620. doi: 10.1101/cshperspect.a019620

Du, X., Tang, W., Chen, X., Zeng, T., Wang, Y., Chen, Z., et al. (2019). Clinical, genetic and immunological characteristics of 40 Chinese patients with CD40 ligand deficiency. *Scand. J. Immunol.* 90:e12798. doi: 10.1111/sji.12798

Duxbury, E. M., Day, J. P., Maria Vespasiani, D., Thüringer, Y., Tolosana, I., Smith, S. C., et al. (2019). Host-pathogen coevolution increases genetic variation in susceptibility to infection. *ELife* 8:e46440. doi: 10.7554/eLife.46440

Eades, C. P., and Armstrong-James, D. P. H. (2019). Invasive fungal infections in the immunocompromised host: mechanistic insights in an era of changing immunotherapeutics. *Med. Mycol.* 57, S307–S317. doi: 10.1093/mmy/myy136

Egri, N., Esteve-Solé, A., Deyà-Martínez, À, Ortiz de Landazuri, I., Vlagea, A., García, A. P., et al. (2021). Primary immunodeficiency and chronic mucocutaneous candidiasis: pathophysiological, diagnostic, and therapeutic approaches. *Allergol. Immunopathol. (Madr.)* 49, 118–127. doi: 10.15586/aei.v49i1.20

Engelhardt, K. R., and Grimbacher, B. (2012). Mendelian traits causing susceptibility to mucocutaneous fungal infections in human subjects. *J. Allergy Clin. Immunol.* 129, 294–305; quiz 306–307. doi: 10.1016/j.jaci.2011.12.966

Espinoza, N., Galdames, J., Navea, D., Farfán, M. J., and Salas, C. (2019). Frequency of the CYP2C19*17 polymorphism in a Chilean population and its effect on voriconazole plasma concentration in immunocompromised children. *Sci. Rep.* 9:8863. doi: 10.1038/s41598-019-45345-2

Ferwerda, B., Ferwerda, G., Plantinga, T. S., Willment, J. A., van Spriel, A. B., Venselaar, H., et al. (2009). Human dectin-1 deficiency and mucocutaneous fungal infections. *N. Engl. J. Med.* 361, 1760–1767. doi: 10.1056/NEJMoa0901053

Fierer, J. (2006). IL-10 and susceptibility to *Coccidioides* immitis infection. *Trends Microbiol.* 14, 426–427. doi: 10.1016/j.tim.2006.07.009

Fisher, M. C., Gurr, S. J., Cuomo, C. A., Blehert, D. S., Jin, H., Stukenbrock, E. H., et al. (2020). Threats posed by the fungal kingdom to humans, wildlife, and agriculture. *mBio* 11:e00449-20. doi: 10.1128/mBio.00449-20

Frey-Jakobs, S., Hartberger, J. M., Fliegauf, M., Bossen, C., Wehmeyer, M. L., Neubauer, J. C., et al. (2018). ZNF341 controls STAT3 expression and thereby immunocompetence. *Sci. Immunol.* 3:eaat4941. doi: 10.1126/sciimmunol.aat4941

Garlanda, C., Hirsch, E., Bozza, S., Salustri, A., De Acetis, M., Nota, R., et al. (2002). Non-redundant role of the long pentraxin PTX3 in anti-fungal innate immune response. *Nature* 420, 182–186. doi: 10.1038/nature01195

Giardino, G., Cicalese, M. P., Delmonte, O., Migliavacca, M., Palterer, B., Loffredo, L., et al. (2017). NADPH oxidase deficiency: a multisystem approach. *Oxid. Med. Cell. Longev.* 2017:4590127. doi: 10.1155/2017/4590127

Glocker, E.-O., Hennigs, A., Nabavi, M., Schäffer, A. A., Woellner, C., Salzer, U., et al. (2009). A homozygous CARD9 mutation in a family with susceptibility to fungal infections. *N. Engl. J. Med.* 361, 1727–1735. doi: 10.1056/NEJMoa0810719

Gnat, S., Łagowski, D., and Nowakiewicz, A. (2021). Genetic predisposition and its heredity in the context of increased prevalence of dermatophytoses. *Mycopathologia* 186, 163–176. doi: 10.1007/s11046-021-00529-1

Goodrich, J. K., Davenport, E. R., Clark, A. G., and Ley, R. E. (2017). The relationship between the human genome and microbiome comes into view. *Annu. Rev. Genet.* 51, 413–433. doi: 10.1146/annurev-genet-110711-155532

Goyal, S., Castrillón-Betancur, J. C., Klaile, E., and Slevogt, H. (2018). The interaction of human pathogenic fungi with C-Type lectin receptors. *Front. Immunol* 9:1261. doi: 10.3389/fimmu.2018.01261

Gresnigt, M. S., Cunha, C., Jaeger, M., Gonçalves, S. M., Malireddi, R. K. S., Ammerdorffer, A., et al. (2018). Genetic deficiency of NOD2 confers resistance to invasive aspergillosis. *Nat. Commun.* 9:2636. doi: 10.1038/s41467-018-04912-3

Griffiths, J. S., Camilli, G., Kotowicz, N. K., Ho, J., Richardson, J. P., and Naglik, J. R. (2021). Role for IL-1 family cytokines in fungal infections. *Front. Microbiol.* 12:633047. doi: 10.3389/fmicb.2021.633047

Guo, X., Zhang, R., Li, Y., Wang, Z., Ishchuk, O. P., Ahmad, K. M., et al. (2020). Understand the genomic diversity and evolution of fungal pathogen *Candida glabrata* by genome-wide analysis of genetic variations. *Methods San Diego Calif.* 176, 82–90. doi: 10.1016/j.ymeth.2019.05.002

Hall, R. A., and Noverr, M. C. (2017). Fungal interactions with the human host: exploring the spectrum of symbiosis. *Curr. Opin. Microbiol.* 40, 58–64. doi: 10.1016/j.mib.2017.10.020

Hamad, M., Mohammad, M. G., and Abu-Elteen, K. H. (2018). Immunity to human fungal infections," in *Fungi Biology and Applications*, 3rd Edn, 275–298. doi: 10.1002/9781119374312.ch11

Hatinguais, R., Willment, J. A., and Brown, G. D. (2020). PAMPs of the fungal cell wall and mammalian PRRs. *Curr. Top. Microbiol. Immunol.* 425, 187–223. doi: 10.1007/82_2020_201

Hawksworth, D. L., and Rossman, A. Y. (1997). Where are all the undescribed fungi? *Phytopathology* 87, 888–891. doi: 10.1094/PHYTO.1997.87.9.888

Hirschhorn, J. N., Lohmueller, K., Byrne, E., and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genet. Med. Off. J. Am. Coll. Med. Genet.* 4, 45–61. doi: 10.1097/00125817-200203000-00002

Hughes, A. L., Welch, R., Puri, V., Matthews, C., Haque, K., Chanock, S. J., et al. (2008). Genome-wide SNP typing reveals signatures of population history. *Genomics* 92, 1–8. doi: 10.1016/j.ygeno.2008.03.005

Huppler, A. R., Bishu, S., and Gaffen, S. L. (2012). Mucocutaneous candidiasis: the IL-17 pathway and implications for targeted immunotherapy. *Arthritis Res. Ther.* 14:217. doi: 10.1186/ar3893

Huseyin, C. E., Rubio, R. C., O'Sullivan, O., Cotter, P. D., and Scanlan, P. D. (2017). The fungal frontier: a comparative analysis of methods used in the study of the human gut mycobiome. *Front. Microbiol.* 8:1432. doi: 10.3389/fmicb.2017.01432

Ibrahim, A. S., and Voelz, K. (2017). The mucormycete-host interface. *Curr. Opin. Microbiol.* 40, 40–45. doi: 10.1016/j.mib.2017.10.010

Iliev, I. D., and Leonardi, I. (2017). Fungal dysbiosis: immunity and interactions at mucosal barriers. *Nat. Rev. Immunol.* 17, 635–646. doi: 10.1038/nri.2017.55

Jackson, B. R., Chow, N., Forsberg, K., Litvintseva, A. P., Lockhart, S. R., Welsh, R., et al. (2019). On the origins of a species: what might explain the rise of *Candida auris*? *J. Fungi Basel Switz.* 5:E58. doi: 10.3390/jof5030058

Jacobsen, I. D. (2019). Fungal infection strategies. *Virulence* 10, 835–838. doi: 10.1080/21505594.2019.1682248

Jaeger, M., Matzaraki, V., Aguirre-Gamboa, R., Gresnigt, M. S., Chu, X., Johnson, M. D., et al. (2019). A genome-wide functional genomics approach identifies susceptibility pathways to fungal bloodstream infection in humans. *J. Infect. Dis.* 220, 862–872. doi: 10.1093/infdis/jiz206

Jeffery-Smith, A., Taori, S. K., Schelenz, S., Jeffery, K., Johnson, E. M., Borman, A., et al. (2018). *Candida auris*: a review of the literature. *Clin. Microbiol. Rev.* 31:e00029-17. doi: 10.1128/CMR.00029-17

Johnson, M. D., Plantinga, T. S., van de Vosse, E., Velez Edwards, D. R., Smith, P. B., Alexander, B. D., et al. (2012). Cytokine gene polymorphisms and the outcome of invasive candidiasis: a prospective cohort study. *Clin. Infect. Dis. Off. Publ. Infect. Dis. Soc. Am.* 54, 502–510. doi: 10.1093/cid/cir827

Jonkers, I. H., and Wijmenga, C. (2017). Context-specific effects of genetic variants associated with autoimmune disease. *Hum. Mol. Genet.* 26, R185–R192. doi: 10.1093/hmg/ddx254

Kasper, L., König, A., Koenig, P.-A., Gresnigt, M. S., Westman, J., Drummond, R. A., et al. (2018). The fungal peptide toxin Candidalysin activates the NLRP3 inflammasome and causes cytolysis in mononuclear phagocytes. *Nat. Commun.* 9:4260. doi: 10.1038/s41467-018-06607-1

Kaufman, L. (1985). "The role of specific antibodies of different immunoglobulin classes in the rapid diagnosis of systemic mycotic infections," in *Rapid Methods and Automation in Microbiology and Immunology*, ed. K. O. Habermehl (Berlin: Springer). doi: 10.1007/978-3-642-69943-6_21

Kawai, T., and Akira, S. (2007). TLR signaling. *Semin. Immunol.* 19, 24–32. doi: 10.1016/j.smim.2006.12.004

Kobayashi, G. S. (1996). "Disease mechanisms of fungi," in *Medical Microbiology*, ed. S. Baron 4th Edn, (Galveston, TX: University of Texas). Medical Branch at Galveston.

Kruger, W., Vielreicher, S., Kapitan, M., Jacobsen, I. D., and Niemiec, M. J. (2019). Fungal-Bacterial interactions in health and disease. *Pathog. Basel Switz.* 8:E70. doi: 10.3390/pathogens8020070

Kumar, V., van de Veerdonk, F. L., and Netea, M. G. (2018). Antifungal immune responses: emerging host-pathogen interactions and translational implications. *Genome Med.* 10:39. doi: 10.1186/s13073-018-0553-2

Kumaresan, P. R., da Silva, T. A., and Kontoyiannis, D. P. (2017). Methods of controlling invasive fungal infections using CD8+ T cells. *Front. Immunol.* 8:1939. doi: 10.3389/fimmu.2017.01939

Kutukculer, N., Aykut, A., Karaca, N. E., Durmaz, A., Aksu, G., Genel, F., et al. (2019). Chronic granulamatous disease: two decades of experience from a paediatric immunology unit in a country with high rate of consangineous marriages. *Scand. J. Immunol.* 89:e12737. doi: 10.1111/sji.12737

Kwizera, R., Musaazi, J., Meya, D. B., Worodria, W., Bwanga, F., Kajumbula, H., et al. (2019). Burden of fungal asthma in Africa: a systematic review and meta-analysis. *PLoS One* 14:e0216568. doi: 10.1371/journal.pone.0216568

Lamoth, F., and Kontoyiannis, D. P. (2018). The *Candida auris* alert: facts and perspectives. *J. Infect. Dis.* 217, 516–520. doi: 10.1093/infdis/jix597

Lanternier, F., Cypowyj, S., Picard, C., Bustamante, J., Lortholary, O., Casanova, J.-L., et al. (2013). Primary immunodeficiencies underlying fungal infections. *Curr. Opin. Pediatr.* 25, 736–747. doi: 10.1097/MOP.0000000000000031

Larcombe, L., Rempel, J. D., Dembinski, I., Tinckam, K., Rigatto, C., and Nickerson, P. (2005). Differential cytokine genotype frequencies among Canadian aboriginal and Caucasian populations. *Genes Immun.* 6, 140–144.

Lehrer, R. I., and Cline, M. J. (1969). Leukocyte myeloperoxidase deficiency and disseminated candidiasis: the role of myeloperoxidase in resistance to *Candida* infection. *J. Clin. Invest.* 48, 1478–1488. doi: 10.1172/JCI106114

León-Lara, X., Hernández-Nieto, L., Zamora, C. V., Rodríguez-D'Cid, R., Gutiérrez, M. E. C., Espinosa-Padilla, S., et al. (2020). Disseminated infectious disease caused by histoplasma capsulatum in an adult patient as first manifestation of inherited IL-12Rβ1 deficiency. *J. Clin. Immunol.* 40, 1051–1054. doi: 10.1007/s10875-020-00828-0

Limon, J. J., Skalski, J. H., and Underhill, D. M. (2017). Commensal fungi in health and disease. *Cell Host & Microbe* 22, 156–165. doi: 10.1016/j.chom.2017.07.002

Lionakis, M. S. (2012). Genetic susceptibility to fungal infections in humans. *Curr. Fungal Infect. Rep.* 6, 11–22. doi: 10.1007/s12281-011-0076-4

Lionakis, M. S., Netea, M. G., and Holland, S. M. (2014). Mendelian genetics of human susceptibility to fungal infection. *Cold Spring Harb. Perspect. Med.* 4:a019638. doi: 10.1101/cshperspect.a019638

Low, C.-Y., and Rotstein, C. (2011). Emerging fungal infections in immunocompromised patients. *F1000 Med. Rep.* 3:14. doi: 10.3410/M3-14

Lupiañez, C. B., Martínez-Bueno, M., Sánchez-Maldonado, J. M., Badiola, J., Cunha, C., Springer, J., et al. (2020). Polymorphisms within the ARNT2 and CX3CR1 genes are associated with the risk of developing invasive aspergillosis. *Infect. Immun.* 88, e882–e819. doi: 10.1128/IAI.00882-19

Martin, E. M., and Fry, R. C. (2018). Environmental influences on the epigenome: exposure- associated DNA methylation in human populations. *Annu. Rev. Public Health* 39, 309–333. doi: 10.1146/annurev-publhealth-040617-014629

Maskarinec, S. A., Johnson, M. D., and Perfect, J. R. (2016). Genetic susceptibility to fungal infections: what is in the genes? *Curr. Clin. Microbiol. Rep.* 3, 81–91. doi: 10.1007/s40588-016-0037-3

Merkhofer, R. M., and Klein, B. S. (2020). Advances in understanding human genetic variations that influence innate immunity to fungi. *Front. Cell. Infect. Microbiol.* 10:69. doi: 10.3389/fcimb.2020.00069

Merkhofer, R. M., O'Neill, M. B., Xiong, D., Hernandez-Santos, N., Dobson, H., Fites, J. S., et al. (2019). Investigation of genetic susceptibility to blastomycosis reveals interleukin-6 as a potential susceptibility locus. *mBio* 10:e01224-19. doi: 10.1128/mBio.01224-19

Minegishi, Y., Saito, M., Morio, T., Watanabe, K., Agematsu, K., Tsuchiya, S., et al. (2006). Human tyrosine kinase 2 deficiency reveals its requisite roles in multiple cytokine signals involved in innate and acquired immunity. *Immunity* 25, 745–755. doi: 10.1016/j.immuni.2006.09.009

Mogensen, T. H. (2009). Pathogen recognition and inflammatory signaling in innate immune defenses. *Clin. Microbiol. Rev.* 22, 240–273. doi: 10.1128/CMR. 00046-08 Table of Contents

Nahum, A. (2017). Chronic mucocutaneous candidiasis: a spectrum of genetic disorders. *LymphoSign J.* 4, 87–99.

Naranjo-Ortiz, M. A., and Gabaldón, T. (2019). Fungal evolution: major ecological adaptations and evolutionary transitions. *Biol. Rev. Camb. Philos. Soc.* 94, 1443–1476. doi: 10.1111/brv.12510

Navarro-Mendoza, M. I., Pérez-Arques, C., Murcia, L., Martínez-García, P., Lax, C., Sanchis, M., et al. (2018). Components of a new gene family of ferroxidases involved in virulence are functionally specialized in fungal dimorphism. *Sci. Rep.* 8:7660. doi: 10.1038/s41598-018-26051-x

Netea, M. G., Schlitzer, A., Placek, K., Joosten, L. A. B., and Schultze, J. L. (2019). Innate and adaptive immune memory: an evolutionary continuum in the host's response to pathogens. *Cell Host Microbe* 25, 13–26. doi: 10.1016/j.chom.2018. 12.006

Netea, M. G., Wijmenga, C., and O'Neill, L. A. J. (2012). Genetic variation in Toll-like receptors and disease susceptibility. *Nat. Immunol.* 13, 535–542. doi: 10.1038/ni.2284

Pana, Z.-D., Farmaki, E., and Roilides, E. (2014). Host genetics and opportunistic fungal infections. *Clin. Microbiol. Infect. Off. Publ. Eur. Soc. Clin. Microbiol. Infect. Dis.* 20, 1254–1264. doi: 10.1111/1469-0691.12800

Panday, A., Sahoo, M. K., Osorio, D., and Batra, S. (2015). NADPH oxidases: an overview from structure to innate immunity-associated pathologies. *Cell. Mol. Immunol.* 12, 5–23. doi: 10.1038/cmi.2014.89

Pathakumari, B., Liang, G., and Liu, W. (2020). Immune defence to invasive fungal infections: a comprehensive review. *Biomed. Pharmacother. Biomed. Pharmacother.* 130:110550. doi: 10.1016/j.biopha.2020.110550

Patin, E. C., Thompson, A., and Orr, S. J. (2019). Pattern recognition receptors in fungal immunity. *In Semin. Cell Dev. Biol.* 89, 4–33.

Pedroza, L. A., Kumar, V., Sanborn, K. B., Mace, E. M., Niinikoski, H., Nadeau, K., et al. (2012). Autoimmune regulator (AIRE) contributes to Dectin-1-induced TNF-α production and complexes with caspase recruitment domain-containing protein 9 (CARD9), spleen tyrosine kinase (Syk), and Dectin-1. *J. Allergy Clin. Immunol.* 129, 464–472, 472.e1–3.

Perez, N. B., Wright, F., and Vorderstrasse, A. (2021). A microbial relationship between irritable bowel syndrome and depressive symptoms. *Biol. Res. Nurs.* 23, 50–64. doi: 10.1177/1099800420940787

Pfavayi, L. T., Sibanda, E. N., and Mutapi, F. (2020). The pathogenesis of fungal-related diseases and allergies in the African population: the state of the evidence and knowledge gaps. *Int. Arch. Allergy Immunol.* 181, 257–269. doi: 10.1159/ 000506009

Plantinga, T. S., Johnson, M. D., Scott, W. K., Joosten, L. A. B., van der Meer, J. W. M., Perfect, J. R., et al. (2012). Human genetic susceptibility to *Candida* infections. *Med. Mycol.* 50, 785–794. doi: 10.3109/13693786.2012.690902

Plantinga, T. S., van der Velden, W. J. F. M., Ferwerda, B., van Spriel, A. B., Adema, G., Feuth, T., et al. (2009). Early stop polymorphism in human DECTIN-1 is associated with increased candida colonization in hematopoietic stem cell transplant recipients. *Clin. Infect. Dis.* 49, 724–732. doi: 10.1086/604714

Plato, A., Hardison, S. E., and Brown, G. D. (2015). Pattern recognition receptors in antifungal immunity. *Semin. Immunopathol.* 37, 97–106. doi: 10.1007/s00281-014-0462-4

Puel, A., Cypowyj, S., Bustamante, J., Wright, J. F., Liu, L., Lim, H. K., et al. (2011). Chronic mucocutaneous candidiasis in humans with inborn errors of interleukin-17 immunity. *Science* 332, 65–68. doi: 10.1126/science.1200439

Rai, L. S., Wijlick, L. V., Bougnoux, M. E., Bachellier-Bassi, S. and d'Enfert, C. (2021). Regulators of commensal and pathogenic life-styles of an opportunistic fungus–Candida albicans. *Yeast* 38, 243–250. doi: 10.1002/yea.3550

Raut, A., and Huy, N. T. (2021). Rising incidence of mucormycosis in patients with COVID-19: another challenge for India amidst the second wave? *Lancet Respir. Med.* 3, 265–264. doi: 10.1016/S2213-2600(21)00265-4

Reid, D. M., Gow, N. A. R., and Brown, G. D. (2009). Pattern recognition: recent insights from Dectin-1. *Curr. Opin. Immunol.* 21, 30–37. doi: 10.1016/j.coi. 2009.01.003

Ren, R., Fedoriw, Y., and Willis, M. (2012). The molecular pathophysiology, differential diagnosis, and treatment of MPO deficiency. *J. Clin. Exp. Pathol.* 2, 2161–2681.

Richmond, J. M., and Harris, J. E. (2014). Immunology and skin in health and disease. *Cold Spring Harb. Perspect. Med.* 4:a015339. doi: 10.1101/cshperspect. a015339

Roilides, E., Dimitriadou-Georgiadou, A., Sein, T., Kadiltsoglou, I., and Walsh, T. J. (1998). Tumor necrosis factor alpha enhances antifungal activities of polymorphonuclear and mononuclear phagocytes against *Aspergillus fumigatus*. *Infect. Immun.* 66, 5999–6003. doi: 10.1128/IAI.66.12.5999-6003. 1998

Rosales, C., and Uribe-Querol, E. (2017). Phagocytosis: a fundamental process in immunity. *BioMed Res. Int.* 2017:9042851. doi: 10.1155/2017/9042851

Rosentul, D. C., Plantinga, T. S., Scott, W. K., Alexander, B. D., van de Geer, N. M. D., Perfect, J. R., et al. (2012). The impact of caspase-12 on susceptibility to candidemia. *Eur. J. Clin. Microbiol. Infect. Dis.* 31, 277–280. doi: 10.1007/ s10096-011-1307-x

Sainz, J., Lupiáñez, C. B., Segura-Catena, J., Vazquez, L., Ríos, R., Oyonarte, S., et al. (2012). Dectin-1 and DC-SIGN polymorphisms associated with invasive pulmonary *Aspergillosis* infection. *PLoS One* 7:e32273. doi: 10.1371/journal. pone.0032273

Sainz, J., Pérez, E., Hassan, L., Moratalla, A., Romero, A., Collado, M. D., et al. (2007). Variable number of tandem repeats of TNF receptor type 2 promoter as genetic biomarker of susceptibility to develop invasive pulmonary *Aspergillosis*. *Hum. Immunol.* 68, 41–50. doi: 10.1016/j.humimm.2006.10.011

Salazar, F., and Brown, G. D. (2018). Antifungal innate immunity: a perspective from the last 10 years. *J. Innate Immun.* 10, 373–397. doi: 10.1159/000488539

Sampaio, E. P., Hsu, A. P., Pechacek, J., Bax, H. I., Dias, D. L., Paulson, M. L., et al. (2013). Signal transducer and activator of transcription 1 (STAT1) gain-of-function mutations and disseminated coccidioidomycosis and histoplasmosis. *J. Allergy Clin. Immunol.* 131, 1624–1634. doi: 10.1016/j.jaci.2013.01.052

Sardinha, J. F. J., Tarlé, R. G., Fava, V. M., Francio, A. S., Ramos, G. B., Ferreira, L. C., et al. (2011). Genetic risk factors for human susceptibility to infections of relevance in dermatology. *An. Bras. Dermatol.* 86, 708–715. doi: 10.1590/s0365-05962011000400013

Sawada, Y., Setoyama, A., Sakuragi, Y., Saito-Sasaki, N., Yoshioka, H., and Nakamura, M. (2021). The role of IL-17-Producing cells in cutaneous fungal infections. *Int. J. Mol. Sci.* 22:5794. doi: 10.3390/ijms22115794

Schuetz, C., Huck, K., Gudowius, S., Megahed, M., Feyen, O., Hubner, B., et al. (2008). An immunodeficiency disease with RAG mutations and granulomas. *N. Engl. J. Med.* 358, 2030–2038. doi: 10.1056/NEJMoa073966

Segal, B. H., and Romani, L. R. (2009). Invasive aspergillosis in chronic granulomatous disease. *Med. Mycol.* 47(Suppl. 1), S282–S290. doi: 10.1080/ 13693780902736620

Smeekens, S. P., van de Veerdonk, F. L., Kullberg, B. J., and Netea, M. G. (2013). Genetic susceptibility to *Candida* infections. *EMBO Mol. Med.* 5, 805–813. doi: 10.1002/emmm.201201678

Sobel, J. D. (2016). Recurrent vulvovaginal candidiasis. *Am. J. Obstet. Gynecol.* 214, 15–21. doi: 10.1016/j.ajog.2015.06.067

Sparber, F., and LeibundGut-Landmann, S. (2019). Interleukin-17 in antifungal immunity. *Pathog. Basel Switz.* 8:E54. doi: 10.3390/pathogens8020054

Speakman, E. A., Dambuza, I. M., Salazar, F., and Brown, G. D. (2020). T cell antifungal immunity and the role of C-Type lectin receptors. *Trends Immunol.* 41, 61–76. doi: 10.1016/j.it.2019.11.007

Spinner, M. A., Sanchez, L. A., Hsu, A. P., Shaw, P. A., Zerbe, C. S., Calvo, K. R., et al. (2014). GATA2 deficiency: a protean disorder of hematopoiesis, lymphatics, and immunity. *Blood* 123, 809–821. doi: 10.1182/blood-2013-07-515528

Stasia, M. J. (2016). CYBA encoding p22(phox), the cytochrome b558 alpha polypeptide: gene structure, expression, role and physiopathology. *Gene* 586, 27–35. doi: 10.1016/j.gene.2016.03.050

Stasia, M. J., Brion, J.-P., Boutonnat, J., and Morel, F. (2003). Severe clinical forms of cytochrome b-negative chronic granulomatous disease (X91-) in 3 brothers with a point mutation in the promoter region of CYBB. *J. Infect. Dis.* 188, 1593–1604. doi: 10.1086/379035

Steenwyk, J. L., Lind, A. L., Ries, L. N. A., Dos Reis, T. F., Silva, L. P., Almeida, F., et al. (2020). Pathogenic allodiploid hybrids of *Aspergillus* fungi. *Curr. Biol. CB* 30, 2495–2507.e7.

Taylor, M. L., Pérez-Mejía, A., Yamamoto-Furusho, J. K., and Granados, J. (1997). Immunologic, genetic and social human risk factors associated to histoplasmosis: studies in the State of Guerrero, Mexico. *Mycopathologia* 138, 137–142. doi: 10.1023/a:1006847630347

Taylor, P. R., Tsoni, S. V., Willment, J. A., Dennehy, K. M., Rosas, M., Findon, H., et al. (2007). Dectin-1 is required for beta-glucan recognition and control of fungal infection. *Nat. Immunol.* 8, 31–38. doi: 10.1038/ni1408

Thompson, A., and Orr, S. J. (2018). Emerging IL-12 family cytokines in the fight against fungal infections. *Cytokine* 111, 398–407. doi: 10.1016/j.cyto.2018.05.019

Tiew, P. Y., Jaggi, T. K., Chan, L. L. Y., and Chotirmall, S. H. (2021). The airway microbiome in COPD, bronchiectasis and bronchiectasis-COPD overlap. *Clin. Respir. J.* 15, 123–133. doi: 10.1111/crj.13294

Tsai, H.-J., and Nelliat, A. (2019). A Double-Edged sword: aneuploidy is a prevalent strategy in fungal adaptation. *Genes* 10:E787. doi: 10.3390/genes10100787

Ulfig, A., and Leichert, L. I. (2021). The effects of neutrophil-generated hypochlorous acid and other hypohalous acids on host and pathogens. *Cell. Mol. Life Sci.* 78, 385–414. doi: 10.1007/s00018-020-03591-y

Underhill, D. M., and Pearlman, E. (2015). Immune interactions with pathogenic and commensal fungi: a two-way street. *Immunity* 43, 845–858. doi: 10.1016/j.immuni.2015.10.023

Urban, C. F., Ermert, D., Schmid, M., Abu-Abed, U., Goosmann, C., Nacken, W., et al. (2009). Neutrophil extracellular traps contain calprotectin, a cytosolic protein complex involved in host defense against *Candida albicans*. *PLoS Pathog.* 5:e1000639. doi: 10.1371/journal.ppat.1000639

Vaid, M., Kaur, S., Sambatakou, H., Madan, T., Denning, D. W., and Sarma, P. U. (2007). Distinct alleles of mannose-binding lectin (MBL) and surfactant proteins A (SP-A) in patients with chronic cavitary pulmonary aspergillosis and allergic bronchopulmonary aspergillosis. *Clin. Chem. Lab. Med.* 45, 183–186. doi: 10.1515/CCLM.2007.033

van Burik, J. A., and Magee, P. T. (2001). Aspects of fungal pathogenesis in humans. *Annu. Rev. Microbiol.* 55, 743–772. doi: 10.1146/annurev.micro.55.1.743

van de Veerdonk, F. L., Kullberg, B. J., van der Meer, J. W., Gow, N.A., and Netea, M. G. (2008). Host-microbe interactions: innate pattern recognition of fungal pathogens. *Curr. Opin. Microbiol.* 11, 305–312. doi: 10.1016/j.mib.2008.06.002

Vautier, S., Sousa, M., da, G., and Brown, G. D. (2010). C-type lectins, fungi and Th17 responses. *Cytokine Growth Factor Rev.* 21, 405–412. doi: 10.1016/j.cytogfr.2010.10.001

Vedula, R. S., Cheng, M. P., Ronayne, C. E., Farmakiotis, D., Ho, V. T., Koo, S., et al. (2021). Somatic GATA2 mutations define a subgroup of myeloid malignancy patients at high risk for invasive fungal disease. *Blood Adv.* 5, 54–60. doi: 10.1182/bloodadvances.2020002854

Verma, A., Wüthrich, M., Deepe, G., and Klein, B. (2015). Adaptive immunity to fungi. *Cold Spring Harb. Perspect. Med.* 5:a019612. doi: 10.1101/cshperspect.a019612

Vijaya Chandra, S. H., Srinivas, R., Dawson, T. L. Jr., and Common, J. E. (2021). Cutaneous *Malassezia*: commensal, pathogen, or protector? *Front. Cell. Infect. Microbiol.* 10:614446. doi: 10.3389/fcimb.2020.614446

Vinh, D. C. (2019). The molecular immunology of human susceptibility to fungal diseases: lessons from single gene defects of immunity. *Expert Rev. Clin. Immunol.* 15, 461–486. doi: 10.1080/1744666X.2019.1584038

Vornholz, L., and Ruland, J. (2020). Physiological and pathological functions of CARD9 signaling in the innate immune system. *Curr. Top. Microbiol. Immunol.* 429, 177–203. doi: 10.1007/82_2020_211

Wang, Z., Zhang, S., Xiao, Y., Zhang, W., Wu, S., Qin, T., et al. (2020). NLRP3 inflammasome and inflammatory diseases. *Oxid. Med. Cell. Longev.* 2020:4063562. doi: 10.1155/2020/4063562

Warris, A., and Ballou, E. R. (2019). Oxidative responses and fungal infection biology. *Semin. Cell Dev. Biol.* 89, 34–46. doi: 10.1016/j.semcdb.2018.03.004

Wójtowicz, A., Bibert, S., Taffé, P., Bernasconi, E., Furrer, H., Günthard, H. F., et al. (2019). IL-4 polymorphism influences susceptibility to *Pneumocystis jirovecii* pneumonia in HIV-positive patients. *AIDS* 33, 1719–1727. doi: 10.1097/QAD.0000000000002283

Wu, S.-Y., Weng, C.-L., Jheng, M.-J., Kan, H.-W., Hsieh, S.-T., Liu, F.-T., et al. (2019). Candida albicans triggers NADPH oxidase-independent neutrophil extracellular traps through dectin-2. *PLoS Pathog.* 15:e1008096. doi: 10.1371/journal.ppat.1008096

Xiong, D., Meece, J. K., and Pepperell, C. S. (2013). Genetic research with hmong-ancestry populations: lessons from the literature and a pilot study. *Hmong Stud. J.* 14, 1–28.

Yanagisawa, K., Wichukchinda, N., Tsuchiya, N., Yasunami, M., Rojanawiwat, A., Tanaka, H., et al. (2020). Deficiency of mannose-binding lectin is a risk of *Pneumocystis jirovecii* pneumonia in a natural history cohort of people living with HIV/AIDS in Northern Thailand. *PLoS One* 15:e0242438. doi: 10.1371/journal.pone.0242438

Zaas, A. K. (2006). Host genetics affect susceptibility to invasive aspergillosis. *Med. Mycol.* 44, S55–S60. doi: 10.1080/13693780600865481

Zahedi, N., Abedian Kenari, S., Mohseni, S., Aslani, N., Ansari, S., and Badali, H. (2016). Is human Dectin-1 Y238X gene polymorphism related to susceptibility to recurrent vulvovaginal candidiasis? *Curr. Med. Mycol.* 2, 15–19. doi: 10.18869/acadpub.cmm.2.3.15

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.

ORIGINAL RESEARCH
published: 30 August 2021
doi: 10.3389/fgene.2021.678957

Epidemiology of Rare Hereditary Diseases in the European Part of Russia: Point and Cumulative Prevalence

Rena A. Zinchenko[1,2]*, *Eugeny K. Ginter*[1], *Andrey V. Marakhonov*[1]*, *Nika V. Petrova*[1], *Vitaly V. Kadyshev*[1], *Tatyana P. Vasilyeva*[2], *Oksana U. Alexandrova*[2], *Alexander V. Polyakov*[1] and *Sergey I. Kutsev*[1]

[1] Research Centre for Medical Genetics, Moscow, Russia, [2] Department of Public Health Research, N.A. Semashko National Research Institute of Public Health, Moscow, Russia

OPEN ACCESS

Edited by:
Elvira Galieva,
Novosibirsk State University, Russia

Reviewed by:
Feng Zhang,
Fudan University, China
Evgenii Chekalin,
Michigan State University,
United States
Nina Oparina,
University of Gothenburg, Sweden
Milan Macek,
Charles University, Czechia

***Correspondence:**
Rena A. Zinchenko
renazinchenko@mail.ru
Andrey V. Marakhonov
marakhonov@generesearch.ru

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: 10 March 2021
Accepted: 15 July 2021
Published: 30 August 2021

Citation:
Zinchenko RA, Ginter EK,
Marakhonov AV, Petrova NV,
Kadyshev VV, Vasilyeva TP,
Alexandrova OU, Polyakov AV and
Kutsev SI (2021) Epidemiology
of Rare Hereditary Diseases
in the European Part of Russia: Point
and Cumulative Prevalence.
Front. Genet. 12:678957.
doi: 10.3389/fgene.2021.678957

The issue of point prevalence, cumulative prevalence (CP), and burden of rare hereditary diseases (RHD), comprising 72–80% of the group of rare diseases, is discussed in many reports and is an urgent problem, which is associated with the rapid progress of genetic technology, the identification of thousands of genes, and the resulting problems in society. This work provides an epidemiological analysis of the groups of the most common RHDs (autosomal dominant, autosomal recessive, and X-linked) and their point prevalence (PP) and describes the structure of RHD diversity by medical areas in 14 spatially remote populations of the European part of Russia. The total size of the examined population is about 4 million. A total of 554 clinical forms of RHDs in 10,265 patients were diagnosed. The CP for all RHDs per sample examined was 277.21/100,000 (1:361 people). It is worth noting that now is the time for characterizing the accumulated data on the point prevalence of RHDs, which will help to systematize our knowledge and allow us to develop a strategy of care for patients with RHDs. However, it is necessary to address the issues of changing current medical classifications and coding systems for nosological forms of RHDs, which have not kept pace with genetic advances.

Keywords: genetic epidemiology, rare hereditary diseases, point prevalence, cumulative prevalence, Russia

INTRODUCTION

The problem of rare diseases (RDs) and their number, birth, point, and cumulative prevalence are actively discussed by many researchers, and this is important for public health and society. Criteria for the definition of "rare diseases" differ from country to country depending on legislation. A review by Richter et al. (2015) provides data on 296 definitions of RDs from 1,109 organizations (Richter et al., 2015) and confirms their quantitative differences in different countries. European legislation defines a prevalence threshold of 1 per 2,000 persons. The United States in 1983 defined the threshold for RD as <200,000 affected people in the country (currently 1 in 1,800 people). Japan considers any disease affecting less than 50,000 people in the country as rare, which is equivalent to less than 1 in 2,500 people. In Russia, one patient per 10,000 people in the population is a sufficient measure for a disease being rare (European Union (EU), 2000; Donnart et al., 2013; Richter et al., 2015; Ferreira, 2019; Wakap et al., 2020). RDs are diagnosed in all fields of medicine and occur in all demographic groups (Pariser and Gahl, 2014).

There is a variation and steady increase in the reported RDs according to the main available sources. The Online Mendelian Inheritance in Man (OMIM) database contains 6,806 phenotypes with known genetic nature[1]. According to the Orphanet portal of RDs, about 10,500 RDs are currently registered[2]. A very detailed analysis of the known number, point, and cumulative prevalences of RDs has been performed by several teams (Ferreira, 2019; Wakap et al., 2020). Reviews on RD number analysis and point prevalence estimation cite data from the "Epidemiology section of Orphanet[3]". It has been shown that 84.5% of the analyzed diseases from the Orphanet database have a point prevalence less than 1 per 1,000,000. However, 77.3–80.7% of the burden of RDs in the population falls on a limited number of diseases, representing only about 4.2% of all identified diseases, with a point prevalence of one to five per 10,000 of the population (Wakap et al., 2020). Ferreira estimates the burden of RDs with manifestation at different periods of life as 6.2% of the total population (Ferreira, 2019). A more conservative estimate by Wakap et al. (2020) demonstrates the cumulative prevalence of RDs in the population to be 3.5–5.9%.

According to various researchers, 72–80% of RDs have a genetic cause, some of which have already been confirmed (Richter et al., 2015; Wakap et al., 2020). The widespread introduction of technologies of whole genome and/or whole-exome analysis into practical healthcare has significantly increased the number of genetically determined diseases in the structure of human morbidity. While by 2012 the molecular nature was identified for 3,650 nosological forms, there has been a greater increase in genetically determined diseases over the past 5 years. According to annual observations of OMIM statistics in 2016, molecular nature was confirmed for 5,888 diseases, in 2017—6,087 (+199), in 2018—6340 (+253), in 2019—6572 (+232), and in 2020—6,800 (+228), i.e., genetic nature is established for an additional 200–250 diseases each year (see text footnote 1). Most of the newly reported forms are rare and found in single families.

Since genetic diseases constitute a high percentage in the RD group, it is advisable to assess the point prevalence of monogenic hereditary diseases in the modern population based on actual data from a specific population survey. This article provides an epidemiological analysis of the point prevalence of rare monogenic hereditary diseases (RHDs) in geographically remote populations of the European part of Russia.

RESULTS

The population of 96 rural areas, 86 small towns, and urban-type settlements were surveyed during this study. Data on patients with presumptive RHDs were obtained using a questionnaire (Zinchenko et al., 2020b) from medical workers from 125 medical clinics and 2,056 rural ambulant clinics. More than 45,000 patients with various presumably hereditary conditions, including patients with structural chromosomal changes, multiple congenital malformations, and isolated anomalies, were examined in total. A total of 554 clinical forms of RHDs were verified in 10,265 patients (including 4,270 children patients).

Table 1 shows the number of identified patients with RHDs in the regions and the variation of the cumulative prevalence for administrative districts within particular regions.

The cumulative prevalence for all RHDs in the sample examined was 277.21 per 100,000 (Zinchenko et al., 2001a,b, 2007, 2009), 558.71 per 100,000 children (Zinchenko et al., 2019, 2020a). The differentiation in the values of cumulative prevalence by region was explained by the peculiarities of the genetic structure of various populations. In the sample under consideration, the main factor of microevolution is genetic drift, migration processes with little influence of natural selection. These results were demonstrated in studies on the role of the genetic structure in the formation of cumulative prevalence in every population (Zinchenko et al., 2000, 2009, 2020b). The number of familial cases averaged 57.82%. There were no statistically significant differences between the cumulative prevalence of AD and AR pathology in the groups of men and women ($p > 0.05$).

After that, we have analyzed the point prevalence values and the number of diseases by groups in the surveyed populations (**Table 2**). The highest number of patients was found in point prevalence class 1 (1:50,000 and more) (59.46%) with the autosomal dominant (AD) type of inheritance—the autosomal recessive (AR) type of inheritance was observed in 34.66% of patients, the X-linked (XL) type of inheritance in 7.88%. The

TABLE 1 | Number of identified patients with RHDs and variation of cumulative prevalence for administrative districts within particular regions (min/max).

Region of the Russian Federation	Surveyed population (number of districts)	Number of identified patients with RHDs	Variation of cumulative prevalence for districts (min-max)
Central part of Russia			
Kostroma region	444,476 (10)	673	1:121-1:545
Kirov region	286,600 (11)	589	1:83-1:548
Bryansk region	88,200 (1)	133	1:324-1:422
Tver region	75,000 (2)	131	1:260-1:405
Republic of Mari El	276,000 (7)	630	1:78-1:286
Chuvash Republic	264,419 (6)	679	1:150-1:550
Republic of Udmurtia	267,655 (6)	794	1:78-1:375
Republic of Tatarstan	264,098 (8)	1516	1:88-1:350
Republic of Bashkortostan	250,110 (8)	1192	1:88-1:389
Northern part of Russia			
Arkhangelsk region	40,000 (5)	104	1:150-1:281
Southern part of Russia and North Caucasus			
Krasnodar territory	426,600 (6)	740	1:202-1:556
Rostov region	497,460 (12)	1481	1:165-1:340
Republic of Adygea	112,400 (4)	233	1:236-1:387
Republic of Karachay-Cherkessia	410,368 (10)	1857	1:85-1:405
Average	**3,703,018 (96)**	10265	

[1] https://omim.org/statistics/geneMap

[2] http://www.orphadata.org/

[3] http://www.orphadata.org/cgi-bin/epidemio.html

TABLE 2 | Distribution of patients with AD, AR, and X-linked inheritance patterns of RHDs depending on the point prevalence values and the number of diseases by groups[1].

Point prevalence	AD		AR		XL		Total	
	Abs. num. of patients (%)	Num. of diseases (%)	Abs. num. of patients (%)	Num. of diseases (%)	Abs. num. of patients (%)	Num. of diseases (%)	Abs. num. of patients (%)	Num. of diseases (%)
1:50,000 and more	3,358 (56.93%)	17 (6.39%)	2,253 (63.32%)	11 (4.80%)	534 (66.01%)	5 (8.47%)	6,145 (59.86%)	33 (5.96%)
1:50,001–1:100,000	742 (12.58%)	16 (6.02%)	205 (5.76%)	5 (2.18%)	76 (9.39%)	3 (5.08%)	1023 (9.97%)	24 (4.33%)
1:100,001–1:200,000	643 (10.90%)	26 (9.77%)	363 (10.20%)	10 (4.37%)	72 (8.90%)	6 (10.17%)	1,078 (10.50%)	42 (7.58%)
1:200,001–1:300,000	481 (8.16%)	34 (12.78%)	193 (5.42%)	14 (6.11%)	43 (5.32%)	6 (10.17%)	717 (6.98%)	54 (9.75%)
1:300,001–1:400,001	214 (3.63%)	22 (8.27%)	110 (3.09%)	11 (4.80%)	10 (1.24%)	2 (3.39%)	334 (3.25%)	35 (6.32%)
1:400,001–1:500,002	56 (0.95%)	7 (2.63%)	64 (1.80%)	8 (3.49%)	4 (0.49%)	1 (1.69%)	124 (1.21%)	16 (2.89%)
1:500,001 and less	404 (6.85%)	144 (54.14%)	370 (10.40%)	170 (74.24%)	70 (8.65%)	36 (61.02%)	844 (8.22%)	350 (63.18%)
Total	5,898	266	3,558	229	809	59	10,265	554

[1] OMIM—Online Mendelian Inheritance in Man; PS—Phenotypic Series for OMIM with heterogeneity of the disease; AD—autosomal dominant type of inheritance; AR—autosomal recessive type of inheritance; XL—X-linked type of inheritance. Point prevalence for X-linked pathology is for 100,000 of the male population.

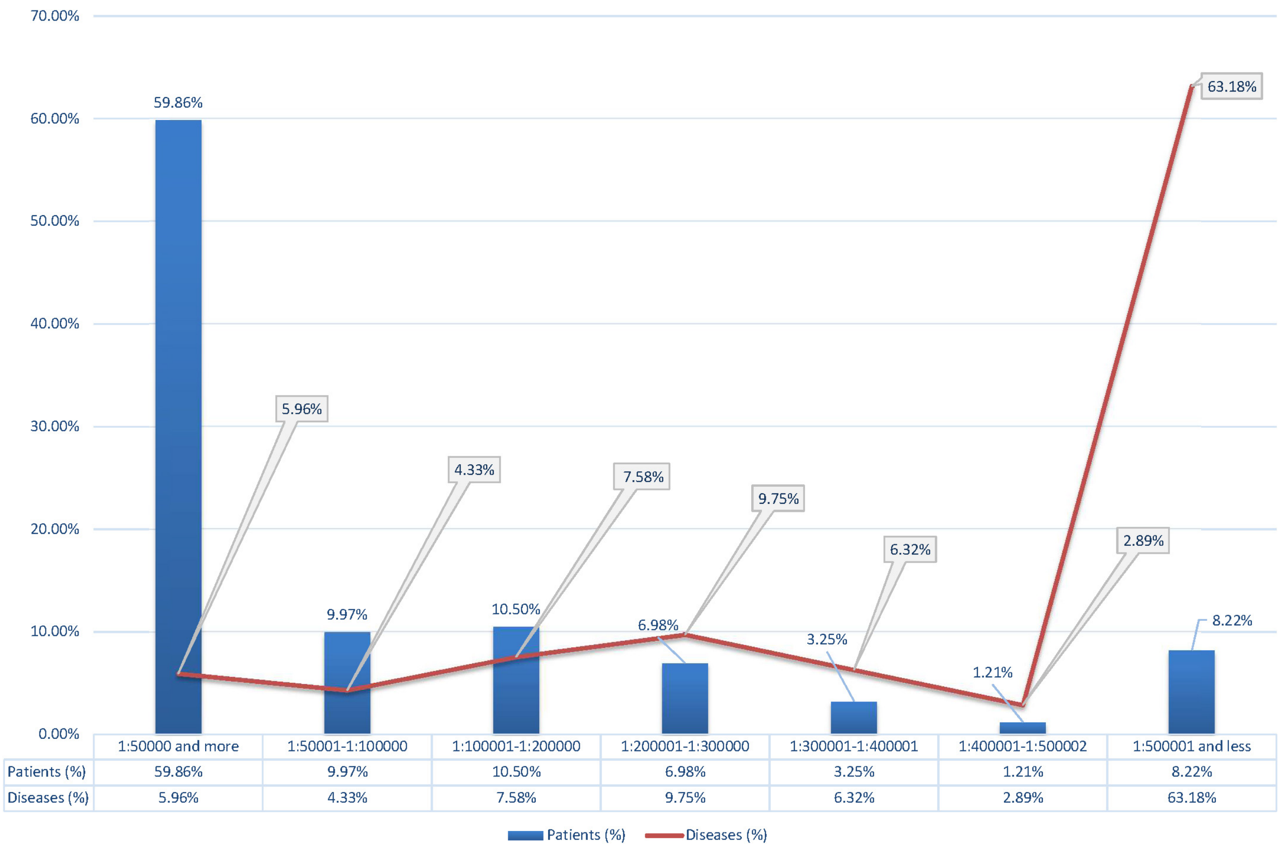
same ratio is observed in the analysis of the number of detected diseases: with the AD type of inheritance—48.01%; with AR—41.34%; and with XL—10.65%. Analysis of the distribution of the number of patients according to the point prevalence values showed that 33 diseases accounted for most of the patients (59.86%), representing only 5.96% of the total number of diseases (**Figure 1**). The class of RHDs with a point prevalence of "1:500,001 and less" represents 350 diseases (63.18% of all detected RHDs) with 844 patients (8.22% of the patients): 6.85% with the AD type of inheritance, 10.40% with the AR type, and 8.65% with the XL type.

Next, we have analyzed reported diseases and the number of patients with a point prevalence of 1:100,000 or more (**Table 3**). **Table 3** also presents disease prevalence values (adjusted for gender differences) and point prevalence in the child population. It is worth noting that for congenital diseases and those with manifestation in childhood, the prevalence in the child population is much higher than in the general population. Most of the diseases are from heterogeneous groups but having a similar clinical picture. For example, about 100 genetic forms with similar clinical manifestations of the disease are known for retinitis pigmentosa. In our studies, we also determined a wide heterogeneity for this disease based on NGS studies—we diagnosed 35 clinical genetic variants, and they are combined under the clinical diagnosis "retinitis pigmentosa" in **Table 3**. Some forms were listed separately, i.e., Stargardt disease, because it has characteristic features and is well diagnosed clinically. The same pattern is observed for Charcot–Marie–Tooth disease, the verification of specific nosological forms in most cases being determined by a molecular genetic study. RHDs vary in prevalence, locus, and allelic heterogeneity depending on the geographic region, which is associated with the specific genetic structure of the population. More than 20 rare genetic variants for Charcot–Marie–Tooth disease associated with different loci and represented by single families have been identified in different regions of Russia (Schagina et al., 2007; Khidiyatova et al., 2013; Dadali et al., 2016; Shchagina et al., 2018, 2020; Murtazina et al., 2020).

Differential diagnosis of "undifferentiated intellectual disability" is particularly difficult, and work is being done in this direction identifying new genes and genetic variants, most of which we identify in single families (Levchenko et al., 2019). In our sample, the majority of cases were familial.

In Russia, we have identified allelic heterogeneity, and regional and ethnic peculiarities, for most diseases included in newborn screening (Petrova et al., 2016, 2019a,b; Gundorova et al., 2018). Our studies identified the highest prevalence of phenylketonuria around the world (birth prevalence 1:850 newborns) in one region of the Russian Federation, and it is associated with the spread of a specific mutation (p.R261* in the *PAH* gene) associated with the founder effect (Gundorova et al., 2018). The study of allelic heterogeneity is necessary because the varying significance of different genetic variants leads to different clinical courses of diseases and treatment correction (Petrova et al., 2016, 2019a,b; Gundorova et al., 2018). For some diseases, the analysis of genotype–phenotype correlations and the mutation spectrum has identified peculiarities of the clinical course of the disease (Andreeva et al., 2016; Marakhonov et al., 2019; Petrova et al., 2019b; Vasilyeva et al., 2021).

In addition to RHDs that are frequent for all populations of the Russian Federation and Europe, regionally specific diseases have been identified. Our study identified diseases endemic to specific regions of the Russian Federation—several



	1:50000 and more	1:50001-1:100000	1:100001-1:200000	1:200001-1:300000	1:300001-1:400001	1:400001-1:500002	1:500001 and less
Patients (%)	59.86%	9.97%	10.50%	6.98%	3.25%	1.21%	8.22%
Diseases (%)	5.96%	4.33%	7.58%	9.75%	6.32%	2.89%	63.18%

FIGURE 1 | Percentage of patients and diseases according to the point prevalence values.

types of hypotrichosis (Kazantseva et al., 2006; Zernov et al., 2016), osteopetrosis (Bliznetz et al., 2009), and a number of RDs identified in the Russian Federation and worldwide in single cases—primary microcephaly (Marakhonov et al., 2018), gnathodiaphyseal dysplasia (Andreeva et al., 2016), metatropic dysplasia (Timkovskaya et al., 2016), etc.

The assessment of the cumulative prevalence and burden of RHDs according to the common medical classification of diseases is important for public health and public policy in Russia and in European countries.

We analyzed the diversity and point prevalence of RHDs according to the conventional classification of the disease by types of affected organs and systems according to their main clinical manifestations—neurological syndrome, ophthalmic syndrome, genodermatoses, skeletal syndrome, hereditary syndrome, and other pathology (hereditary diseases of metabolism, blood, hearing, etc.). **Table 3** presents the number of patients, the number of diseases, and the point prevalence of RHDs distributed according to the medical classification of diseases.

Analysis of **Table 4** shows that the maximum number of patients (23.56%) was found to have neurological and psychiatric pathology, the cumulative point prevalence of which is 65.30/100,000 or 1:1530 people. The next most numerous groups were hereditary syndromes—18.04% (cumulative point prevalence 50.01/100,000 or 1:2,000) and other pathology (hereditary diseases of metabolism, blood, hearing, etc.)—17.11% (47.42/100,000 or 1:2,109 people). For the first two groups of diseases, the maximum incidence was 18.38 and 36.94%, respectively.

Analysis of the age category of patients with RHDs showed that the majority of patients (about 45.3%) belonged to the category of childhood (from newborn to 17 full years), despite the fact that the share of the child population in the regions is only 20.64%; for the reproductive and post-reproductive age, 54.7% of all patients. The results obtained demonstrate the need to develop preventive programs specifically among children.

DISCUSSION

The issue of point prevalence, cumulative prevalence, and burden of RHDs is discussed in many articles and is an urgent problem that is associated with the rapid progress of genetic technology, the identification of thousands of genes, and the problems in society, public health, and social structures of states. According to OMIM statistics, genetic nature has been determined for 6,806 phenotypes so far, and the genetic nature is established for another 200–250 diseases each year (see text footnote 1). Current medical classifications have not kept pace with genetic progress; nosological forms of RHDs are underrepresented in

TABLE 3 | Reported diseases with high point prevalence of 1:100,000 or more and the number of patients (ranging in decreasing number of total point prevalence).

OMIM #	Diagnosis	Number of patients	Point prevalence per 100,000 (including men/women)	Point prevalence per 100,000 children (newborn-17 years old/boys*)
Autosomal dominant RHDs				
Point prevalence 1:50,000 and more				
#146700	Ichthyosis vulgaris	646	17.45 (8.61; 8.83)	34.15
PS130000	Ehlers–Danlos syndrome	459	12.40 (6.13; 6.27)	28.00
#148700	Keratosis palmoplantaris	304	8.21 (3.81; 4.40)	14.52
PS118220	Charcot–Marie–Tooth disease	232	6.27 (3.02; 3.24)	5.23
PS116200	Congenital hereditary cataract	215	5.81 (2.89; 2.92)	15.05
#162200	Neurofibromatosis, type I	199	5.37 (2.67; 2.70)	10.60
PS156200	Undifferentiated intellectual disability	177	4.78 (2.40; 2.38)	9.42
#146000	Hypochondroplasia	164	4.43 (2.08; 2.35)	3.14
PS124900	Deafness, autosomal dominant	133	3.59 (1.50; 2.09)	3.14
178300	Ptosis, hereditary congenital	126	3.40 (1.76; 1.65)	12.04
PS268000	Retinitis pigmentosa	122	3.29 (1.62; 1.67)	4.97
PS174200	Polydactyly, postaxial, type A1	112	3.02 (1.57; 1.46)	8.50
151900	Lipomatosis, multiple	108	2.92 (1.38; 1.54)	0
#154700	Marfan syndrome	105	2.84 (1.40; 1.48)	8.23
PS166200	Osteogenesis imperfecta	100	2.70 (1.32; 1.38)	7.98
#185900	Syndactyly, type I	84	2.27 (1.08; 1.19)	7.46
PS163950	Noonan syndrome 1	72	2.03 (0.85; 1.12)	5.40
Point prevalence 1:500,01–1:100,000				
#160900	Dystrophia myotonica 1	65	1.76 (0.80; 0.95)	1.05
181800	Scoliosis, idiopathic	69	1.86 (0.86; 1.00)	5.10
#133700	Exostoses, multiple, type I	56	1.51 (0.81; 0.70)	4.19
#100800	Achondroplasia	54	1.46 (0.70; 076)	4.19
#143100	Huntington disease	51	1.38 (0.65; 0.73)	0
#120200	Coloboma, ocular	49	1.32 (0.48; 0.85)	4.58
PS310700	Nystagmus, congenital	49	1.32 (0.53; 0.80)	4.58
PS183600	Split-hand/foot malformation 1	45	1.22 (0.23; 0.22)	3.79
PS174400	Polydactyly, preaxial I	44	1.19 (0.57; 0.62)	3.27
PS303350	Spastic paraplegia, autosomal dominant	42	1.13 (0.59; 0.54)	1.57
#110100	Blepharophimosis, ptosis	41	1.11 (0.57; 0.54)	1.83
126070	Albinoidism, oculocutaneous, autosomal dominant	39	1.05 (0.51; 0.54)	0.92
#158900	Facioscapulohumeral muscular dystrophy 1A	39	1.05 (0.49; 0.57)	0.79
PS165500	Optic atrophy 1	38	1.03 (0.57; 0.46)	1.44
#186000	Synpolydactyly 1	37	1.00 (0.46; 0.54)	2.75
#106210	Aniridia	37	1.00 (0.49; 0.51)	3.27
Autosomal recessive RHDs				
Point prevalence 1:50,000 and more				
PS220290	Deafness, autosomal recessive	776	20,96 (9.91; 11.05)	59.27
PS249500	Undifferentiated intellectual disability	431	11.64 (5.40; 6.24)	29.31
PS251200	Microcephaly, primary autosomal recessive	155	4.19 (2.11; 2.08)	17.14
PS268000	Retinitis pigmentosa	150	4.05 (1.92; 2.13)	2.88
#261600	Phenylketonuria	145	3.92 (2.00; 1.92)	15.05
PS116200	Congenital hereditary cataract	105	2.84 (1.40; 1.43)	8.24
#242100	Ichthyosiform erythroderma, congenital	84	2.27 (1.08; 1.19)	5.23
PS253600	Muscular dystrophy, limb-girdle	83	2.24 (1.11; 1.13)	1.70
PS203100	Albinism, oculocutaneous	81	2.19 (1.05; 1.13)	7.85
#253300	Spinal muscular atrophy, types I–III	72	2.03 (1.01; 1.01)	8.23

(Continued)

TABLE 3 | Continued

OMIM #	Diagnosis	Number of patients	Point prevalence per 100,000 (including men/women)	Point prevalence per 100,000 children (newborn-17 years old/boys*)
PS262400	Growth hormone deficiency	72	2.03 (1.02; 1.00)	1.70
Point prevalence 1:50,001–1:100,000				
PS276900	Usher syndrome	52	1.40 (0.68; 0.73)	2.36
PS204000	Leber congenital amaurosis	44	1.19 (0.50; 0.66)	1.44
#248200	Stargardt disease 1	43	1.16 (0.62; 0.57)	1.83
#604379	Hypotrichosis, total, Mari type	39	1.05 (0.52; 0.53)	1.75
#219700	Cystic fibrosis	37	1.00 (0.49; 0.51)	4.45
X-linked RHDs				
Point prevalence 1:50,000 and more				
PS309530	Undifferentiated intellectual disability, X-linked	226	12.21 (12.21; 0)	30.09*
#308100	Ichthyosis, X-linked	124	6.70 (6.70; 0)	15.70*
#306700	Hemophilia A	78	4.21 (4.21; 0)	12.04*
#310200	Muscular dystrophy, Duchenne type	55	2.97 (2.97; 0)	10.21*
PS310700	Nystagmus, congenital, X-linked	51	2.75 (2.75; 0)	11.51*
Point prevalence 1:50,001–1:100,000				
#305400	Faciogenital dysplasia	32	1.73 (1.67; 0)	7.59*
#300376	Muscular dystrophy, Becker type	24	1.30 (1.30; 0)	1.24*
#302800	Charcot–Marie–Tooth disease, X-linked dominant	20	1.08 (0.53; 0.55)	4.71

**Point prevalence for X-linked pathology is estimated for 100,000 boys.*

TABLE 4 | Structure of the diversity and point prevalence of the RHDs in accordance with the main medical classification of diseases.

Types of the hereditary disease	Patient data		Disease data	
	Abs. num. of patients (%)	Point prevalence per 100,000	Num. of diseases	%
Neurological and psychiatric	2418 (23.56%)	65.30	102	18.38%
Ophthalmic	1524 (14.85%)	41.16	73	13.15%
Genodermatoses	1510 (14.71%)	40.78	39	7.03%
Skeletal	1205 (11.74%)	32.54	87	15.68%
Hereditary syndromes	1852 (18.04%)	50.01	204	36.94%
Other pathology (hereditary diseases of metabolism, blood, hearing, etc.)	1756 (17.11%)	47.42	49	8.83%

coding systems (e.g., International Classification of Disease, ICD-11). Data collection from different researchers lacks a unified methodological approach, which makes comparative analysis difficult. Insufficient organization of the process, lack of knowledge, lack of diagnostic expertise, lack of information on point prevalence, distribution of RHDs by medical areas, and cumulative burden of RHDs prevent the full development of a public health strategy. In addition to public health questions, it is a priority to provide medical care for patients with specific RHDs and to identify the most common RHDs for the necessary prioritization and development of regional, national, and global health programs.

This study presents an epidemiological analysis of the point prevalence of RHDs based on actual material based on a total survey of several regions of the European part of Russia. Russia is a multinational country with a population of 146,238,185, which makes it difficult to choose a unified strategy in public health. Fourteen regions (Northern, Central, and Southern Russia) were chosen for the survey and subsequent analysis, both populations with different ethnic extractions, and regions of a single ethnic origin. Selection of different ethnically diverse territories in a multinational country was necessary to be able to identify groups of the most common RHDs, determine point prevalence, and describe the structure of RHDs diversity by medical areas. The study was performed by a single team, and the collection and processing of material remained unchanged throughout the study.

Despite the listed limitations of the method (see section "Features and Limitations of the Method") in our study, every 351 people have a hereditary disease. Remarkably higher values of cumulative prevalence were obtained in children—the proportion of children out of the total number of patients with RHDs was 43.3%. This number is remarkably higher than the proportion of children in the general population (20.64%). The main reason for this age distribution is that up to 70% of RHDs manifest in childhood according to the Orphanet database. However, the distribution according to the inheritance type was uneven: 39.17% with AD pathology, 45.24% with AR and 52.28% X-linked. Moreover, the proportion of child patients (of pre-productive age) varied from 38 to 51% in different surveyed populations (Zinchenko et al., 2019, 2020a). The proportion of patients in the reproductive age (18–45 years) out of the total number of patients with RHDs was 37.2% (from 31.91 to 40% by population); for the post-reproductive age (46 and older), only 17.5% (variation 17–22%) out of all patients (Elchinova et al., 2017). Analysis of the sample showed that some diseases do not occur in older age groups because they have high mortality in childhood and middle age (the rate is not constant and varies depending on the population, the causing gene, and the mutation). The lower point prevalence in the older age group is mostly due to a milder and stable course of a limited number of RHDs, with fitting approaching 1, which reduces their relevance and referral to medical facilities.

The number of familial cases and individual cases in family (including sporadic cases for AD pathology) varied depending on the type of inheritance and the average size of the family in a particular region. Among families with the AD type of inheritance, the familial case rate ranged from 70 to 80% in different populations. In families with the AR type, familial cases were 26–34%; for those X-linked, 15–20%. The distribution of men and women did not differ for autosomal pathology. We assume that due to the limitations detailed in the materials and methods, the data obtained may be an underestimate. However, it should be noted that there was a variation by region and by location. The cumulative prevalence and diversity of RHDs was determined for each region, and the most common RHDs were identified. A general disease registry was then compiled. The analysis showed that a single class of the most common RHDs was identified for all populations, with insignificant variation by region for most diseases. The most significant differences in prevalence were found among the child population for congenital and hereditary diseases with early onset. The largest number of patients was detected in the point prevalence class "1:50,000 and more"—33 clinical forms of RHDs (5.96% of the total number of detected diseases) accounting for 59.86% of the patients. Most of the diseases are from heterogeneous groups of RHDs, and there is locus and allelic heterogeneity by region. The smallest number of patients, 8.22%, was identified in the class of RHDs with a point prevalence of "1:500,001 or less"—350 diseases (63.18% of all identified RHDs). These RHDs were mostly detected in single families.

Similar data were obtained by Wakap et al. (2020) when analyzing the Orphanet database epidemiological data (see text footnote 3) for RDs, not all of which are of monogenic nature, but only some of them (70–80%) (Wakap et al., 2020). Their study showed that out of 5,304 diseases—84.5% have a point prevalence <1/1,000,000, and 77.3–80.7% of the burden of RDs in the population accounts for 4.2% (n = 149) of the diseases. The analysis performed in our study on a real contemporary population demonstrates the rarity of most RHDs affecting single families and highlights the difficulty of detection and diagnosis by physicians.

Wakap et al. (2020) also highlighted that it would be a benefit to perform an analysis of distribution of diseases in real populations according to the medical classification for organization of real medical care. An analysis of the pattern of diversity of RHDs by medical specialties, which has been performed in the current study, showed a predominance of patients (41.60% of patients) with neurological, psychiatric, and hereditary syndromes with a point prevalence of 115/100,000 people. Most diseases of neurological, psychiatric, and hereditary syndromes; hereditary metabolism diseases; and hereditary blood diseases are characterized by reduced fitness, multisystemic lesions, disability, and reduced life expectancy and quality because of the lack of effective treatment. Most hereditary skin, eye, ear, skeletal, and treatable metabolic diseases could affect quality of life (including possible disability) but have adaptation in society and average life expectancy. The findings require a comprehensive public health approach.

CONCLUSION

In conclusion, the present time is characterized by the accumulation of data on the point prevalence of RHDs, and this will help to systematize our knowledge and allow us to develop a strategy of healthcare for patients with RHDs.

MATERIALS AND METHODS

We analyzed long-term studies (1985–2020) on the epidemiology of rare monogenic hereditary diseases in the European part of Russia. This approach allows us to identify disease peculiarities in a given region and provides insight into the diversity of rare hereditary pathologies and their point prevalence.

Surveyed Population

Russia is a multinational state with more than 190 ethnic groups, with Russians constituting about 80%. The study covered various remote areas of European Russia—Central, Northern, Southern, and the North Caucasus. The relative location of the studied regions is shown in **Figure 2**. The selection of survey areas in each region was focused on the indigenous population (the history and migration flows of population formation were studied). **Table 5** shows the populations surveyed, the size of child populations, the ethnic composition of the populations, number of medical organizations (hospital, medical ambulance, paramedic, and obstetric centers), and the number medical workers (physicians, nurses, paramedics) participated in the study. The total size of the surveyed population was 3,703,018 people from 96 rural districts, including 764,260 children (from newborns to 17 full years of age). The sex distribution of the

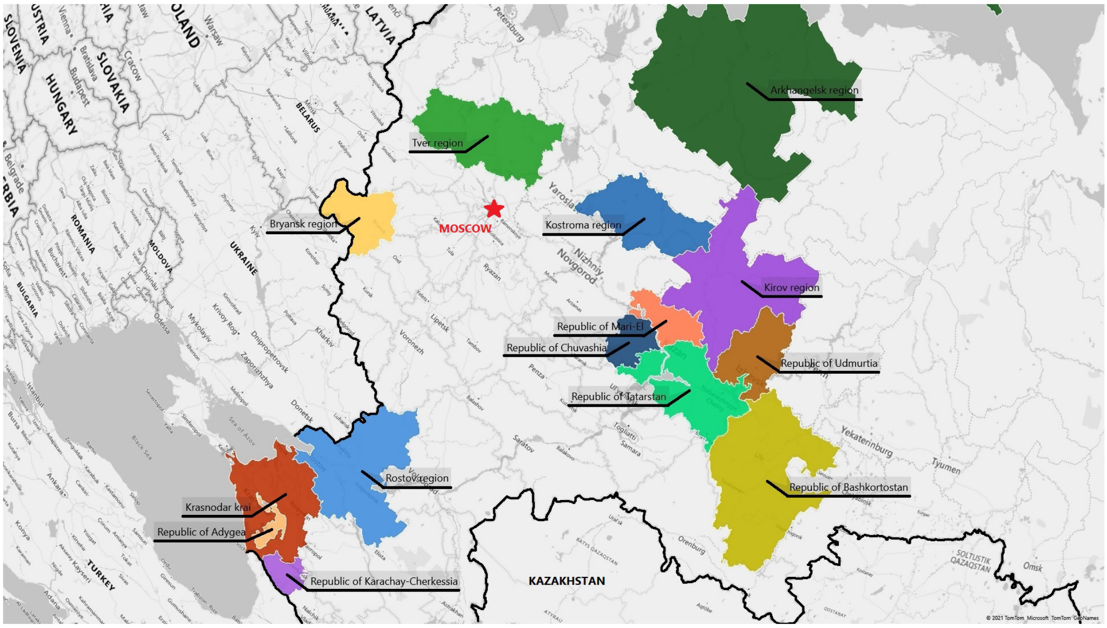


FIGURE 2 | Map of Russian Federation with regions included into the study. The border of the Russia is in bold.

total surveyed population was 1,716,298 (46.35%) men/1,987,087 (53.66%) women[4]. There was an uneven distribution in age: up to 35 years 1:1 (M/W), 36–50 years 0.8:1 (M/W), and the post-reproductive period 0.6:1 (M/W). In total, from the surveyed regions, the study involved 2,181 medical organization and 6,370 medical workers. Each of the studied districts of the particular region has one central district hospital, which includes in-patient and medical ambulance, and several rural ambulances and paramedic–obstetric centers in the villages. The number of rural medical organizations depends on the number of villages in the region and is determined by the Department of Medical Statistics of the hospital. The medical workers permanently residing in the area has full medical information about the attached population.

Survey and Protocol

The survey was carried out by a single team of the Research Centre for Medical Genetics (RCMG) in accordance with the protocol of medical and genetic examination of small populations described previously in detail (Zinchenko et al., 2020b). A survey of the investigated populations was conducted regardless of ethnicity, age, and gender structure. The protocol includes the study of populations through different genetic systems simultaneously: (1) population survey, study of the point and cumulative prevalence of hereditary diseases in the particular population; (2) study of the genetic structure using standard methods of population statistics (analysis of the marriages and migration structure, demographic data, analysis of the frequency of surnames to obtain indicators of random inbreeding through isonymy); and (3) study of the genetic structure through the neutral DNA loci of the nuclear genome.

The survey was conducted in three consecutive steps in accordance with the protocol. At the first step, a lecture course is conducted for all medical workers of each surveyed districts of regions (physicians of different specialties, nurses, paramedics) to explain the goals and objectives of the study. During the lecture, the medical personnel are given an information card–questionnaire (**Figure 3A**) with request for filling in the information on patients (**Figure 3B**), including those at the initial stage of disease and with minimal clinical manifestations[5]. The questionnaire contains easily detectable clinical symptoms of RHDs, almost each of which is characteristic for a group of diseases (isolated, syndromic forms). This protocol allowed the detection of the maximum possible number of nosological forms of RHDs known to date. The card–questionnaire with the listed symptoms allows to reveal practically all "portrait" syndromes. Registration of families presumably with RHDs was done through affected persons in the family. The "Multiple registration" method was used (Cavalli-Sforza and Bodmer, 1999). In addition to the questionnaire card, other sources of information are involved: (i) data on disabled persons (from childhood to adults) in the district provided by the hospital; (ii) information from special schools for the blind and visually impaired, deaf, and hard

[4] https://infotables.ru/statistika/31-rossijskaya-federatsiya/200-sootnoshenie-chislennosti-muzhchin-i-zhenshchin-rossii-v-2010-i-2002-godakh-tablitsa

[5] Transfer of patients' data was performed in accordance with Article 13 of Federal Law No. 323-FL "On the basics of health protection of citizens of the Russian Federation" which allowed the submission of information about patients in the following cases: the exchange of information by medical organizations, including those deposited in medical information systems, for the purpose of providing medical care, taking into account the requirements of the legislation of the Russian Federation (paragraph 8), as well as for the purpose of accounting and control within the mandatory medical insurance system (paragraph 9) [http://docs.cntd.ru/document/902312609].

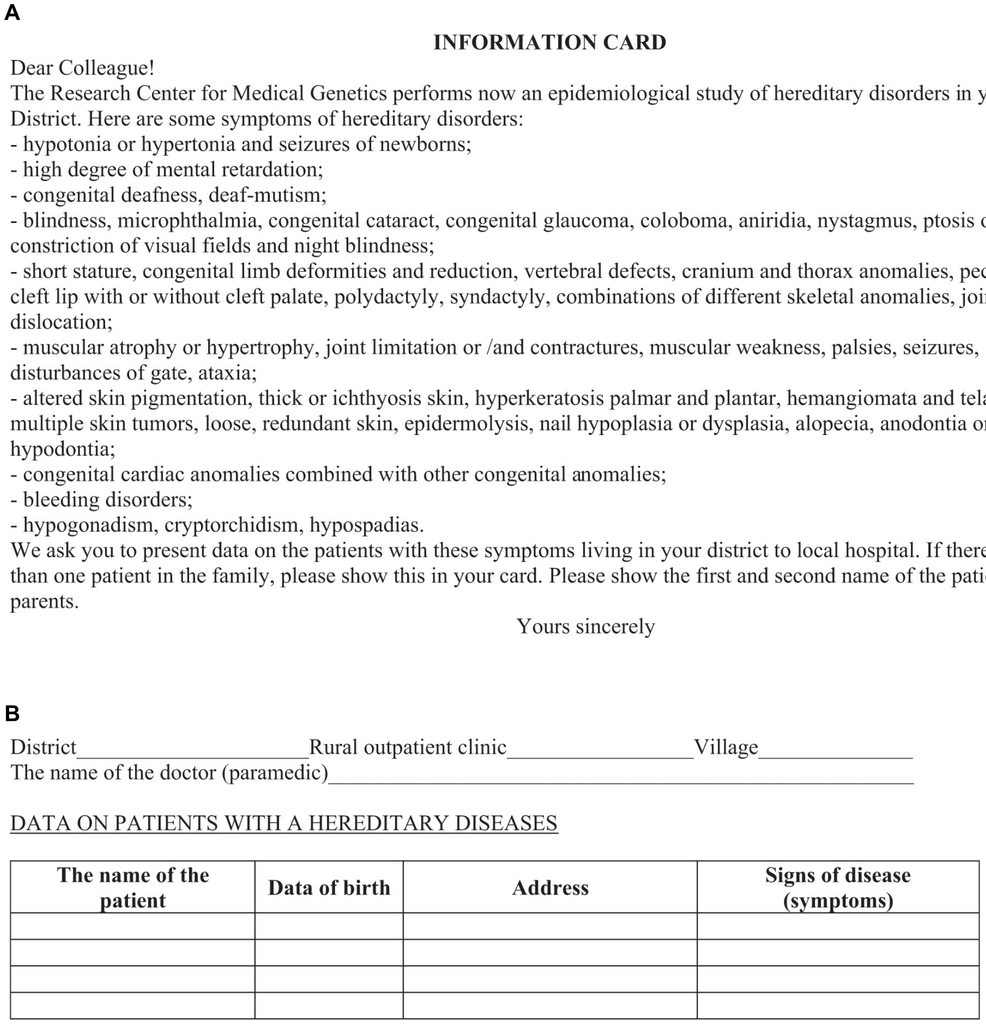
TABLE 5 | Number, ethnic composition of the surveyed populations, number of organization, and medical workers who participated in the study.

Region of the Russian Federation	Size of the region population	Surveyed population/size of children (number of districts)	Main ethnic groups	Number of medical organization which participated in the study	Number medical workers who participated in the study
Central part of Russia					
Kostroma region	637,267	444,476/80,895 (10)	Russians (>90%), other	177	586
Kirov region	1,272,109	286,600/51,051 (11)	Russians (>90%), other	201	619
Bryansk region	1,200,187	88,200/14,906 (1)	Russians (>90%), other	44	64
Tver region	1,269,636	75,000/51051 (2)	Russians (>90%), other	38	81
Republic of Mari El	728,000	276,000/51051 (7)	Maris (62.16%), Russians (32.14%), other (5.7%)	145	271
Chuvash Republic	1,314,000	264,419/67,863 (6)	Chuvashes (67.59%), Russians (25.27%), other (7.14%)	241	504
Republic of Udmurtia	1,570,000	267,655/60,197 (6)	Udmurts (58%), Russians (31.43%), other (10.57%)	272	513
Republic of Tatarstan	3,838,230	264,098/57,648 (8)	Tatars (79.24%), Russians (10.24%), other (10.52%)	253	577
Republic of Bashkortostan	4,093,795	250,110/64,935 (8)	Bashkirs (69.48%), Russians (14.14%), other (16.38%)	255	675
Northern part of Russia					
Arkhangelsk region	1,128,099	40,000/7,440 (5)	Russians (>90%), other	22	172
Southern part of Russia and North Caucasus					
Krasnodar territory	5,124,400	426,600/78,921 (6)	Russians (>90%), other	153	446
Rostov region	4,406,700	497,460/101,845 (12)	Russians (>90%), other	161	582
Republic of Adygea	447,000	112,400/21,581 (4)	Adygeans (57.83%), Russians (36.83%), other (5,34%)	56	124
Republic of Karachay-Cherkessia	470,000	410,368/90,739 (10)	Karachays (39.58%), Russians (32.84%), Cherkess (12.38%), Abazins (8.11%), Nogais (3.59%), other (3.5%)	163	1156
Average	**31,273,391 (21.39%)**	**3,703,018/764,260 (96)**		**2 181**	6370

of hearing and schools for children with intellectual disability; and (iii) data from the genetic counseling unit. Thus, registration of the same patient was possible from several sources of information, i.e., being multiple, but recorded in a single database as one case for further examination. A one-time examination of congenital and hereditary pathology in each region is carried out. In the aggregate, the detection rate of patients from all sources of registration reaches 80%.

At the second step, the patients were examined by clinical geneticists from the RCMG. The patient with the presumed RHDs is given a medical card, which contains personal information of the proband and his family members, a brief medical history, genealogical data, and detailed phenotyping data. In the process of data collection, pedigrees and the possibility of consanguinity were analyzed. Consanguinity was revealed in rare cases. If necessary, a cytogenetic study is conducted to rule out chromosomal abnormalities. In complicated cases, additional examinations were prescribed for patients to be able to verify the diagnoses (biochemical, radiological, electromyographic, and other methods). As a result of this step, a significant proportion of patients (usually, more than half) from different sources is excluded from the sample because of the external cause of the disease (injuries, infections, isolated congenital pathology, etc.). The remainder families represent a list of families presumably with RHDs for further research.

At the third step, clinical investigations were performed by specialists from leading federal research institutes (geneticist, neurologist, ophthalmologist, dermatologist, pediatrician, otolaryngologist, and orthopedist), which ensured unification of diagnostic criteria. In some cases, blood is collected from patients for molecular genetic diagnosis. Written informed consent was obtained from all identified and examined families for voluntary participation in the study. From 1 to 5% of patients refuse to be



A

INFORMATION CARD

Dear Colleague!

The Research Center for Medical Genetics performs now an epidemiological study of hereditary disorders in your District. Here are some symptoms of hereditary disorders:

- hypotonia or hypertonia and seizures of newborns;
- high degree of mental retardation;
- congenital deafness, deaf-mutism;
- blindness, microphthalmia, congenital cataract, congenital glaucoma, coloboma, aniridia, nystagmus, ptosis of eyelids, constriction of visual fields and night blindness;
- short stature, congenital limb deformities and reduction, vertebral defects, cranium and thorax anomalies, peculiar faces, cleft lip with or without cleft palate, polydactyly, syndactyly, combinations of different skeletal anomalies, joints dislocation;
- muscular atrophy or hypertrophy, joint limitation or /and contractures, muscular weakness, palsies, seizures, disturbances of gate, ataxia;
- altered skin pigmentation, thick or ichthyosis skin, hyperkeratosis palmar and plantar, hemangiomata and telangiectasia, multiple skin tumors, loose, redundant skin, epidermolysis, nail hypoplasia or dysplasia, alopecia, anodontia or hypodontia;
- congenital cardiac anomalies combined with other congenital anomalies;
- bleeding disorders;
- hypogonadism, cryptorchidism, hypospadias.

We ask you to present data on the patients with these symptoms living in your district to local hospital. If there is more than one patient in the family, please show this in your card. Please show the first and second name of the patient's parents.

Yours sincerely

B

District_____________________Rural outpatient clinic_________________Village_______________
The name of the doctor (paramedic)____

DATA ON PATIENTS WITH A HEREDITARY DISEASES

The name of the patient	Data of birth	Address	Signs of disease (symptoms)

FIGURE 3 | Information card. Front **(A)** and reverse **(B)** sides.

examined for various reasons. The study was approved by the Ethical Committee of the Research Centre for Medical Genetics (Protocol No. 17/2006 dated 02.02.2006).

Statistical Methods

Given the heterogeneity of many diseases, the type of inheritance was also confirmed by segregation analysis used in multiple-family registration (Cavalli-Sforza and Bodmer, 1999). Segregation frequency is calculated by Weinberg's proband method (Morton, 1959). Using this method, segregation frequencies are calculated (separately for families with AD and AR pathology) by the ratio of probands; for actually detected patients in families after examination, probability of registration. In our case, the calculated segregation frequencies corresponded to the expected ones—0.25 for the AR type of inheritance and 0.5 for AD. However, it should be noted that the probability of registration differed from 100% and was 83% for the group of families with AR pathology and 72% for AD. The results suggest that irrespective of the total screening performed, a certain number of patients might not have been registered by us.

Point prevalence was calculated according to the procedural document on epidemiology of RDs in Orphanet (2019) as the number of reported cases in the population at a given time point per 100,000 people (all age categories were considered) (European Union (EU), 2009). The prevalence of X-linked pathology was calculated for the male population in the surveyed regions. The average male population in the regions was 46.44% (variation 45.6–46.9%).

A nosological registry of detected RHDs based on clinical diagnoses was compiled. For RHD diversity analyses, seven groups were selected with a point prevalence interval of (i) 1:50,000 and more frequently, (ii) 1:50,001–1:100,000, (iii) 1:100,001–1:200,000, (iv) 1:200,001–1:300,000, (v) 1:300,001–1:400,001, (vi) 1:400,001–1:500,002, and (vii) 1:500,001 and less frequently.

Based on clinical manifestations, we additionally analyzed point prevalence according to the generally accepted classification of diseases: neurological syndrome, ophthalmic syndrome, genodermatoses, skeletal syndrome, hereditary syndrome, and other pathology (hereditary diseases of metabolism, blood, hearing, etc.).

The methods for collection and processing of medical genetic material remained unchanged throughout all the studies, which allows comparison of newly obtained data with results from the previously surveyed populations of the country.

Molecular Genetic Analysis

Confirmatory DNA diagnostics was carried out in the laboratories of the RCMG: Laboratory of Genetic Epidemiology (head—R.A. Zinchenko), Laboratory of Epigenetics (head—Sci. V.V. Strelnikov), and Laboratory of DNA Diagnostics (head—A.V. Polyakov). A variety of methods were used for DNA diagnosis—Sanger sequencing, MLPA, RFLP, AFLP, and whole-exome sequencing—depending on the studied nosology according to the protocols published elsewhere by the authors of the current manuscript (Zinchenko et al., 2020b).

Features and Limitations of the Method

Monogenic hereditary diseases listed in the OMIM and Orphanet databases with AD, AR, and X-linked (XL) types of inheritance were included in the analysis. Patients with mitochondrial disorders and chromosomal rearrangements were excluded from the analysis after cytogenetic and molecular genetic studies.

Some patients may not be identified or missed because of the following reasons: (i) patients with subclinical forms of the disease; (ii) patients in the initial stage of the disease with late onset; (iii) patients who refused from examination (from 1 to 5%) for various reasons: observation and treatment by a specific physician, unwillingness to disclose their diagnosis, and others; (iv) patients who are not registered in a medical organization of the region; (v) patients who have not passed this examination and do not live in the region (even in familial cases of the disease); (vi) patients who died by the time of the survey; (vii) child patients under 1 year of age with severe hereditary metabolic diseases unable to pass the survey; and (viii) we also assume that due to the increase in the number of confirmed phenotypes at the genetic level in recent years, the recognition of these diseases by physicians of various specialties is not yet possible. There is a lack of knowledge and diagnostic experience, which must inevitably lead to the omission of sporadic cases of rare hereditary diseases. Doctors in all countries face these problems.

The resulting values of the probability of registration are 83% for the group of families with AR pathology and 72% for AD. We were also unable to estimate the annual incidence (number of newly diagnosed cases in a population within 1 year) due to the lack of this information in medical organizations. In consequence of the above, certain corrections should be made taking into account the limited number of nosological forms envisaged by our study.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: Data are available upon request. Requests to access these datasets should be directed to RZ, renazinchenko@mail.ru.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethical Committee of the Research Centre for Medical Genetics. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

RZ and EG: conceptualization, methodology, visualization, and funding acquisition. RZ, TV, and OA: validation. RZ, NP, VK, AP, and AM: formal analysis. RZ: resources, data curation, and writing—original draft preparation. AM: writing—review and editing. EG and SK: supervision. RZ, EG, and SK: project administration. All authors have read and agreed to the published version of the manuscript.

FUNDING

This study was supported by the Russian Science Foundation, grant no. 17-15-01051 and the state assignment of the Ministry of Science and Higher Education of the Russian Federation.

ACKNOWLEDGMENTS

We thank Richard H. Lozier for interest in our work and useful assistance.

REFERENCES

Andreeva, T. V., Tyazhelova, T. V., Rykalina, V. N., Gusev, F. E., Goltsov, A. Y., Zolotareva, O. I., et al. (2016). Whole exome sequencing links dental tumor to an autosomal-dominant mutation in ANO5 gene associated with gnathodiaphyseal dysplasia and muscle dystrophies. *Sci. Rep.* 6:26440.

Bliznetz, E. A., Tverskaya, S. M., Zinchenko, R. A., Abrukova, A. V., Savaskina, E. N., Nikulin, M. V., et al. (2009). Genetic analysis of autosomal recessive osteopetrosis in Chuvashiya: the unique splice site mutation in TCIRG1 gene spread by the founder effect. *Eur. J. Hum. Genet.* 17, 664–672. doi: 10.1038/ejhg.2008.234

Cavalli-Sforza, L. L., and Bodmer, W. F. (1999). *The Genetics of Human Populations*. Mineola, NY: Dover Publications.

Dadali, E. L., Makaov, A. K., Galkina, V. A., Konovalov, F. A., Polyakov, A. V., Bulakh, M. V., et al. (2016). Hereditary motor and sensory neuropathy, caused by mutations in the NEFL gene in a family from Karachaevo-Cherkessia. *Neuromuscul. Dis.* 6, 47–51. (In Russ.), doi: 10.17650/2222-8721-2016-6-2-47-51

Donnart, A., Viollet, V., and Roinet-Tournay, M. (2013). Les maladies rares, définitions et épidémiologie. *Soins Pédiatr. Puéric.* 34, 14–16.

Elchinova, G. I., Makaov, A. K., Bikanov, R. A., Gavrilina, S. G., Petrin, A. N., Marakhonov, A. V., et al. (2017). Analysis of the age and sex structure of the population of Karachay-Cherkessia Republic. *Mod. Probl. Sci. Educ.* 2:52.

European Union (EU) (2000). Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products. *OJEC* L18, 1–5.

European Union (EU) (2009). Council recommendation of 8 June 2009 on an action in the field of rare diseases (2009/C 151/02). *OJEU* C151, 7–10.

Ferreira, C. R. (2019). The burden of rare diseases. *Am. J. Med. Genet. A* 179, 885–892.

Gundorova, P., Zinchenko, R. A., Kuznetsova, I. A., Bliznetz, E. A., Stepanova, A. A., and Polyakov, A. V. (2018). Molecular-genetic causes for the high frequency of phenylketonuria in the population from the North Caucasus. *PLoS One* 13:e0201489. doi: 10.1371/journal.pone.0201489

Kazantseva, A., Goltsov, A., Zinchenko, R., Grigorenko, A. P., Abrukova, A. V., Moliaka, Y. K., et al. (2006). Human hair growth deficiency is linked to a genetic defect in the phospholipase gene LIPH. *Science* 314, 982–985. doi: 10.1126/science.1133276

Khidiyatova, I. M., Skachkova, I. A., Saifullina, E. V., Magzhanov, R. V., Schagina, O. A., Zinchenko, R. A., et al. (2013). [MFN2 gene analysis in patients with hereditary motor and sensory neuropathy from Bashkortostan Republic]. *Genetika* 49, 884–890. doi: 10.7868/s0016675813060040

Levchenko, O., Dadali, E. L., Bessonova, L., Demina, N., Rudenskaya, G. E., Matyushchenko, G., et al. (2019). Exome sequencing of 100 patients with intellectual disability. *Eur. J. Hum. Genet.* 27, S1390–S1391.

Marakhonov, A. V., Konovalov, F. A., Makaov, A. K., Vasilyeva, T. A., Kadyshev, V. V., Galkina, V. A., et al. (2018). Primary microcephaly case from the Karachay-Cherkess Republic poses an additional support for microcephaly and Seckel syndrome spectrum disorders. *BMC Med. Genomics* 11:8. doi: 10.1186/s12920-018-0326-1

Marakhonov, A. V., Vasilyeva, T. A., Voskresenskaya, A. A., Sukhanova, N. V., Kadyshev, V. V., Kutsev, S. I., et al. (2019). LMO2 gene deletions significantly worsen the prognosis of Wilms' tumor development in patients with WAGR syndrome. *Hum. Mol. Genet.* 28, 3323–3326. doi: 10.1093/hmg/ddz168

Morton, N. E. (1959). Genetic tests under incomplete ascertainment. *Am. J. Hum. Genet.* 11, 1–16.

Murtazina, A. F., Shchagina, O. A., Milovidova, T. B., Dadali, E. L., Rudenskaya, G. E., Kurbatov, S. A., et al. (2020). Clinical and genetic characteristics of Charcot–Marie–Tooth disease type 4D (type Lom) in Russia. *Neuromuscul. Dis.* 10, 39–45. (In Russ.), doi: 10.17650/2222-8721-2020-10-2-39-45

Wakap, S. N., Lambert, D. M., Olry, A., Rodwell, C., Gueydan, C., Lanneau, V., et al. (2020). Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet* 28, 165–173. doi: 10.1038/s41431-019-0508-0

Orphanet (2019). Procedural Document on Epidemiology of Rare Disease in Orphanet (Prevalence, incidence and number of published cases or families). February 2019, Version 01. Available online at: https://www.orpha.net/orphacom/cahiers/docs/GB/Epidemiology_in_Orphanet_R1_Ann_Epi_EP_05.pdf

Pariser, A. R., and Gahl, W. A. (2014). Important role of translational science in rare disease innovation, discovery, and drug development. *J. Gen. Intern. Med.* 29 Suppl 3, S804–S807.

Petrova, N. V., Kashirskaya, N. Y., Saydaeva, D. K., Polyakov, A. V., Adyan, T. A., Simonova, O. I., et al. (2019a). Spectrum of CFTR mutations in Chechen cystic fibrosis patients: high frequency of c.1545_1546delTA (p.Tyr515X; 1677delTA) and c.274G>A (p.Glu92Lys, E92K) mutations in North Caucasus. *BMC Med. Genet.* 20:44. doi: 10.1186/s12881-019-0785-z

Petrova, N. V., Kashirskaya, N. Y., Vasilyeva, T. A., Timkovskaya, E. E., Voronkova, A. Y., Shabalova, L. A., et al. (2016). High prevalence of W1282x mutation in cystic fibrosis patients from Karachay-Cherkessia. *J. Cyst. Fibros* 15, e28–e32.

Petrova, N. V., Marakhonov, A. V., Vasilyeva, T. A., Kashirskaya, N. Y., Ginter, E. K., Kutsev, S. I., et al. (2019b). Comprehensive genotyping reveals novel CFTR variants in cystic fibrosis patients from the Russian Federation. *Clin. Genet.* 95, 444–447. doi: 10.1111/cge.13477

Richter, T., Nestler-Parr, S., Babela, R., Khan, Z. M., Tesoro, T., Molsen, E., et al. (2015). Rare disease terminology and definitions-A systematic global review: report of the ISPOR rare disease special interest group. *Value Health* 18, 906–914. doi: 10.1016/j.jval.2015.05.008

Schagina, O. A., Dadali, E. L., Fedotov, V. P., Osipova, E. V., Zinchenko, R. A., Ginter, E. K., et al. (2007). New allelic variant of hereditary motor and sensory neuropathy in big family from Udmurtia. *Med. Genet.* 6, 33–37. (In Russ.),

Shchagina, O. A., Dadali, E. L., Fedotov, V. P., Milovidova, T. B., Ryzhkova, O. P., and Polyakov, A. V. (2018). Family case of hereditary motor-sensory neuropathy due to the INF2-mutation and hereditary motor and sensory neuropathy with nephrotic syndrome in Russia. *Nevrologicheskii Zhurnal* 23, 121–127. (In Russ.),

Shchagina, O. A., Milovidova, T. B., Murtazina, A. F., Rudenskaya, G. E., Nikitin, S. S., Dadali, E. L., et al. (2020). HINT1 gene pathogenic variants: the most common cause of recessive hereditary motor and sensory neuropathies in Russian patients. *Mol. Biol. Rep.* 47, 1331–1337. doi: 10.1007/s11033-019-05238-z

Timkovskaya, E. E., Makaov, A. H.-M., Mikhailova, L. K., Vasilyeva, T. A., Marakhonov, A. V., Galkina, V. A., et al. (2016). Metatropic dysplasia: clinical and molecular diagnostics, genetic counseling. *Med. News North Cauc.* 11, 173–176.

Vasilyeva, T. A., Marakhonov, A. V., Voskresenskaya, A. A., Kadyshev, V. V., Kasmann-Kellner, B., Sukhanova, N. V., et al. (2021). Analysis of genotype-phenotype correlations in PAX6-associated aniridia. *J. Med. Genet.* 58, 270–274. doi: 10.1136/jmedgenet-2019-106172

Zernov, N. V., Skoblov, M. Y., Marakhonov, A. V., Shimomura, Y., Vasilyeva, T. A., Konovalov, F. A., et al. (2016). Autosomal recessive hypotrichosis with woolly hair caused by a mutation in the keratin 25 gene expressed in hair follicles. *J. Invest. Dermatol.* 136, 1097–1105. doi: 10.1016/j.jid.2016.0.037

Zinchenko, R. A., El'chinova, G. I., Balanovskaya, E. V., and Al, E. (2000). The influence of population genetic structure on the burden of monogenetic hereditary diseases in Russian Populations. *Vestn. Ross. Akad. Med. Nauk.* 5, 5–10.

Zinchenko, R. A., El'chinova, G. I., Baryshnikova, N. V., Polyakov, A. V., and Ginter, E. K. (2007). Prevalences of hereditary diseases in different populations of Russia. *Russ. J. Genet.* 43, 1038–1045. doi: 10.1134/s1022795407090104

Zinchenko, R. A., Elchinova, G. I., Gavrilina, S. G., and Ginter, E. K. (2001a). Analysis of diversity of autosomal recessive diseases in populations of Russia. *Russ. J. Genet.* 37, 1312–1322.

Zinchenko, R. A., El'chinova, G. I., and Ginter, E. K. (2009). Factors determining the distribution of hereditary diseases in Russian populations. *Med. Genet.* 8, 7–23.

Zinchenko, R. A., Elchinova, G. I., Nurbaev, S. D., and Ginter, E. K. (2001b). Diversity of autosomal dominant diseases in populations of Russia. *Russ. J. Genet.* 37, 290–301.

Zinchenko, R. A., Kadyshev, V. V., Galkina, V. A., Dadali, E. L., Mikhailova, L. K., Marakhonov, A. V., et al. (2019). Clinical population genetics of hereditary diseases among children of the karachay-cherkess republic. *Russ. J. Genet.* 55, 1033–1040. doi: 10.1134/s1022795419080180

Zinchenko, R. A., Kadyshev, V. V., Galkina, V. A., El'chinova, G. I., Marakhonov, A. V., Alexandrova, O. Y., et al. (2020a). The load and diversity of monogenic hereditary pathology among the children's population of the Kirov region. *Russ. J. Genet.* 54, 1530–1534. doi: 10.1134/s1022795420120157

Zinchenko, R. A., Makaov, A. K., Marakhonov, A. V., Galkina, V. A., Kadyshev, V. V., El'chinova, G. I., et al. (2020b). Epidemiology of hereditary diseases in the karachay-cherkess republic. *Int. J. Mol. Sci.* 21:325.

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 MM declared a past co-authorship with the authors RZ, EG, AM, NP, TV, and SK to the handling editor.

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

REVIEW
published: 17 September 2021
doi: 10.3389/fgene.2021.709388

A Role of Variance in Interferon Genes to Disease Severity in COVID-19 Patients

*Leonid Gozman[1], Kellie Perry[2], Dimitri Nikogosov[3], Ilya Klabukov[4], Artem Shevlyakov[3] and Ancha Baranova[2,3,5]**

[1]Sackler School of Medicine, Tel Aviv University, Ramat Aviv, Israel, [2]School of System Biology, George Mason University, Fairfax, VA, United States, [3]Atlas Biomed Group Limited, London, United Kingdom, [4]Department of Regenerative Technologies and biofabrication, National Medical Research Radiological Center of the Ministry of Health of the Russian Federation, Obninsk, Russia, [5]Research Center for Medical Genetics, Moscow, Russia

The rapid rise and global consequences of the novel coronavirus disease 19 (COVID-19) have again brought the focus of the scientific community on the possible host factors involved in patient response and outcome to exposure to the virus. The disease severity remains highly unpredictable, and individuals with none of the aforementioned risk factors may still develop severe COVID-19. It was shown that genotype-related factors like an ABO Blood Group affect COVID-19 severity, and the risk of infection with SARS-CoV-2 was higher for patients with blood type A and lower for patients with blood type O. Currently it is not clear which specific genes are associated with COVID-19 severity. The comparative analysis of COVID-19 and other viral infections allows us to predict that the variants within the interferon pathway genes may serve as markers of the magnitude of immune response to specific pathogens. In particular, various members of Class III interferons (lambda) are reviewed in detail.

Keywords: COVID-19, interferons, SARS-CoV-2, signaling, type I interferon, differential activity

Edited by:
Ranajit Das,
Yenepoya University, India

Reviewed by:
Eaaswarkhanth Muthukrishnan,
Dasman Diabetes Institute, Kuwait
Evgenii Chekalin,
Michigan State University,
United States

****Correspondence:***
Ancha Baranova
abaranov@gmu.edu

Specialty section:
This article was submitted to
Evolutionary and Population Genetics,
a section of the journal
Frontiers in Genetics

Received: *13 May 2021*
Accepted: *01 September 2021*
Published: *17 September 2021*

Citation:
Gozman L, Perry K, Nikogosov D, Klabukov I, Shevlyakov A and Baranova A (2021) A Role of Variance in Interferon Genes to Disease Severity in COVID-19 Patients.
Front. Genet. 12:709388.
doi: 10.3389/fgene.2021.709388

INTRODUCTION

Coronavirus disease 2019 (COVID-19) is a rapidly emerging infectious disease caused by SARS-CoV-2 virus, a member of the Coronaviridae family. Since the discovery of first cases in December 2019 in Wuhan, China (World Health Organization, 2020a), the number of infected patients worldwide has been increasing logarithmically, and by December 2020 had surpassed 70 million reported cases and over 1.5 million deaths globally since the start of the pandemic (World Health Organization, 2020b).

According to WHO, mild to moderate respiratory symptoms such as fever, dry cough, upper airway congestion and sore throat are among the most common symptoms of COVID-19 (World Health Organization, 2020c; World Health Organization, 2021) which may develop over the course of 2 weeks after the exposure. However, approximately 20% of the patients develop severe or critical COVID-19 (Wu and McGoogan, 2020), characterized by pneumonia and acute respiratory distress syndrome which require hospitalization. While the overall mortality of severe COVID-19 is estimated between 1 and 4% (Ruan, 2020), in-hospital mortality in severe cases is substantially higher, reaching 28–62%, and even surpassing that in patients requiring mechanical ventilation (Weiss and Murdoch, 2020).

Multiple studies have been performed to establish the factors influencing the susceptibility to, severity and mortality of COVID-19 (Du et al., 2020; Gong et al., 2020; Vardavas and Nikitara, 2020;

Verity et al., 2020). It has been shown that the severe or critical course of the disease is more likely in older adults, especially those with underlying health conditions (Centers for Disease Control and Prevention, 2020a; Garg et al., 2020), as 80% of deaths associated with COVID-19 were among adults aged 65 years or older, or those with severe comorbidities (Centers for Disease Control and Prevention, 2020b). Other proposed risk factors include smoking (Vardavas and Nikitara, 2020) and blood type (Zhao et al., 2020).

However, disease severity remains highly unpredictable, and individuals with none of the aforementioned risk factors may still develop severe COVID-19. This is particularly evident in the demographics of the disease in the United States, where even in the early stages of the outbreak 20% of the hospital admissions and 12% of the ICU admissions were attributed to people aged 20–44 years (Centers for Disease Control and Prevention, 2020b). Coupled with the first reported case of identical twins both dying of severe COVID-19 (BBC, 2020) within several days of each other, as well as data obtained in pilot studies on the heritability of COVID-19 symptoms (Williams et al., 2020), it strongly suggests that inherited DNA variants play a significant role in the severity of the disease.

ACE2 and Viral Entry

The rapid rise and global consequences of the novel coronavirus disease 19 (COVID-19) has brought the focus of the research on the possible contribution of the host factors to patient response and outcome of exposure to the virus. SARS-coronavirus 2 (SARS-CoV-2), the pathogenic cause of the disease, relies on similar mechanisms of cellular entry as SARS-CoV; namely, the SARS-CoV-2 receptor angiotensin-converting enzyme II (ACE2) and the serine protease TMPRSS2, which facilitates the priming of spike (S) protein for viral entry (Hoffmann et al., 2020; Zhou et al., 2020).

Early on during the rise of the pandemic, there was a hope that variations in the *ACE2* gene may account for resistance or susceptibility to COVID-19 in different populations. It was shown that populations in East Asia had higher allele frequencies in the expression quantitative trait loci (eQTL) in *ACE2,* which might have led to increased expression of the enzyme (Cao et al., 2020). Likewise, given that *ACE2* is located on chromosome X, a hope was expressed for explaining the gender differences in response to the disease, namely the fact that men were disproportionately more susceptible to the SARS-CoV-2 virus (Majdic, 2020).

Early studies attempting to connect variation in *ACE2* and *TMPRSS2* loci on the risks of contracting COVID-19 in any form, so far, have produced inconclusive results, ranging from single SNP associations uncovered in small cohorts (Latini et al., 2020) to the lack thereof. The latter, possibly, is due to the dual role of ACE2, which serves both as an entry into the wells and a lung-protective molecule (Dalan et al., 2020; Nagy et al., 2021).

A large study by Lopera Maya et al. (2020) used the Lifelines cohort data to analyze the association between the variants within *ACE2* or *TMPRSS2* loci and cardiac, pulmonary, renal and other quantitative phenotypes, which are also pertinent to COVID-19. Despite the large sample size and wide variety of variants and quantitative phenotypes examined, no statistically significant association was detected. The study found, however, an intriguing association between the use of angiotensin receptor blockers (ARBs) and non-steroidal anti-inflammatory drugs (NSAIDs) and variants at the *ACE2* and loci. As the diseases associated with the use of these medications are commonly comorbid with COVID-19 (Lopera Maya et al., 2020), these findings may eventually prove their relevance to COVID-19 severity. Drugs based on the inhibition or blockage of TMPRSS2 protease are undergoing clinical trials as a therapeutic option for COVID-19 treatment (Abbasi et al., 2021). Thus, there remains continued interest in studying the *ACE2* and *TMPRSS2* genes as determinants of susceptibility to SARS-CoV-2.

ABO Blood Group and Disease Severity

After the SARS-CoV outbreak in Hong Kong in 2003, researchers showed a relationship between the blood type of the participants that had been exposed to the virus and the chance of contracting infection (Cheng et al., 2005). It appeared that exposed individuals with blood group O phenotype were less susceptible to SARS-CoV infection, even while previous studies had shown that they had increased susceptibility to infection with either Norwalk virus or *H. pylori* (Cheng et al., 2005). The association of ABO blood groups with the risk to contract coronavirus disease has also been noted during the current pandemic. Retrospective study conducted in China showed that patients with blood group O had a significantly lower risk of infection and hospitalization with SARS-CoV-2, while patients with blood group A had a higher risk of infection and hospitalization (Li et al., 2020).

Further research has both confirmed this association and shed more light on it. Retrospective studies conducted in various regions of China, New York, Italy, Spain, and Turkey have shown a higher odds ratio for being infected with SARS-CoV-2 for patients with blood type A phenotype as well as a lower one for blood type O patients when (Focosi, 2020). A genome wide association study conducted in Italy and Spain regarding the genetic associations between individuals infected with COVID-19 and respiratory failure, confirmed that patients with blood group A had a higher risk of COVID-19-induced respiratory failure while blood group O granted patients a protective effect (The Severe Covid-19 GWAS Group, 2020). Two loci with a genome-wide significance were found, namely, the rs11385942 insertion-deletion GA at locus 3p21.31 and the rs657152 A at locus 9q34.2. The association signal at 9q34.2 coincided with the *ABO* locus, further implicating the connection between patient's ABO blood group and the course and danger of the disease (The Severe Covid-19 GWAS Group, 2020). Later, a multicenter study performed in Canada showed that COVID-19 patients with blood group A or AB are at increased risk for requiring mechanical ventilation and prolonged ICU admission compared with patients with blood group O or B (Hoiland et al., 2020), thus, supporting *in silico* GWAS results by patient's ward observations.

Interestingly, the viral infectivity features due to the ACE2 receptor binding, and due to contribution of the blood antigens may be related to each other. In case of SARS-CoV, the presence of anti-A antibodies, which is a characteristic of groups O and B, inhibits the adhesion of the virus to the ACE2 receptor (Guillon et al., 2008). It is tempting to speculate that this finding may be directly relevant to SARS-CoV-2 as well, given that these findings are consistent with the host response to other viruses such as measles and HIV (Arendrup et al., 1991; Preece et al., 2002) and a trend in increased efficiency of the transfusion of the convalescent plasma from O or B group donors (Hacibekiroğlu et al., 2021).

3p21.31 Locus

In addition to findings reported from Italy/Spain (The Severe Covid-19 GWAS Group, 2020), a separate study comprising 3,199 hospitalized patients with COVID-19 and control individuals was released by the COVID-19 Host Genetics Initiative in which the region on chromosome 3 was the only major genetic risk factor for severe symptoms after SARS-CoV-2 infection and hospitalization at the genome-wide level (The COVID-19 Host Genetics Initiative, 2020). It is not clear which specific gene within the region identified on chromosome 3 is associated with COVID-19 severity. In particular, this region harbors *CXCR6* and *CCR1* genes, encoding important chemokines, which control the movement of immune cells and are critical for the innate immune system to function properly (Sokol and Luster, 2015). Another gene of this region, *SLC6A20*, encodes a protein that functions as a proline transporter expressed in alveolar cells, kidney and small intestine (SIT1), which is known to bind to ACE2 (Camargo et al., 2009; Vuille-dit-Bille et al., 2015; Wang et al., 2020). Notably, the entire fragment may have been inherited from Neanderthals, entering the human genome during the period of interbreeding between the two groups (Zeberg and Pääbo, 2020), and is differentially represented in human population samples.

Other Loci

Interferons (IFNs) are central to antiviral immunity. Previously it was shown that type I IFN deficiency in the blood could be a hallmark of severe COVID-19 and provide a rationale for combined therapeutic approaches (Hadjadj et al., 2020).

Additional studies have shown the importance of other loci in determining the genetic susceptibility of hosts to COVID-19, particularly in determining which patients are susceptible to severe manifestations of the illness. A recent study of patients with life-threatening COVID-19 pneumonia looked at thirteen loci involved in either the TLR3 or IRF7 dependent pathways for the amplification of type I IFN, and found that 3.5% of patients had deleterious variants (pLOF) in eight of the tested loci, underlining how impairment of the production of type I IFNs can lead to critical SARS-CoV-2 infection (Zhang et al., 2020). Similarly, a recent study of critically-ill COVID-19 patients in the United Kingdom used Mendelian randomization to show the potential for a causal relationship between the *IFNAR2* gene which codes for a receptor subunit in interferon signalling and disease severity, and concluded likewise that the administration of interferons may aid in patient recovery, while acknowledging that it is as yet unclear when during the course of the illness they may provide therapeutic benefit (Pairo-Castineira et al., 2021). Moreover, the study was able to replicate the results of a previous study on the 3p21.31 locus, and a transcriptome-wide association study that they performed on the patient pool showed that the variant in oligoadenylate synthetase (*OAS*), rs10735079 affected expression of *OAS3*, which codes mediator involved in the degradation pathway of double-stranded RNA, which is itself involved in the replication pathway of coronaviruses (Pairo-Castineira et al., 2021).

ACE2 as an Interferon-Responsive Gene

Notably, in humans, the *ACE2* belongs to a family of interferon-stimulated genes (ISGs), which typically serve to promote a complex and uniform response to an infection-related spike in interferon levels (Schneider et al., 2014). Moreover, in human lung epithelial cells, the levels of *ACE2* mRNA are co-correlated with that of TMPRSS4, and many immune response pathways, including proinflammatory interleukins and IFI16 (Wruck and Adjaye, 2020). Specifically in human nasal epithelial cells, *ACE2* expression is upregulated by type I (IFN-α and IFN-β) and type II (IFN-γ) interferons (Ziegler et al., 2020). The efficacy of this process may be affected by genetic variations in any part of this cascade. However, in this review, we would like to bring attention to a particular component of the interferon response, which has been massively implicated in the natural and therapeutic outcomes for other viral diseases, namely, the *IFNL4* gene. This gene encodes a type III interferon IFN-λ4, capable of blocking some of the interferon signaling, resulting in poor response to HCV treatment with IFN (Sung et al., 2017). Notably, type III IFNs have been proposed as more viable therapeutic option for prevention and treatment of COVID-19 than type I IFNs, particularly because they cause fewer and milder systemic side effects (Muir et al., 2014; Prokunina-Olsson et al., 2020). It has also been shown that Type III IFNs are highly effective at preventing the viral spread from the nasal epithelium to the upper respiratory tract (Klinkhammer et al., 2018). Additional studies may be warranted to explore the mechanisms of interaction between SARS-CoV-2 and type III interferons, and to estimate how they are affected by the status of the IFNL4 gene.

It is anticipated that influence of ACE2 in COVID-19 can potentially be exploited for the rational design of effective SARS-CoV-2 therapeutics (Ni et al., 2020; Barros et al., 2021).

Role of Human Interferons in Viral and Non-Viral Liver Disease

Interferons are a class of cytokines that mediate the host immune response to infection by viral and non-viral pathogens (Crosse et al., 2017; Bogdan et al., 2004; Seliger et al., 2008). They are categorized into three types based on their protein sequence (O'Brien et al., 2014) (**Table 1**). Type I interferons are rapidly produced when viral envelope glycoproteins, CpG DNA, or dsRNA interact with host cell receptors such as mannose receptors, toll-like receptors, and cytosolic receptors (Malmgaard, 2004). Type 1 interferons can directly activate natural killers (NKs), antigen-presenting dendritic cells as well as CD4 and CD8 T cells (Hervas-Stubbs et al., 2011). All type I

TABLE 1 | Classification of interferons.

Interferon (IFN)Type	Receptor type	Protein structure	Genes	Gene location	Tissue expression pattern
Type I	IFN α receptor that consists of IFNAR1 and IFNAR2 chains	α–helix	IFN-α 2a and 2b	Chr. 9	Leukocytes, macrophages, endothelial cells, tumor cells, keratinocytes, and mesenchymal cells
			IFN-b		Fibroblasts, endothelial cells, macrophages, and epithelial cells
			IFN-ω		T lymphocytes
			IFN-ε		Cerebral tissues
			IFN-κ		Not known
Type II	IFNGR consisting of IFNGR1 and IFNGR2 chains	Core of six α– helices and an extended unfolded sequence in the C-terminal region	IFN-γ	Chr. 12	T and Natural Killer cells
Type III	Receptor complex consisting of IL10R2 and IFNLR1 chains	Structurally similar to the IL-10 family, despite functionally being an IFN	IFN-λ		Dendritic cells and macrophages

interferons signal through a common receptor interferon alpha receptor (IFNAR). The IFNAR induces the Janus activated kinase-signal transducer and activation of transcription (JAK-STAT) pathway that control a large collection of genes through regulated expression of various signaling intermediaries (Guan et al., 2014; Messina et al., 2016; Olex et al., 2016).

Type I interferons are rapidly produced when viral factors, such as envelope glycoproteins, CpG DNA, or dsRNA, interact with cellular pattern-recognition receptors (PRRs), such as mannose receptors, toll-like receptors (TLRs), and cytosolic receptors (Malmgaard, 2004). These interferons directly activate natural killers (NKs), antigen-presenting dendritic cells (DC) as well as CD4 and CD8 T cells (Hervas-Stubbs et al., 2011). In T cells, the signaling through the IFNAR is critical for the acquisition of effector functions (Kole et al., 2013).

Type II interferons are represented by pleiotropic Th1-type cytokine interferon-γ. The IFN-γ is induced in response to a variety of cytokines, including interleukin-2 (IL-2), IL-18, Type I IFNs alpha/beta, or by stimulation through T cell receptors (TCRs) or NK cell receptors (Malmgaard, 2004). Similar to Type I interferons, IFN-γ stimulates the JAK/STAT pathway. In addition, a number of other pathways, including MAP kinase, PI3-K, CaMKII, and NF-kappaB cross-talk with JAK-STAT signaling to fine-tune the multifaceted effects of IFNγ, which are exerted in a gene- and cell type-specific manner (Gough et al., 2008).

The type III family of interferons are comprised of IFN-λ1, IFN-λ2, and IFN-λ3 or IL-29, IL-28A, and IL-28B, respectively (Kotenko et al., 2003; Gad et al., 2009; Lin and Young., 2014; O'Brien et al., 2014). These interferons signal through a receptor complex composed of the IFN-λR1 chain (also known as IL-28RA) and the IL-10R2 chain, which is also a part of the receptor complexes for IL-10, IL-22, and IL-26. (Sheppard et al., 2003; Gad et al., 2009; Donnelly and Kotenko, 2010; Lopušná et al., 2013).

In 2013, a new member of the interferon λ (lambda) family, IFN-λ4, was described which signals through the IFNλR1 and IL-10R2 receptor chains (Hamming et al., 2013). The IFN-λ4 is encoded by the gene IFNL4, whose expression has been shown to be upregulated in response to HCV infection, but not to HBV infection (Estep et al., 2014).

Recent studies point that IFN dysregulation may be the key to determining COVID-19 pathogenesis (Andreakos and Tsiodras, 2020; Lopez et al., 2020; Meffre and Iwasaki, 2020). There is evidence that the response to class I interferons in COVID-19 is impaired. In the blood of patients with severe COVID-19, amounts of class I IFNs are much lower when compared to that of patients infected with highly pathogenic influenza viruses. Nevertheless, in the lungs, in bronchoalveolar lavage in some seriously ill COVID-19 patients, local induction of IFN genes becomes noticeable. A dysregulated interferon response is considered part of the immunomodulatory strategies used by some coronaviruses, including SARS-CoV-2 (Acharya et al., 2020). Nevertheless, a recent pan-ancestry exome-wide association study of rare genetic protein-coding variants and various t COVID-19 outcomes didn't find any significant associations in any of the 13 interferon pathway genes (Kosmicki et al., 2021).

Since the beginning of the pandemics, interferons were repeatedly seen as a viable option for boosting the host's defences against SARS-CoV-2. Indeed, early evidence suggests that SARS-CoV-2 may be more susceptible to pretreatment with type I IFNs, even more so than SARS-CoV (Lokugamage et al., 2020; Sallard et al., 2020). Later, in human intestinal cells, the treatment with interferon-lambda and respective responses showed efficiency at controlling SARS-CoV-2 replication (Stanifer et al., 2020). In this light, a renewed attention was paid to type III interferons, which have being tested as therapeutics in COVID-19 outpatients, either with no success (Jagannathan et al., 2021) or with limited virological response detected (Feld et al., 2021). The difference in outcomes of the interferon-lambda based therapeutics may be explained by the varied presence of neutralizing IFNL3 autoantibodies pre-existing in patients that later develop severe COVID-19 (Credle et al., 2021).

The *IFNL4* Locus

The *IFNL4* gene is located on chromosome 19q13, just over 1 kb upstream of, and in the same orientation as, the gene encoding IFN-λ3 (**Figure 1**). It is extremely conserved in all mammals, indicating its functional importance (Key et al., 2014). The

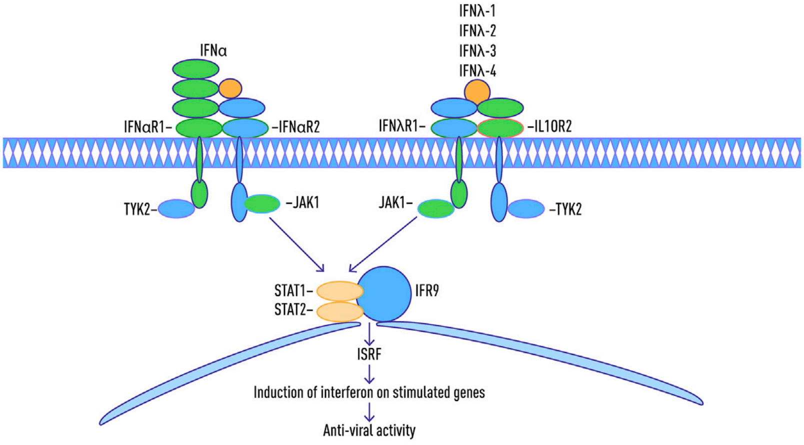


FIGURE 1 | Location of common SNPs in IFNL4 Locus on Chromosome 19, and the map of IFNL4 exons. Adapted from: Stephen M. Laidlaw and Lynn B. Dustin, 2014, with changes.

ancestral allele of *IFNL4,* contains a guanine residue at position 342 of the coding sequence (referred to as "ΔG"). It encodes a functional IFN-λ4 peptide.

The *IFNL4* locus is known to contain a number of medically relevant single nucleotide polymorphisms (SNPs). One of these, rs368234815 or ss469415590 (TT) is characterized by the substitution of the G nucleotide with two thymine residues (TT) resulting in a nonsense mutation. As a result, IFN-λ4 can be generated only by individuals, who carry the ΔG allele, To date, the majority of the studies of *IFNL4* locus have been performed in the context of hepatitis C virus (HCV). In contrast to other IFNs, expression of IFNλ4 is associated with decreased clearance of HCV in the human population; by contrast, a natural frameshift mutation that abrogates IFNλ4 production improves HCV clearance. The ΔG allele is associated with adverse outcomes of infection and interferon-based treatments (Amanzada et al., 2013; Franco et al., 2014 AIDS; Aka et al., 2014; Nozawa et al., 2014; Jouvin-Marche et al., 2014; Stättermayer et al., 2014) while the TT allele is associated with the spontaneous clearance of HCV and interferon responsiveness. It is presently the strongest known host factor for predicting clearance of HCV (O'Brien et al., 2014). Another SNP, known as rs12979860, located within the intron of IFNL4 gene, is closely linked to the rs368234815 allele and is significantly associated with sustained viral response (SVR) in HCV patients (Younossi et al., 2012).

In a genome-wide association study published in 2009, the presence of a rs12979860 with a "C" allele was strongly associated with spontaneous viral clearance and treatment response (Ge et al., 2009). Patients who were homozygous for the presence of "C" allele had a greater than 2-fold increase in rates of SVR as compared to patients with heterozygosity of this locus (C/T allele combination) and homozygous state T/T (Younossi et al., 2012; Meissner et al., 2014; Stättermayer et al., 2014). In addition to increased SVR rates, patients homozygous for C allele (C/C) were more likely to demonstrate spontaneous clearance of HCV (Thomas et al., 2009). Additionally, the presence of SVR-promoting rs12979860 allele of *IL28B* locus was associated with lower baseline inflammation and possible suppression of apoptosis in peripheral blood mononuclear cell (PBMCs) evaluated during early phase of the treatment as compared to the presence of deleterious allele (Younossi et al., 2012). These findings were confirmed in the 2013 GWAS performed in 13 international multicenter study sites (Duggal et al., 2013).

Interpretation of these findings relies on the proximity linkage of rs12979860 (*IL28B*) to rs368234815 (*IFNL4*) that is functionally responsible for effects of both variants. Due to shorter average size of haplotype blocks in individuals of African ancestry, rs368234815 is more strongly associated with HCV clearance in these ethnicities, whereas in Europeans and Asians it performs similarly to rs12979860 (Prokunina-Olsson et al., 2013).

Non-functional rs368234815-TT allele is specific for humans and is common in all human populations. In HapMap collection, it is detected in 93% of Asians genomes, 68% of European genomes, and 23% of Africans genomes (Prokunina-Olsson et al., 2013).

Similar frequencies of distribution were observed in the 1000 Genomes Project samples: the TT allele is present in 89.8–95.2% of Chinese genomes, in 68.9% of European genomes and in 29.3% of African genomes (The 1000 Genomes Project Consortium, 2015).

Linkage disequilibrium between rs368234815-TT allele and rs12979860-C allele results in the same frequencies distribution pattern for the latter genetic variant: C allele is present in 89.8–95.2% Chinese genomes, in 69.1% of European genomes and in 33.1% of African genomes (The 1000 Genomes Project Consortium, 2015). Frequency of the rs12979860 C/C genotype in IFNL4 gene was significantly lower in COVID-19 patients ($p <$ 0.001) (Saponi-Cortes et al., 2021).

Linkage between these two genetic variations rises from Africans (R^2 = 0.8318) through Europeans (R^2 = 0.9815) and is absolute in Chinese populations (R^2 = 1.0) (Machiela and Chanock, 2015).

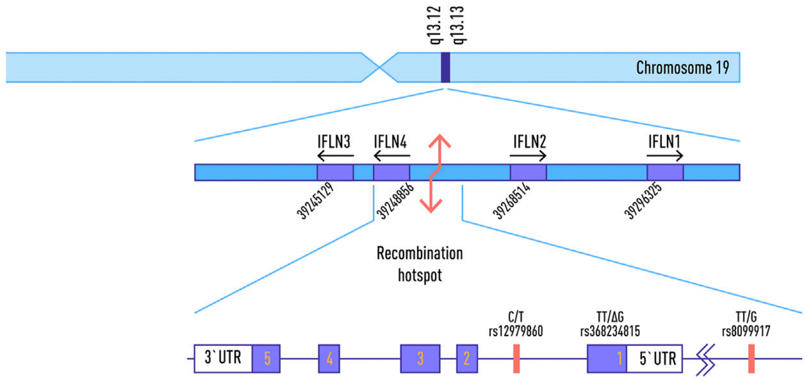


FIGURE 2 | Schematic map of Jak-STAT pathway during an immune response with type 1-3 interferon antiviral activity.

It is theorized that it emerged right before the onset of the “out-of-Africa” migration and was immediately supported in its spread by positive selection in European and Asian populations resulting in the high frequency observed today (Key et al., 2014).

It is most likely that the selective force that was driving elimination of IFN-λ4 in a majority of human populations was an exposure to a certain pathogen (or pathogens), most likely a virus. However, this pathogen is unlikely to be HCV, which is known for its relatively slow progress toward symptomatic phase (Fan et al., 2016). There is also no association between *IFNL4* polymorphisms and HBV susceptibility or natural clearance (Fan et al., 2016) and the advantages or disadvantages of IFN-λ4 expression in case of infection with a majority of non-HCV non-HBV viruses remain unknown.

A number of recent studies showed that IFN-λ4 possesses strong antiviral activity toward HCV and coronaviruses (Hamming et al., 2013; Prokunina-Olsson et al., 2013). When over-expressed in a hepatoma cells, *IFNL4* induces STAT1/STAT2 phosphorylation and expression of interferon-stimulated genes (Prokunina-Olsson et al., 2013; O’Brien et al., 2014; Randall and Goodburn, 2008; Ank et al., 2006) (**Figure 2**). Interestingly, when studied against either HCV or coronavirus (HCoV-229E and MERS-CoV) challenges tested in either human ciliated airway epithelial cell (HAE) or hepatocyte cultures, the antiviral activities of recombinant IFNλ3 and IFNλ4 were similar (Hamming et al., 2013). Another recent comparative study of Type III interferons, this time performed using transcriptome sequencing, also failed to reveal any crucial differences between particular members of this family, with the majority of the identified genes being similarly regulated in hepatocytes as well as airway epithelial cells (Lauber et al., 2015). Hence, it looks like the differences in mode of action for various IFN-λ may be due to their direct binding to some cellular or viral targets rather than to the transcription programs they stimulate.

IFNLR is expressed at relatively high levels in respiratory epithelial cells, and mice treated with IFN-λ prior to infection with human metapneumovirus (HMPV) develop lower viral titers and reduced inflammatory responses. On the other hand, Ifnlr1 −/− mice exhibit increased susceptibility to respiratory viral infections, including influenza virus, HMPV, respiratory syncytial virus, and SARS coronavirus (Lazear et al., 2015).

In contrast, Prokunina-Olsson proffered the hypothesis that functional IFN-λ4 protein may compete with the IL28B/IFN-λ3 receptors and apparently cause a pre-activation of the interferon-dependent genes, thus, reducing overall responsiveness to Type I and III interferon (Prokunina-Olsson et al., 2013). This hypothesis is a good agreement with previous findings that SVR-promoting alleles are associated with lower baseline inflammation (Younossi et al., 2012). Notably, in a small study of rs12979860 allele distribution in COVID-19 patients and controls, the “C” allele, previously associated with favourable HCV outcomes and lower baseline inflammation, showed association both with higher susceptibility to coronavirus and with poorer outcomes of SARS-CoV-2 disease (Agwa et al., 2021).

There is some evidence that the polymorphisms in IFNλ4 may influence outcomes of non-HCV non-coronavirus types of acute and chronic infections. In particular, solid-organ transplant recipients homozygous for the active, ancestral rs368234815 allele (ΔG) are more susceptible to CMV replication, especially in absence of antiviral prophylaxis (Egli et al., 2014; Manuel et al., 2015). Another study showed that the same allele is associated with increased susceptibility to AIDS-related CMV retinitis (Bibert et al., 2014).

Findings related to IFN-lambda gene variants in patients with HIV infection remain controversial. One study showed that, in Caucasian populations, the CC genotype of rs12979860, which is associated with favourable HCV outcomes, is also associated with spontaneous control of human immunodeficiency virus (HIV) viremia (Machmach et al., 2013). In cohorts of African Americans these findings, however, were not replicated (Sajadi et al., 2011; Salgado et al., 2011). In a study of Real and co-authors, pseudogenized allele rs368234815-TT that protects against infection with HCV was also associated with decreased likelihood of HIV-1 infection in male intravenous drug users [odds ratio (OR): 0.3; p = 0.006], and this association was not modified by the genotype of CCR5 (Real et al., 2015). Another recent study of rs368234815 variant showed that carriers of its active, ancestral variant ΔG have a higher occurrence of AIDS-defining illnesses and lower CD4 T-cell counts (Machmach et al., 2015). These results suggest that genetic susceptibility to HCV and HIV-1 infection may share a common molecular pathway (Real et al., 2015).

It should be noted that the relationships between pre-existing HIV infection and COVID-19 are still unclear (Centers for Disease Control and Prevention, 2020b), most likely due to limited cross-testing (Jones et al., 2020). While the use of protease inhibitors such as lopinavir and ritonavir had a positive effect on patients with MERS-CoV, recent research suggests that in patients with SARS-CoV-2 these compounds do not work (Jones et al., 2020; Jothimani et al., 2020). Attempts to utilize known anti-HCV treatments in COVID-19 wards had failed in a similar way (Huang et al., 2020).

Nevertheless, recent events have unequivocally shown that the coronaviruses, in general, and the SARS-CoV-2, in particular, should be regarded as yet another evolutionary driver for the fine-tuning of human interferon response to existing and emerging pathogens. Population frequencies of *IFNL4* and other interferon-encoding gene variants may reflect a sum of past exposures to the various pathogens, and the subsequent bottlenecks which may or may not be related to epidemic events.

CONCLUSION

Genetic resistance to severe viral diseases shares common molecular pathways for various viral infections. Interferon expression pathways in host cells make a crucial contribution to the proinflammatory response to infectious agent appearance. The presence of some variants in the loci of interferon gene sequences reduces the natural immunity, and stimulates a susceptibility to severe viral disease.

AUTHOR CONTRIBUTIONS

LG and AB substantially contributed to the conception of the review. LG and KP drafted the outline of the review. DN and AS contributed to the genetic and clinical aspects of the review. AB, LG, and IK contributed to the immunological aspects of the review. AB, DN, AS, and IK reviewed and edited the draft with input from all authors. DN prepared and created figures. AB supervised and managed the review planning and execution. All have read and approved the submitted version.

REFERENCES

Abbasi, A. Z., Kiyani, D. A., Hamid, S. M., Saalim, M., Fahim, A., and Jalal, N. (2021). Spiking Dependence of SARS-CoV-2 Pathogenicity on TMPRSS2. *J. Med. Virol.* 93, 4205–4218. doi:10.1002/jmv.26911

Acharya, D., Liu, G., and Gack, M. U. (2020). Dysregulation of Type I Interferon Responses in COVID-19. *Nat. Rev. Immunol.* 20 (7), 397–398. doi:10.1038/s41577-020-0346-x

Agwa, S. H. A., Kamel, M. M., Elghazaly, H., Abd Elsamee, A. M., Hafez, H., Girgis, S. A., et al. (2021). Association between Interferon-Lambda-3 Rs12979860, TLL1 Rs17047200 and DDR1 Rs4618569 Variant Polymorphisms with the Course and Outcome of SARS-CoV-2 Patients. *Genes* 12 (6), 830. doi:10.3390/genes12060830

Aka, P. V., Kuniholm, M. H., Pfeiffer, R. M., Wang, A. S., Tang, W., Chen, S., et al. (2014). Association of the IFNL4-Δg Allele with Impaired Spontaneous Clearance of Hepatitis C Virus. *J. Infect. Dis.* 209 (3), 350–354. doi:10.1093/infdis/jit433

Amanzada, A., Kopp, W., Spengler, U., Ramadori, G., and Mihm, S. (2013). Interferon-λ4 (IFNL4) Transcript Expression in Human Liver Tissue Samples. *PloS one* 8 (12), e84026. doi:10.1371/journal.pone.0084026

Andreakos, E., and Tsiodras, S. (2020). COVID -19: Lambda Interferon against Viral Load and Hyperinflammation. *EMBO Mol. Med.* 12 (6), e12465. doi:10.15252/emmm.202012465

Ank, N., West, H., and Paludan, S. R. (2006). IFN-λ: Novel Antiviral Cytokines. *J. Interferon Cytokine Res.* 26 (6), 373–379. doi:10.1089/jir.2006.26.373

Arendrup, M., Hansen, J.-E. S., Clausen, H., Nielsen, C., Mathiesen, L. R., and Nielsen, J. O. (1991). Antibody to Histo-Blood Group A Antigen Neutralizes HIV Produced by Lymphocytes from Blood Group A Donors but Not from Blood Group B or O Donors. *Aids* 5 (4), 441–444. doi:10.1097/00002030-199104000-00014

Barros, E. P., Casalino, L., Gaieb, Z., Dommer, A. C., Wang, Y., Fallon, L., et al. (2021). The Flexibility of ACE2 in the Context of SARS-CoV-2 Infection. *Biophysical J.* 120 (6), 1072–1084. doi:10.1016/j.bpj.2020.10.036

BBC (2020). *Coronavirus: Twin sisters Katy and Emma Davis die with Covid-19*. Southampton: BBC News. Available at: https://www.bbc.com/news/uk-england-hampshire-52409765.

Bibert, S., Wojtowicz, A., Taffé, P., Manuel, O., Bernasconi, E., Furrer, H., et al. (2014). The IFNL3/4 ΔG Variant Increases Susceptibility to Cytomegalovirus Retinitis Among HIV-Infected Patients. *AIDS* 28 (13), 1885–1889. doi:10.1097/qad.0000000000000379

Bogdan, C., Mattner, J., and Schleicher, U. (2004). The Role of Type I Interferons in Non-viral Infections. *Immunol. Rev.* 202, 33–48. doi:10.1111/j.0105-2896.2004.00207.x

Camargo, S. M. R., Singer, D., Makrides, V., Huggel, K., Pos, K. M., Wagner, C. A., et al. (2009). Tissue-Specific Amino Acid Transporter Partners ACE2 and Collectrin Differentially Interact with Hartnup Mutations. *Gastroenterology* 136 (3), 872–882. doi:10.1053/j.gastro.2008.10.055

Cao, Y., Li, L., Feng, Z., Wan, S., Huang, P., Sun, X., et al. (2020). Comparative Genetic Analysis of the Novel Coronavirus (2019-nCoV/SARS-CoV-2) Receptor ACE2 in Different Populations. *Cell Discov.* 6, 11. doi:10.1038/s41421-020-0147-1

Centers for Disease Control and Prevention (2020a). COVID-19 and HIV. HIV. Available at: https://www.cdc.gov/hiv/basics/covid-19.html (Accessed November 11, 2020).

Centers for Disease Control and Prevention (2020b). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep.* 69, 343–346. Available at: https://www.cdc.gov/mmwr/volumes/69/wr/mm6912e2.htm. doi:10.15585/mmwr.mm6912e2

Cheng, Y., Cheng, G., Chui, C. H., Lau, F. Y., Chan, P. K. S., Ng, M. H. L., et al. (2005). ABO Blood Group and Susceptibility to Severe Acute Respiratory Syndrome. *JAMA* 293 (12), 1447–1451. doi:10.1001/jama.293.12.1450-c

Credle, J. J., Gunn, J., Sangkhapreecha, P., Monaco, D. R., Zheng, X. A., Tsai, H. J., et al. (2021). Neutralizing IFNL3 Autoantibodies in Severe COVID-19 Identified Using Molecular Indexing of Proteins by Self-Assembly. *bioRxiv*, 432977, 2021 . [Preprint]. doi:10.1101/2021.03.02.432977

Crosse, K. M., Monson, E. A., Beard, M. R., and Helbig, K. J. (2017). Interferon-Stimulated Genes as Enhancers of Antiviral Innate Immune Signaling. *J. Innate Immun.* 10 (2), 85–93. doi:10.1159/000484258

Dalan, R., Bornstein, S. R., El-Armouche, A., Rodionov, R. N., Markov, A., Wielockx, B., et al.(2020). The ACE-2 in COVID-19: Foe or Friend?. *Horm. Metab. Res.* 52 (5), 257–263. doi:10.1055/a-1155-0501

Donnelly, R. P., and Kotenko, S. V. (2010). Interferon-lambda: a New Addition to an Old Family. *J. Interferon Cytokine Res.* 30 (8), 555–564. doi:10.1089/jir.2010.0078

Du, R.-H., Liang, L.-R., Yang, C.-Q., Wang, W., Cao, T.-Z., Li, M., et al. (2020). Predictors of Mortality for Patients with COVID-19 Pneumonia Caused by SARS-CoV-2: a Prospective Cohort Study. *Eur. Respir. J.* 55 (5), 2000524. doi:10.1183/13993003.00524-2020

Duggal, P., Thio, C. L., Wojcik, G. L., Goedert, J. J., Mangia, A., Latanich, R., et al. (2013). Genome-Wide Association Study of Spontaneous Resolution of Hepatitis C Virus Infection: Data from Multiple Cohorts. *Ann. Intern. Med.* 158 (4), 235. doi:10.7326/0003-4819-158-4-201302190-00003

Egli, A., Levin, A., Santer, D. M., Joyce, M., O'Shea, D., Thomas, B. S., et al. (2014). Immunomodulatory Function of Interleukin 28B during Primary Infection with Cytomegalovirus. *J. Infect. Dis.* 210 (5), 717–727. doi:10.1093/infdis/jiu144

Estep, M., Perry, K., Tavakolian, K., Younoszai, Z., Stepanova, M., Noorzad, A., and Younossi, Z. (2014). Interferon Lambda-4 (IFNL4) TT Allele Is Associated with Lower Expression of Genes Associated with Early Inflammation after Initiation of Treatment: 1835. *65th Annual Meeting of the American Association for the Study of Liver Diseases: The Liver Meeting 2014*, 182A. doi:10.1002/hep.27415

Fan, J. H., Hou, S. H., Qing-Ling, L., Hu, J., Peng, H., and Guo, J. J. (2016). Association of HLA-DQ and IFNL4 Polymorphisms with Susceptibility to Hepatitis B Virus Infection and Clearance. *Ann. Hepatol.* 15 (4), 532–539. doi:10.5604/16652681.1202946

Feld, J. J., Kandel, C., Biondi, M. J., Kozak, R. A., Zahoor, M. A., Lemieux, C., et al. (2021). Peginterferon Lambda for the Treatment of Outpatients with COVID-19: a Phase 2, Placebo-Controlled Randomised Trial. *Lancet Respir. Med.* 9 (5), 498–510. doi:10.1016/S2213-2600(20)30566-X

Focosi, D. (2020). Anti-A Isohaemagglutinin Titres and SARS-CoV-2 Neutralization: Implications for Children and Convalescent Plasma Selection. *Br. J. Haematol.* 190 (3), 148–150. doi:10.1111/bjh.16932

Franco, S., Aparicio, E., Parera, M., Clotet, B., Tural, C., and Martinez, M. A. (2014). IFNL4 Ss469415590 Variant Is a Better Predictor Than Rs12979860 of Pegylated Interferon-Alpha/ribavirin Therapy Failure in Hepatitis C virus/HIV-1 Coinfected Patients. *AIDS* 28 (1), 133–136. doi:10.1097/QAD.0000000000000052

Gad, H. H., Dellgren, C., Hamming, O. J., Vends, S., Paludan, S. R., and Hartmann, R. (2009). Interferon-λ Is Functionally an Interferon but Structurally Related to the Interleukin-10 Family. *J. Biol. Chem.* 284 (31), 20869–20875. doi:10.1074/jbc.m109.002923

Garg, S., Kim, L., Whitaker, M., O'Halloran, A., Cummings, C., Holstein, R., et al. (2020). Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States, March 1–30, 2020. *MMWR Morb Mortal Wkly Rep.* 69, 458–464. doi:10.15585/mmwr.mm6915e3

Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., et al. (2009). Genetic Variation in *IL28B* Predicts Hepatitis C Treatment-Induced Viral Clearance. *Nature* 461, 399–401. doi:10.1038/nature08309

Gong, J., Ou, J., Qiu, X., Jie, Y., Chen, Y., Yuan, L., et al. (2020). A Tool to Early Predict Severe Corona Virus Disease 2019 (COVID-19) : A Multicenter Study Using the Risk Nomogram in Wuhan and Guangdong, China. *Clin. Infect. Dis* 71, 833–840. Cold Spring Harbor Laboratory. doi:10.1093/cid/ciaa443

Gough, D. J., Levy, D. E., Johnstone, R. W., and Clarke, C. J. (2008). IFNgamma Signaling-Does it Mean JAK-STAT?. *Cytokine Growth Factor. Rev.* 19 (5-6), 383–394. doi:10.1016/j.cytogfr.2008.08.004

Guan, J., Miah, S. M. S., Wilson, Z. S., Erick, T. K., Banh, C., and Brossay, L. (2014). Role of Type I Interferon Receptor Signaling on NK Cell Development and Functions. *PLoS ONE* 9 (10), e111302. doi:10.1371/journal.pone.0111302

Guillon, P., Clément, M., Sébille, V., Rivain, J.-G., Chou, C.-F., Ruvoën-Clouet, N., et al. (2008). Inhibition of the Interaction between the SARS-CoV Spike Protein and its Cellular Receptor by Anti-histo-blood Group Antibodies. *Glycobiology* 18 (12), 1085–1093. doi:10.1093/glycob/cwn093

Hacibekiroğlu, T., Kalpakçı, Y., Genç, A. C., Hacibekiroğlu, İ., Sunu, C., Saricaoğlu, A., et al. (2021). Efficacy of Convalescent Plasma According to Blood Groups in COVID-19 Patients. *Turk J. Med. Sci.* 51 (1), 45–48. doi:10.3906/sag-2007-59

Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., et al. (2020). Impaired Type I Interferon Activity and Inflammatory Responses in Severe COVID-19 Patients. *Science* 369 (6504), 718–724. doi:10.1126/science.abc6027

Hamming, O. J., Terczyńska-Dyla, E., Vieyres, G., Dijkman, R., Jørgensen, S. E., Akhtar, H., et al. (2013). Interferon Lambda 4 Signals via the IFNλ Receptor to Regulate Antiviral Activity against HCV and Coronaviruses. *EMBO J.* 32 (23), 3055–3065. doi:10.1038/emboj.2013.232

Hervas-Stubbs, S., Perez-Gracia, J. L., Rouzaut, A., Sanmamed, M. F., Le Bon, A., and Melero, I. (2011). Direct Effects of Type I Interferons on Cells of the Immune System. *Clin. Cancer Res.* 17 (9), 2619–2627. doi:10.1158/1078-0432.CCR-10-1114

Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* 181 (2), 271–280. doi:10.1016/j.cell.2020.02.052

Hoiland, R. L., Fergusson, N. A., Mitra, A. R., Griesdale, D. E. G., Devine, D. V., Stukas, S., et al. (2020). The Association of ABO Blood Group with Indices of Disease Severity and Multiorgan Dysfunction in COVID-19. *Blood Adv.* 4 (20), 4981–4989. doi:10.1182/bloodadvances.2020002623

Huang, Y. Q., Tang, S. Q., Xu, X. L., Zeng, Y. M., He, X. Q., Li, Y., et al. (2020). No Statistically Apparent Difference in Antiviral Effectiveness Observed Among Ribavirin Plus Interferon-Alpha, Lopinavir/Ritonavir Plus Interferon-Alpha, and Ribavirin Plus Lopinavir/Ritonavir Plus Interferon-Alpha in Patients with Mild to Moderate Coronavirus Disease 2019: Results of a Randomized, Open-Labeled Prospective Study. *Front. Pharmacol.* 11, 1071. doi:10.3389/fphar.2020.01071

Jagannathan, P., Andrews, J. R., Bonilla, H., Hedlin, H., Jacobson, K. B., Balasubramanian, V., et al. (2021). Peginterferon Lambda-1a for Treatment of Outpatients with Uncomplicated COVID-19: a Randomized Placebo-Controlled Trial. *Nat. Commun.* 12 (1), 1967. doi:10.1038/s41467-021-22177-1

Jones, R., Nelson, M., Bracchi, M., Asboe, D., and Boffito, M. (2020). COVID-19 in Patients with HIV. *The Lancet HIV.* 7 (6), e383. doi:10.1016/s2352-3018(20)30139-9

Jothimani, D., Venugopal, R., Abedin, M. F., Kaliamoorthy, I., and Rela, M. (2020). COVID-19 and the Liver. *J. Hepatol.* 73 (5), 1231–1240. doi:10.1016/j.jhep.2020.06.006

Jouvin-Marche, E., Macek Jílková, Z., Thelu, M.-A., Marche, H., Fugier, E., Van Campenhout, N., et al. (2014). Lymphocytes Degranulation in Liver in Hepatitis C Virus Carriers Is Associated with IFNL4 Polymorphisms and ALT Levels. *J. Infect. Dis.* 209 (12), 1907–1915. doi:10.1093/infdis/jiu016

Key, F. M., Peter, B., Dennis, M. Y., Huerta-Sánchez, E., Tang, W., Prokunina-Olsson, L., et al. (2014). Selection on a Variant Associated with Improved Viral Clearance Drives Local, Adaptive Pseudogenization of Interferon Lambda 4 (IFNL4). *Plos Genet.* 10 (10), e1004681. doi:10.1371/journal.pgen.1004681

Klinkhammer, J., Schnepf, D., Ye, L., Schwaderlapp, M., Gad, H. H., Hartmann, R., et al. (2018). IFN-λ Prevents Influenza Virus Spread from the Upper Airways to the Lungs and Limits Virus Transmission. *eLife* 7, e33354. doi:10.7554/eLife.33354

Kole, A., He, J., Rivollier, A., Silveira, D. D., Kitamura, K., Maloy, K. J., et al. (2013). Type I IFNs Regulate Effector and Regulatory T Cell Accumulation and Anti-inflammatory Cytokine Production during T Cell-Mediated Colitis. *J. Immunol.* 191 (5), 2771–2779. doi:10.4049/jimmunol.1301093

Kosmicki, J. A., Horowitz, J. E., Banerjee, N., Lanche, R., Marcketta, A., Maxwell, E., et al. (2021). Pan-ancestry Exome-wide Association Analyses of COVID-19 Outcomes in 586,157 Individuals. *Am. J. Hum. Genet.* 108 (7), 1350–1355. doi:10.1016/j.ajhg.2021.05.017

Kotenko, S. V., Gallagher, G., Baurin, V. V., Lewis-Antes, A., Shen, M., Shah, N. K., et al. (2003). IFN-lambdas Mediate Antiviral protection through a Distinct Class II Cytokine Receptor Complex. *Nat. Immunol.* 4 (1), 69–77. doi:10.1038/ni875

Latini, A., Agolini, E., Novelli, A., Borgiani, P., Giannini, R., Gravina, P., et al. (2020). COVID-19 and Genetic Variants of Protein Involved in the SARS-CoV-2 Entry into the Host Cells. *Genes (Basel)* 11 (9), 1010. doi:10.3390/genes11091010

Lauber, C., Vieyres, G., Terczyńska-Dyla, E., Dijkman, R., Gad, H. H., Akhtar, H., et al. (2015). Transcriptome Analysis Reveals a Classical Interferon Signature Induced by IFNλ4 in Human Primary Cells. *Genes Immun.* 16, 414–421. doi:10.1038/gene.2015.23

Lazear, H. M., Nice, T. J., and Diamond, M. S. (2015). Interferon-λ: Immune Functions at Barrier Surfaces and beyond. *Immunity* 43 (1), 15–28. doi:10.1016/j.immuni.2015.07.001

Li, J., Wang, X., Chen, J., Cai, Y., Deng, A., and Yang, M. (2020). Association between ABO Blood Groups and Risk of SARS-CoV-2 Pneumonia. *Br. J. Haematol.* 190 (1), 24–27. doi:10.1111/bjh.16797

Lin, F. C., and Young, H. A. (2014). Interferons: Success in Anti-viral Immunotherapy. *Cytokine Growth Factor. Rev.* 25 (4), 369–376. doi:10.1016/j.cytogfr.2014.07.015

Lokugamage, K. G., Hage, A., de Vries, M., Valero-Jimenez, A. M., Schindewolf, C., Dittmann, M., et al. (2020). Type I Interferon Susceptibility Distinguishes

SARS-CoV-2 from SARS-CoV. *J. Virol.* 94, e01410, 2020 . Cold Spring Harbor Laboratory. doi:10.1128/JVI.01410-20

Lopera Maya, E. A., van der Graaf, A., Lanting, P., van der Geest, M., Fu, J., Swertz, M., et al. (2020). Lack of Association between Genetic Variants at ACE2 and TMPRSS2 Genes Involved in SARS-CoV-2 Infection and Human Quantitative Phenotypes. *Front. Genet.* 11, 613. doi:10.3389/fgene.2020.00613

Lopez, L., Sang, P. C., Tian, Y., and Sang, Y. (2020). Dysregulated Interferon Response Underlying Severe COVID-19. *Viruses* 12 (12), 1433. doi:10.3390/v12121433

Lopušná, K., Režuchová, I., Betáková, T., Skovranová, L., Tomašková, J., Lukáčiková, L., et al. (2013). Interferons Lambda, New Cytokines with Antiviral Activity. *Acta Virol.* 57 (2), 171–179. doi:10.4149/av_2013_02_171

Machiela, M. J., and Chanock, S. J. (2015). LDlink: a Web-Based Application for Exploring Population-specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants: Fig. 1. *Bioinformatics* 31 (21), 3555–3557. doi:10.1093/bioinformatics/btv402

Machmach, K., Abad-Molina, C., Romero-Sánchez, M. C., Abad, M. A., Ferrando-Martínez, S., Genebat, M., et al. (2013). IL28B Single-Nucleotide Polymorphism Rs12979860 Is Associated with Spontaneous HIV Control in White Subjects. *J. Infect. Dis.* 207 (4), 651–655. doi:10.1093/infdis/jis717

Machmach, K., Abad-Molina, C., Romero-Sánchez, M. C., Dominguez-Molina, B., Moyano, M., Rodriguez, M. M., et al. (2015). IFNL4 Ss469415590 Polymorphism Is Associated with Unfavourable Clinical and Immunological Status in HIV-Infected Individuals. *Clin. Microbiol. Infect.* 21 (3), 289. doi:10.1016/j.cmi.2014.10.012

Majdic, G. (2020). Could Sex/Gender Differences in ACE2 Expression in the Lungs Contribute to the Large Gender Disparity in the Morbidity and Mortality of Patients Infected with the SARS-CoV-2 Virus?. *Front. Cell Infect. Microbiol.* 10, 327. doi:10.3389/fcimb.2020.00327

Malmgaard, L. (2004). Induction and Regulation of IFNs during Viral Infections. *J. Interferon Cytokine Res.* 24 (8), 439–454. doi:10.1089/1079990041689665

Manuel, O., Wójtowicz, A., Bibert, S., Mueller, N. J., van Delden, C., Hirsch, H. H., et al. (2015). Influence of IFNL3/4 Polymorphisms on the Incidence of Cytomegalovirus Infection after Solid-Organ Transplantation. *J. Infect. Dis.* 211 (6), 906–914. doi:10.1093/infdis/jiu557

Meffre, E., and Iwasaki, A. (2020). Interferon Deficiency Can lead to Severe COVID. *Nature* 587 (7834), 374–376. doi:10.1038/d41586-020-03070-1

Meissner, E. G., Bon, D., Prokunina-Olsson, L., Tang, W., Masur, H., O'Brien, T. R., et al. (2014). IFNL4-ΔG Genotype Is Associated with Slower Viral Clearance in Hepatitis C, Genotype-1 Patients Treated with Sofosbuvir and Ribavirin. *J. Infect. Dis.* 209 (11), 1700–1704. doi:10.1093/infdis/jit827

Messina, N. L., Clarke, C. J., and Johnstone, R. W. (2016). Constitutive IFNα/β Signaling Maintains Expression of Signaling Intermediaries for Efficient Cytokine Responses. *JAK-STAT* 5 (1), e1173804. doi:10.1080/21623996.2016.1173804

Muir, A. J., Arora, S., Everson, G., Flisiak, R., George, J., Ghalib, R., et al. (2014). A Randomized Phase 2b Study of Peginterferon Lambda-1a for the Treatment of Chronic HCV Infection. *J. Hepatol.* 61 (6), 1238–1246. doi:10.1016/j.jhep.2014.07.022

Nagy, B., Jr, Fejes, Z., Szentkereszty, Z., Sütő, R., Várkonyi, I., Ajzner, É., et al. (2021). A Dramatic Rise in Serum ACE2 Activity in a Critically Ill COVID-19 Patient. *Int. J. Infect. Dis.* 103, 412–414. doi:10.1016/j.ijid.2020.11.184

Ni, W., Yang, X., Yang, D., Bao, J., Li, R., Xiao, Y., et al. (2020). Role of Angiotensin-Converting Enzyme 2 (ACE2) in COVID-19. *Crit. Care.* 24, 422. doi:10.1186/s13054-020-03120-0

Nozawa, Y., Umemura, T., Katsuyama, Y., Shibata, S., Kimura, T., Morita, S., et al. (2014). Genetic Polymorphism in IFNL4 and Response to Pegylated Interferon-α and Ribavirin in Japanese Chronic Hepatitis C Patients. *Tissue Antigens* 83 (1), 45–48. doi:10.1111/tan.12264

O'Brien, T. R., Prokunina-Olsson, L., and Donnelly, R. P. (2014). IFN-λ4: the Paradoxical New Member of the Interferon Lambda Family. *J.Iinterferon Cytokine Res.: official J. Int. Soc. Interferon Cytokine Res.* 34 (11), 829–838. doi:10.1089/jir.2013.0136

Olex, A. L., Turkett, W. H., Brzoza-Lewis, K. L., Fetrow, J. S., and Hiltbold, E. M. (2016). Impact of the Type I Interferon Receptor on the Global Gene Expression Program during the Course of Dendritic Cell Maturation Induced by Polyinosinic Polycytidylic Acid. *J. interferon Cytokine Res. : official J. Int. Soc. Interferon Cytokine Res.* 36 (6), 382–400. doi:10.1089/jir.2014.0150

Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., et al. (2021). Genetic Mechanisms of Critical Illness in COVID-19. *Nature* 591, 92–98. doi:10.1038/s41586-020-03065-y

Preece, A. F., Strahan, K. M., Devitt, J., Yamamoto, F., and Gustafsson, K. (2002). Expression of ABO or Related Antigenic Carbohydrates on Viral Envelopes Leads to Neutralization in the Presence of Serum Containing Specific Natural Antibodies and Complement. *Blood* 99 (7), 2477–2482. doi:10.1182/blood.v99.7.2477

Prokunina-Olsson, L., Muchmore, B., Tang, W., Pfeiffer, R. M., Park, H., Dickensheets, H., et al. (2013). A Variant Upstream of IFNL3 (IL28B) Creating a New Interferon Gene IFNL4 Is Associated with Impaired Clearance of Hepatitis C Virus. *Nat. Genet.* 45 (2), 164–171. doi:10.1038/ng.2521

Prokunina-Olsson, L., Alphonse, N., Dickenson, R. E., Durbin, J. E., Glenn, J. S., Hartmann, R., et al. (2020). COVID-19 and Emerging Viral Infections: The Case for Interferon Lambda. *J. Exp. Med.* 217 (5), e20200653. doi:10.1084/jem.20200653

Randall, R. E., and Goodbourn, S. (2008). Interferons and Viruses: an Interplay between Induction, Signalling, Antiviral Responses and Virus Countermeasures. *J. Gen. Virol.* 89 (Pt 1), 1–47. doi:10.1099/vir.0.83391-0

Real, L. M., Herrero, R., Rivero-Juárez, A., Camacho, Á., Macías, J., Vic, S., et al. (2015). IFNL4 Rs368234815 Polymorphism Is Associated with Innate Resistance to HIV-1 Infection. *AIDS* 29 (14), 1895–1897. doi:10.1097/qad.0000000000000773

Ruan, S. (2020). Likelihood of Survival of Coronavirus Disease 2019. *Lancet Infect. Dis.* 20 (6), 630–631. doi:10.1016/s1473-3099(20)30257-7

Sajadi, M. M., Shakeri, N., Talwani, R., Howell, C. D., Pakyz, R., Redfield, R. R., et al. (2011). IL28B Genotype Does Not Correlate with HIV Control in African Americans. *Clin. Translational Sci.* 4 (4), 282–284. doi:10.1111/j.1752-8062.2011.00307.x

Salgado, M., Kirk, G. D., Cox, A., Rutebemberwa, A., Higgins, Y., Astemborski, J., et al. (2011). Protective interleukin-28B Genotype Affects Hepatitis C Virus Clearance, but Does Not Contribute to HIV-1 Control in a Cohort of African-American Elite Controllers/suppressors. *AIDS (London, England)* 25 (3), 385–387. doi:10.1097/QAD.0b013e328341b86a

Sallard, E., Lescure, F.-X., Yazdanpanah, Y., Mentre, F., and Peiffer-Smadja, N. (2020). Type 1 Interferons as a Potential Treatment against COVID-19. *Antiviral Res.* 178, 104791. doi:10.1016/j.antiviral.2020.104791

Saponi-Cortes, J. M. R., Rivas, M. D., Calle, F., Sanchez Muñoz-Torrero, J. F., Costo, A., Martin, C., et al. (2021). IFNL4 Genetic Variant Can Predispose to COVID-19. *medRxiv* Cold Spring Harbor Laboratory. doi:10.1101/2021.03.01.21252696

Schneider, W. M., Chevillotte, M. D., and Rice, C. M. (2014). Interferon-stimulated Genes: a Complex Web of Host Defenses. *Annu. Rev. Immunol.* 32, 513–545. doi:10.1146/annurev-immunol-032713-120231

Seliger, B., Ruiz-Cabello, F., and Garrido, F. (2008). IFN Inducibility of Major Histocompatibility Antigens in Tumors. *Adv. Cancer Res.* 101, 249–276. doi:10.1016/S0065-230X(08)00407-7

Sheppard, P., Kindsvogel, W., Xu, W., Henderson, K., Schlutsmeyer, S., Whitmore, T. E., et al. (2003). IL-28, IL-29 and Their Class II Cytokine Receptor IL-28R. *Nat. Immunol.* 4 (1), 63–68. doi:10.1038/ni873

Sokol, C. L., and Luster, A. D. (2015). The Chemokine System in Innate Immunity. *Cold Spring Harbor Perspect. Biol.* 7 (5), a016303. doi:10.1101/cshperspect.a016303

Stanifer, M. L., Kee, C., Cortese, M., Zumaran, C. M., Triana, S., Mukenhirn, M., et al. (2020). Critical Role of Type III Interferon in Controlling SARS-CoV-2 Infection in Human Intestinal Epithelial Cells. *Cell Rep.* 32 (1), 107863. doi:10.1016/j.celrep.2020.107863

Stättermayer, A. F., Strassl, R., Maieron, A., Rutter, K., Stauber, R., Strasser, M., et al. (2014). Polymorphisms of Interferon-Λ4 and IL28B - Effects on Treatment Response to Interferon/ribavirin in Patients with Chronic Hepatitis C. *Aliment. Pharmacol. Ther.* 39 (1), 104–111. doi:10.1111/apt.12547

Sung, P. S., Hong, S. H., Chung, J. H., Kim, S., Park, S. H., Kim, H. M., et al. (2017). IFN-λ4 Potently Blocks IFN-α Signalling by ISG15 and USP18 in Hepatitis C Virus Infection. *Sci. Rep.* 7, 3821. doi:10.1038/s41598-017-04186-7

The 1000 Genomes Project Consortium (2015). Corresponding authors., Auton, A. *et al.* A global reference for human genetic variation. *Nature* 526, 68–74. doi:10.1038/nature15393

The COVID-19 Host Genetics Initiative (2020). The COVID-19 Host Genetics Initiative, a Global Initiative to Elucidate the Role of Host Genetic Factors in Susceptibility and Severity of the SARS-CoV-2 Virus Pandemic. *Eur. J. Hum. Genet.* 28, 715–718. doi:10.1038/s41431-020-0636-6

The Severe Covid-19 GWAS Group (2020). Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *New Engl. J. Med.* 383 (16), 1522–1534. doi:10.1056/nejmoa2020283

Thomas, D. L., Thio, C. L., Martin, M. P., Qi, Y., Ge, D., O'Huigin, C., et al. (2009). Genetic Variation in IL28B and Spontaneous Clearance of Hepatitis C Virus. *Nature* 461 (7265), 798–801. doi:10.1038/nature08463

Vardavas, C. I., and Nikitara, K. (2020). COVID-19 and Smoking: A Systematic Review of the Evidence. *Tob. Induc Dis.* 18, 20. doi:10.18332/tid/119324

Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., et al. (2020). Estimates of the Severity of Coronavirus Disease 2019: a Model-Based Analysis. *Lancet Infect. Dis.* 20 (6), 669–677. doi:10.1016/s1473-3099(20) 30243-7

Vuille-dit-Bille, R. N., Camargo, S. M., Emmenegger, L., Sasse, T., Kummer, E., Jando, J., et al. (2015). Human Intestine Luminal ACE2 and Amino Acid Transporter Expression Increased by ACE-Inhibitors. *Amino Acids.* 47, 693–705. doi:10.1007/s00726-014-1889-6

Wang, A., Chiou, J., Poirion, O. B., Buchanan, J., Valdez, M. J., Verheyden, J. M., et al. (2020). Single-cell Multiomic Profiling of Human Lungs Reveals Cell-type-specific and Age-Dynamic Control of SARS-CoV2 Host Genes. *ELife* 9. doi:10.7554/elife.62522

Weiss, P., and Murdoch, D. R. (2020). Clinical Course and Mortality Risk of Severe COVID-19. *The Lancet* 395 (10229), 1014–1015. doi:10.1016/s0140-6736(20) 30633-4

Williams, F. M., Freidin, M. B., Mangino, M., Couvreur, S., Visconti, A., Bowyer, R. C., et al. (2020). Self-reported Symptoms of Covid-19 Including Symptoms Most Predictive of SARS-CoV-2 Infection, Are Heritable. *Twin. Res. Hum. Genet.* 23, 316–321. Cold Spring Harbor Laboratory. doi:10.1017/thg.2020.85

World Health Organization (2021). Coronavirus. Available at: https://www.who.int/health-topics/coronavirus#tab=tab_1 (Accessed May 11, 2021).

World Health Organization (2020c). Coronavirus Disease (COVID-19). Available at: https://www.who.int/news-room/q-a-detail/q-a-coronaviruses (Accessed November 11, 2020).

World Health Organization (2020a). Novel Coronavirus - China. Available at: https://www.who.int/csr/don/12-january-2020-novel-coronavirus-china/en/ (Accessed November 11, 2020).

World Health Organization (2020b). Weekly Epidemiological and Operational Updates December 2020. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (Accessed November 11, 2020).

Wruck, W., and Adjaye, J. (2020). SARS-CoV-2 Receptor ACE2 Is Co-expressed with Genes Related to Transmembrane Serine Proteases, Viral Entry, Immunity and Cellular Stress. *Sci. Rep.* 10 (1), 21415. doi:10.1038/s41598-020-78402-2

Wu, Z., and McGoogan, J. M. (2020). Characteristics of and Important Lessons from the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases from the Chinese Center for Disease Control and Prevention. *JAMA* 323 (13), 1239–1242. doi:10.1001/jama.2020.2648

Younossi, Z. M., Birerdinc, A., Estep, M., Stepanova, M., Afendy, A., Baranova, A., et al. (2012). The Impact of IL28B Genotype on the Gene Expression Profile of Patients with Chronic Hepatitis C Treated with Pegylated Interferon Alpha and Ribavirin. *J. Transl Med.* 10, 25. doi:10.1186/1479-5876-10-25

Zeberg, H., and Pääbo, S. (2020). The Major Genetic Risk Factor for Severe COVID-19 Is Inherited from Neandertals. *Nature* 587, 610–612. Cold Spring Harbor Lab.. doi:10.1038/s41586-020-2818-3

Zhang, Q., Bastard, P., Liu, Z., Le Pen, J., Moncada-Velez, M., Chen, J., et al. (2020). Inborn Errors of Type I IFN Immunity in Patients with Life-Threatening COVID-19. *Science* 370 (6515), eabd4570. doi:10.1126/science.abd4570

Zhao, J., Yang, Y., Huang, H., Li, D., Gu, D., Lu, X., et al. (2020). Relationship between the ABO Blood Group and the COVID-19 Susceptibility. Cold Spring Harbor Lab.. doi:10.1101/2020.03.11.20031096

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020). A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature* 579, 270–273. doi:10.1038/s41586-020-2012-7

Ziegler, C. G. K., Allon, S. J., Nyquist, S. K., Mbano, I. M., Miao, V. N., Tzouanas, C. N., et al. (2020). SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. *Cell* 181 (5), 1016–1035. doi:10.1016/j.cell.2020.04.035

Conflict of Interest: Authors DN, AS, and AB were employed by the company Atlas Biomed Group Limited.

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.

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: 07 October 2021
doi: 10.3389/fgene.2021.709419

Targeted Sequencing of 242 Clinically Important Genes in the Russian Population From the Ivanovo Region

Vasily E. Ramensky[1,2]*, Alexandra I. Ershova[1], Marija Zaicenoka[3], Anna V. Kiseleva[1], Anastasia A. Zharikova[1,2], Yuri V. Vyatkin[1,4], Evgeniia A. Sotnikova[1], Irina A. Efimova[1], Mikhail G. Divashuk[1,5], Olga V. Kurilova[1], Olga P. Skirko[1], Galina A. Muromtseva[1], Olga A. Belova[6], Svetlana A. Rachkova[6], Maria S. Pokrovskaya[1], Svetlana A. Shalnova[1], Alexey N. Meshkov[1]† and Oxana M. Drapkina[1]†

[1] National Medical Research Center for Therapy and Preventive Medicine, Moscow, Russia, [2] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [3] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow, Russia, [4] Novosibirsk State University, Novosibirsk, Russia, [5] All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia, [6] Cardiology Dispensary, Ivanovo, Russia

We performed a targeted sequencing of 242 clinically important genes mostly associated with cardiovascular diseases in a representative population sample of 1,658 individuals from the Ivanovo region northeast of Moscow. Approximately 11% of 11,876 detected variants were not found in the Single Nucleotide Polymorphism Database (dbSNP) or reported earlier in the Russian population. Most novel variants were singletons and doubletons in our sample, and virtually no novel alleles presumably specific for the Russian population were able to reach the frequencies above 0.1–0.2%. The overwhelming majority (99.3%) of variants detected in this study in three or more copies were shared with other populations. We found two dominant and seven recessive known pathogenic variants with allele frequencies significantly increased compared to those in the gnomAD non-Finnish Europeans. Of the 242 targeted genes, 28 were in the list of 59 genes for which the American College of Medical Genetics and Genomics (ACMG) recommended the reporting of incidental findings. Based on the number of variants detected in the sequenced subset of ACMG59 genes, we approximated the prevalence of known pathogenic and novel or rare protein-truncating variants in the complete set of ACMG59 genes in the Ivanovo population at 1.4 and 2.8%, respectively. We analyzed the available clinical data and observed the incomplete penetrance of known pathogenic variants in the 28 ACMG59 genes: only 1 individual out of 12 with such variants had the phenotype most likely related to the variant. When known pathogenic and novel or rare protein-truncating variants were considered together, the overall rate of confirmed phenotypes was about 19%, with maximum in the subset of novel protein-truncating variants. We report three novel protein truncating variants in *APOB* and one in *MYH7* observed in individuals with hypobetalipoproteinemia and hypertrophic cardiomyopathy, respectively. Our results provide a valuable reference for the clinical interpretation of gene sequencing in Russian and other populations.

Keywords: genetic testing, rare variants, secondary findings, pathogenic variants, penetrance

OPEN ACCESS

Edited by:
Tatiana V. Tatarinova,
University of La Verne, United States

Reviewed by:
Meng Wu,
Editas Medicine, Inc., United States
Evgenii Chekalin,
Michigan State University,
United States

****Correspondence:***
Vasily E. Ramensky
ramensky@gmail.com

†These authors have contributed equally to this work and share last authorship

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 May 2021*
Accepted: *31 August 2021*
Published: *07 October 2021*

Citation:
Ramensky VE, Ershova AI, Zaicenoka M, Kiseleva AV, Zharikova AA, Vyatkin YV, Sotnikova EA, Efimova IA, Divashuk MG, Kurilova OV, Skirko OP, Muromtseva GA, Belova OA, Rachkova SA, Pokrovskaya MS, Shalnova SA, Meshkov AN and Drapkina OM (2021) Targeted Sequencing of 242 Clinically Important Genes in the Russian Population From the Ivanovo Region. Front. Genet. 12:709419. doi: 10.3389/fgene.2021.709419

INTRODUCTION

The next-generation sequencing projects of the last decade revealed the complicated spectrum of rare and ultra-rare allelic variation in most human genes (Tennessen et al., 2012; Lek et al., 2016; Van Hout et al., 2020). The sequencing of large population cohorts greatly enriched our understanding of the prevalence of presumably pathogenic variants in reportedly healthy people (Dorschner et al., 2013; Amendola et al., 2015; Dewey et al., 2016) and initiated studies aimed at reviewing the implied pathogenicity of genetic variants (Dewey et al., 2014; Shah et al., 2018) and penetrance of variants classified as pathogenic (Cooper et al., 2013; Chen et al., 2016; Wright et al., 2019; Van Rooij et al., 2020). This activity was shaped by the standards and guidelines for the classification of sequence variants using criteria informed by expert opinion developed by The American College of Medical Genetics and Genomics (ACMG) (Richards et al., 2015; Nykamp et al., 2017).

The ACMG also suggested the list of genes for which the reporting of known and expected pathogenic variants, also termed incidental findings, may be recommended (Green et al., 2013). The updated guidelines presented ACMG59, the list of the mostly dominant 59 genes associated with medically actionable disorders (Kalia et al., 2017). However, limited information on the penetrance of many variants even in this relatively small set of genes makes interpretation challenging. This uncertainty is gradually decreasing as more studies evaluate the carrier status of pathogenic gene variants in general populations (Amendola et al., 2016; Wright et al., 2019; Van Rooij et al., 2020). Numerous papers reported the prevalence of known and expected pathogenic variants in various populations, for example, Dutch (Haer-Wigman et al., 2019; Van Rooij et al., 2020), Qatari (Jain et al., 2018), Korean (Kwak et al., 2017), Australian (Lacaze et al., 2020), British (Van Hout et al., 2020), and Taiwanese (Kuo et al., 2020). The present study investigates the incidental findings of pathogenic variants in 28 genes from the ACMG59 list in the Russian population from the Ivanovo region.

Russia is one of the most ethnically diverse countries in the world; however, it is underrepresented in the large human genome sequencing projects (Lek et al., 2016; Van Hout et al., 2020). Earlier genome-wide studies of the Russian population were mostly based on the limited number of samples and focused on the genetic history and admixture patterns (Mallick et al., 2016; Wong et al., 2017). Recently, the whole-genome variation was analyzed for the 264 healthy adult participants of the Genome Russia Project (Zhernakova et al., 2020) demonstrating adaptive, clinical, and functional consequences. Approximately 3–4% of the called variants were classified as novel when compared to the dbSNP database (Sherry et al., 2001); in many cases, allele frequencies demonstrated divergence from the neighboring populations. Barbitoff et al. (2019) reported variant frequencies in a larger subset of 694 exomes from the northwest Russia. The results indicated that 9.3% of discovered variants were not present in dbSNP. Moreover, this whole-exome study demonstrated the overrepresentation of several disease-causing variants for Mendelian disorders, such as phenylketonuria (*PAH*, rs5030858), Wilson’s disease (*ATP7B*, rs76151636), factor VII deficiency (*F7*, rs36209567), and the kyphoscoliosis type of Ehlers–Danlos syndrome (*FKBP14*, rs542489955). For the Russian population, however, pathogenic variant frequencies were reported mostly for relatively small cohorts including patients and their families and targeted at specific genes and disorders, for example, familial hypercholesterolemia (Meshkov et al., 2021; Miroshnikova et al., 2021); cystic fibrosis, phenylketonuria, alpha-1 antitrypsin deficiency, and sensorineural hearing loss (Kiseleva et al., 2020; Petrova et al., 2020); cardiomyopathy (Marakhonov et al., 2019; Zaklyazminskaya et al., 2019; Kulikova et al., 2021; Shestak et al., 2021); and breast and ovarian cancer (Brovkina et al., 2018; Solodskikh et al., 2019).

The aim of this work was to take a step forward in the study of the genetic makeup of the Russian population, with the emphasis on rare variation, in particular, known and expected pathogenic variants in a subset of clinically important genes. We developed the gene panel which included 242 protein-coding genes associated with cardiovascular diseases and high risk of early or sudden cardiac death and used the population samples collected in the course of the ESSE-RF cardiovascular epidemiology project (Boitsov et al., 2013). Below, we report the results of the targeted sequencing of 1,685 unrelated participants from Ivanovo, one of the ESSE-RF regions. The availability of clinical data made it possible to evaluate the carriers of known and expected novel pathogenic variants for certain phenotypes.

MATERIALS AND METHODS

Selection of Participants and Clinical Data

The individuals for our study were selected from the study “Epidemiology of Cardiovascular Diseases and Risk Factors in Regions of the Russian Federation (ESSE-RF).” The ESSE-RF is a multicenter population-based study, conducted from 2012 to 2014, covering 13 regions of Russia, differing in climatic, geographic, economic, and demographic characteristics (Boitsov et al., 2013). About 1,600–1,900 people, aged 25–64, were randomly selected from every region, including Ivanovo, a region approximately 300 km northeast of Moscow with predominant Russian population. Please see more details regarding sample selection in the **Supplementary Methods**.

The sequenced set contained 1,685 individuals: 1,056 females with median age 52 at the moment of enrollment and 629 males with median age 44, respectively. We used PLINK v1.90 (Chang et al., 2015) to calculate identity by state (IBS) values and estimate the identity by descent (IBD) proportion (PI_HAT) for all possible pairs of individuals. To ensure that our dataset does not include closely related participants, we removed all pairs with

Abbreviation: AC, allele count; ACMG59, genes recommended by the ACMG for return of incidental findings; AF, allele frequency; CAD, coronary artery disease; CI, confidence interval; dbSNP, the Single Nucleotide Polymorphism Database; ECG, electrocardiogram; IBD, identity by descent; IBS, identity by state; KP, known pathogenic; LDL-C, low-density lipoprotein cholesterol; NFE, Non-Finnish European population; NWR, northwest Russia; PTV, protein-truncating variant; VEP, ENSEMBL Variant Effect Predictor; VUS, variant of unknown significance.

PI_HAT > 0.33. The average PI_HAT value across all 1,418,770 pairs of 1,685 individuals was 0.027.

Clinical data were obtained from several major sources: (1) questionnaires administered face-to-face at the beginning of the ESSE-RF study (2012), (2) fasting venous blood samples, (3) electrocardiogram (ECG) records, (4) coronary artery disease (CAD) validation in 2013, and (5) endpoint data. Additionally, (6) the available medical records of 10 patients with variants of interest in the ACMG59 genes observed at the Ivanovo Cardiology dispensary were analyzed. Blood samples for genetic analysis were stored at the Biobank of the National Medical Research Center for Therapy and Preventive Medicine (Moscow, Russia) (Pokrovskaya et al., 2019; Anisimov et al., 2021). Demographic characteristics, information about the most common non-communicable diseases, ECG parameters, and multiple blood biochemical parameters including lipid profiles are included into analysis. Endpoints such as all-cause mortality, myocardial infarctions, strokes, new cases of coronary artery disease, and revascularization are collected annually in the ESSE-RF. Clinical and endpoint data were available for all patients. Coronary artery disease was verified for participants with positive Rose Angina Questionnaire or the patient's positive answer to questions concerning previous diagnoses of either coronary artery disease or myocardial infarction. It is worth noting, however, that even with these data at hand, we were able to evaluate the individuals with known or expected pathogenic mutations for a limited number of phenotypes, for example, hypobetalipoproteinemia and prolonged QT interval.

Ethics Statement

The study was reviewed and approved by the Independent Ethic Committee of the National Medical Research Center for Therapy and Preventive Medicine (Protocol number 07-03/12 from 03.07.2012) and conducted according to the principles of the Helsinki Declaration. The participants provided their written informed consent to take part in this study.

Target Panel Design and Sequencing

We developed the targeted panel including coding exons of 242 protein-coding genes associated with cardiovascular diseases and high risk of early or sudden cardiac death, in particular, cardiomyopathy (for example, *MYBPC3*, *MYH7*, *DSP*, *LMNA,* and *DES*), channelopathy (*KCNQ1*, *KCNH2, SCN5A*, and others), and hypercholesterolemia and hypobetalipoproteinemia (*APOB*, *PCSK9*, *LDLR*, *LDLRAP1*, *ANGPTL3*, and others). Exon coordinates were expanded by 25 bp to include splice region variants. The sequenced target contained 4,311 exons with the total length of 1.04 Mbp. The following 28 genes from our panel are in the list of 59 well-characterized medically actionable disease genes recommended by the American College of Medical Genetics and Genomics for the return of incidental findings in clinical genomic sequencing (Kalia et al., 2017): *ACTA2*, *ACTC1*, *APOB*, *DSC2*, *DSG2*, *DSP*, *GLA*, *KCNH2*, *KCNQ1*, *LDLR*, *LMNA*, *MYBPC3*, *MYH11*, *MYH7*, *MYL2*, *MYL3*, *PCSK9*, *PKP2*, *PRKAG2*, *RYR1*, *RYR2*, *SCN5A*, *SDHD*, *SMAD3*, *TMEM43*, *TNNI3*, *TNNT2*, and *TPM1*. These genes comprise the cardiovascular disorder-related part of the ACMG59 panel. We refer to this subset of our target as ACMG59 genes. The full list of targeted genes with phenotypes is shown in **Supplementary Table 1**.

Sequencing Data Analysis and Variant Annotation

Target sequencing data processing and quality control evaluation were performed with the custom- designed pipeline based on GATK 3.8 and generally following the Broad institute best practices for variant calling with both standard GATK hard filters and VQSR (details in the **Supplementary Methods**). Paired-end reads were aligned to the hg19 genome sequence which is the major human reference in this project. Since the VQSR quality evaluation is more effective for the hg38, we also ran the pipeline with this reference and used the resulting filters independently: a variant was accepted for downstream analysis only if both hg37 and hg38 GATK filters flagged it as PASS and all non-reference genotype quality values were maximal (GQ = 99). This filtering retained 11,876 (84.3%) variants of the initial set of 14,087 variants and is a conservative approach aimed at minimizing the false positive variant calls.

Single-nucleotide variants and short indels were annotated with ENSEMBL Variant Effect Predictor (VEP) v.100 (McLaren et al., 2016) and cross-checked against the standalone versions of ClinVar (2021/01/10) and gnomAD (v2.1.1) databases that accumulate information on variant clinical significance and population frequencies (Landrum et al., 2016; Karczewski et al., 2020). We used ClinVar as the major source of known pathogenic and likely pathogenic variants (KP). VEP identified known variants present in the dbSNP database (Sherry et al., 2001). For each annotated variant, VEP aggregates maximal alternative allele frequency across 1000 Genomes, ESP and ExAC/gnomAD, and reports in the MAX_AF field. Allele frequencies in our dataset were also compared, where available, to those in the Russian northwest population (NWR) (Barbitoff et al., 2019).

We used LOFTEE (Karczewski et al., 2020), SIFT (Ng and Henikoff, 2006), and PolyPhen-2 (Adzhubei et al., 2010) VEP plugins to flag low-confidence protein-truncating variants (PTVs, high- impact variants in VEP) and predict potentially damaging missense variants, respectively. Missense variants simultaneously predicted as deleterious by SIFT and damaging by PolyPhen-2 are referred to as strictly damaging. We used VarSome (Kopanos et al., 2019) to perform automated variant classification according to the ACMG guidelines (Richards et al., 2015).

RESULTS

Novel Variants

We explored the spectrum of genome variation in the Russian population by sequencing 242 clinically important protein-coding genes in 1,658 unrelated individuals from the Ivanovo region. The strict quality control and filtering procedure described in the section "Materials and Methods" filtered out 2,209 variants out of the initial 14,087 keeping 11,423 SNVs and 453 short indels in the targeted regions. The observed transition/transversion Ti/Tv ratio was 3.10 for all SNVs in the exons and flanking regions and 3.48 for missense and synonymous variants suggesting the paucity of false positive variants after filtering.

Missense variants comprise the largest group by annotation (40.8%), followed by synonymous (27.2%) and intron variants (21.3%). We split the bulk of the observed variation into five groups based on their annotation and potential impact on gene function: protein-truncating variants (PTVs), strictly damaging missense variants (predicted to be damaging both by SIFT and PolyPhen-2), other missense variants, in-frame indels, and other variants, including synonymous, UTR, intron, and other variants with the lowest expected impact (**Table 1** and **Figure 1**). The fraction of strictly damaging predictions is maximal among singleton missense variants (31.9%) and gradually reduces with increasing frequency, reaching 15.1% for missense variants with allele frequency in our sample AF > 1%. This confirms the observation of a negative correlation between the alternative allele frequency and the ratio of missense variants predicted as damaging in large cohorts (Tennessen et al., 2012). Out of the total 11,876 high-quality SNV and short indels discovered in our dataset, 7,582 (63.8%) were singletons or doubletons (allele count AC < 3) with estimated alternative allele frequency not exceeding approximately 0.1%. The overwhelming majority (99.3%) of variants detected in our study in three or more copies are reported in the dbSNP and shared with other populations. The allele frequencies of these variants are very close to those in the European non-Finnish (NFE) gnomAD exomes with Pearson's correlation coefficient $R = 0.997$ (**Figure 2**).

A total of 1,356 (11.4%) variants out of 11,876 were novel, that is, not present in dbSNP 153. The fraction of novel variants was highest among the protein-truncating ones (32.4%) apparently due to the purifying selection acting on specific types of variants enriched with deleterious alleles (Lek et al., 2016). In-frame indels were not that abundant but include the second largest fraction of novel variants (18.4%). The majority (95.9%) of novel variants were not reported earlier in the survey of 694 exomes from the northwest Russia (NWR) (Barbitoff et al., 2019), the largest report on the whole exome sequencing in the Russian population to date. This is not surprising, taking in view the fact that our set is 2.4 times larger and thus allows the surveying of a wider spectrum of rare variation.

TABLE 1 | Overview of variants in the 242 targeted genes in the Ivanovo population.

	Allele count AC < 3		Allele count AC ≥ 3	
	Known	Novel (Not in NWR)	Known	Novel (Not in NWR)
Protein truncating variants	112	70 (69)	34	0 (0)
Strictly damaging missense variants	907	193 (190)	346	7 (5)
Other missense	1,957	395 (379)	1,170	4 (4)
In-frame indels	49	15 (15)	22	1 (1)
Other variants	3,227	657 (635)	2,696	14 (3)
Total	6,252	1,330	4,268	26

Novel: not in dbSNP 153.
NWR: 694 exomes from the northwest Russia (Barbitoff et al., 2019).
AC, allele count.

The novel variants in the Ivanovo population were found mostly among singletons and doubletons with only 26 variants with three-to-seven copies of alternative alleles (**Supplementary Table 2**). There were no novel protein-truncating variants other than singletons or doubletons. In **Table 2**, we report 12 selected novel variants with three or more alleles in our dataset and that were annotated as moderate impact by VEP. All these variants were missense except for a complex non-frameshifting variant in *TRIM63* that was reported by GATK as three adjacent short indels. This complex haplotype was confirmed by the visual inspection of read alignments with IGV software (Robinson et al., 2011) and reported in **Table 2** as a single in-frame indel spanning protein residues 98–109. Seven missense variants were predicted as damaging both by SIFT and PolyPhen-2 and were denoted with asterisk in **Table 2**. The remarkably large *SYNE2* gene harbored two such novel strictly damaging variants, Thr3804Asn and Glu4972Lys, each observed in three distinct individuals. Only two variants (Ser291Phe in *DMD* and Lys4037Glu in *SYNE1*) were found earlier in NWR exomes. The relatively high prevalence of these variants in the Russian population should be taken into account in the course of exome or genome sequencing in the clinical context (Richards et al., 2015).

The observed variant counts by type are tabulated for each gene in **Supplementary Table 1**. We found two genes that were significantly enriched with novel variants: *MIB1* associated with left ventricular non-compaction (OMIM: 615092), with 13 novel and 21 known variants, and *PHKA1* associated with muscle glycogenosis (OMIM:300559), with 12 novel and 17 known variants. The enrichment was validated by Fisher's exact test that compared the numbers of novel and known variants in a particular gene vs. all genes taken together, with 1,356 novel and 10,520 known variants. The *P*-values for *MIB1* and *PHKA1* were 5.2×10^{-5} and 4.0×10^{-5}, respectively. The next hit with excess of novel variants is *LAMA2* with 27 novel and 107 known variants; however, unlike the first two hits, its resulting $P = 0.002$ apparently does not pass any reasonable threshold even after the most lenient multiple test correction. Taking into account only rare known variants in comparison instead of all known variants gives similar results. The observed relative excess of rare variants may be a signature of population-specific demographic events at certain loci.

ACMG59 Genes

Our sequencing target included 28 of the 59 genes known as the ACMG59 genes, for which the reporting of known pathogenic variants was recommended by the American College of Medical Genetics and Genomics (Kalia et al., 2017). Numerous studies reported the prevalence estimates for known and expected pathogenic variants in these genes (Amendola et al., 2015; Haer-Wigman et al., 2019; Kuo et al., 2020; Lacaze et al., 2020; Van Hout et al., 2020).

In the Ivanovo sample of 1,685 individuals, six genes (*DSP, KCNQ1, MYBPC3, MYH7, RYR2, TMEM43*) harbored eight

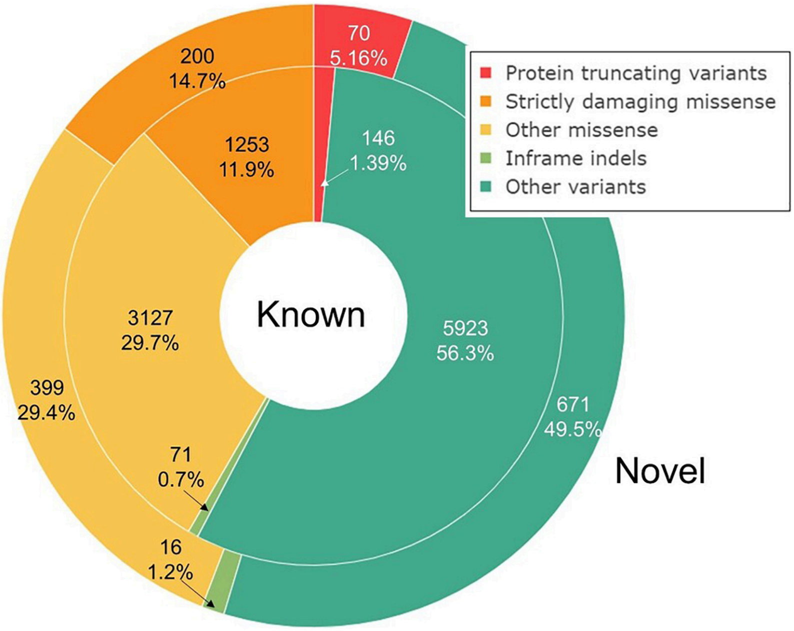


FIGURE 1 | Variant annotation. Known and novel variants are represented by inner and outer rings, respectively. Variant annotation was performed by ENSEMBL VEP. Strictly damaging variants are those predicted as damaging both by SIFT and PolyPhen-2.

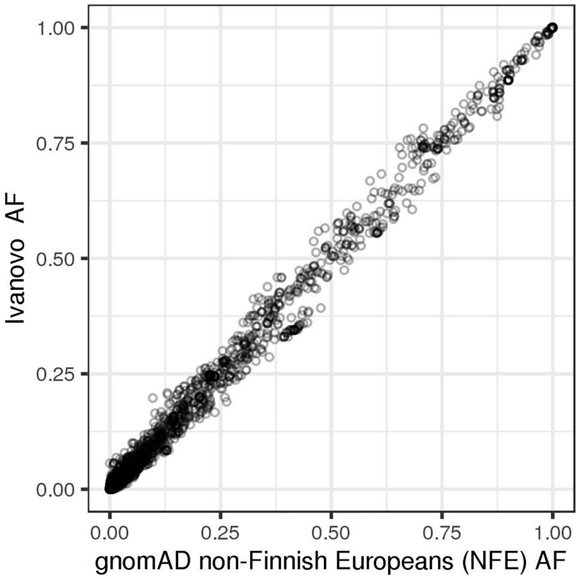


FIGURE 2 | Allele frequencies in the Ivanovo population (y-axis) compared to gnomAD non-Finnish Europeans (x-axis).

known pathogenic or likely pathogenic variants (KP). Besides, nine novel high confidence protein-truncating variants were found in seven genes: *APOB, DSP, KCNQ1, MYH7, MYH11, PCSK9,* and *RYR2* (**Table 3**). All observed genotypes involving these variants were heterozygous, and none of the 17 variants were observed in the NWR exomes.

We used VarSome (Kopanos et al., 2019) to classify KP and novel PTVs according to the ACMG criteria (Richards et al., 2015). All such variants in the ACMG59 genes were evaluated as pathogenic or likely pathogenic with two exceptions classified as variants of unknown significance (VUS). The first one is the missense substitution Thr96Arg in *KCNQ1* reported by one submitter as likely pathogenic without any supporting evidence and predicted as benign/tolerated by SIFT and PolyPhen-2. The second one is the novel splice acceptor variant in the *PCSK9* gene (ENST00000302118.5:c.1864-2A > T), considered high quality by LOFTEE and predicted deleterious by multiple tools. The LDL-C value of the carrier is 2.64 mmol/l which is in the normal range. In both cases, variants were classified by VarSome as VUS but not excluded from our analysis.

For all seven genes with novel PTVs, protein truncation is known to be an established disease-causing mechanism: in particular, no benign PTVs were reported in ClinVar for *KCNQ1*, *MYH11,* and *MYH7* and only one instance of benign PTV in *DSP*. In *PCSK9*, truncating mutations are associated with low plasma levels of low-density lipoprotein cholesterol (LDL-C) (Cohen et al., 2005). Six of eight known KPs were observed as singletons and 2 in 3 individuals each, giving 12 individuals in total which comprises 0.71% of the total sample of 1,685 participants. The initial set of variant calls included a KP variant in *RYR2* (rs794728721), but it was filtered out by both hg37 and hg38 GATK filters and was not included in the filtered variant set. Eight novel truncating variants were singletons, and one was observed in two participants, which gave the extra 10 participants harboring novel protein-truncating variants. The genes also harbored nine rare PTVs with worldwide maximal allele frequency below 0.1% as reported by VEP and the total of

TABLE 2 | Novel variants with three or more alleles in our dataset annotated by VEP as high or moderate impact.

Gene	Variant	HGVS	VarSome	Carriers
HDL deficiency, OMIM: 604091				
ABCA1	chr9:107581932_G/A	ENSP00000363868.3:p.Pro1059Leu*	LP	3
Dilated cardiomyopathy, OMIM: 302045 (XL)				
DMD	chrX:32716075_G/A	ENSP00000354923.3:p.Ser291Phe*	VUS	5
Hypertrophic cardiomyopathy				
ILK	chr11:6625538_G/T	ENSP00000379975.2:p.Ala13Ser	VUS	3
MYOM1	chr18:3086053_T/A	ENSP00000348821.4:p.Thr1412Ser	VUS	3
TRIM63	chr1:26392766-26392799	N/A	N/A	7
TRIM63	chr1:26392801_A/T	ENSP00000363390.3:p.Leu97Gln	VUS	7
TTN	chr2:179455887_T/C	ENSP00000467141.1:p.Met20189Val	VUS	3
Arrhythmogenic right ventricular dysplasia, OMIM: 600996 (AD)				
RYR2	chr1:237777580_A/G	ENSP00000355533.2:p.Arg1718Gly*	LP	3
TGFB3	chr14:76429816_C/A	ENSP00000238682.3:p.Asp257Tyr	VUS	4
Emery–Dreifuss muscular dystrophy, OMIM: 612998 (AD)				
SYNE1	chr6:152665332_T/C	ENSP00000356224.5:p.Lys4037Glu*	VUS	3
SYNE2	chr14:64548225_C/A	ENSP00000350719.3:p.Thr3804Asn*	VUS	3
SYNE2	chr14:64606729_G/A	ENSP00000350719.3:p.Glu4972Lys*	VUS	3

Genes are grouped based on associated disorders.
HGVS: variant description; strictly damaging missense variants are denoted with asterisk.
VarsSome: variant classification by VarSome according to the ACMG criteria.
P, pathogenic; LP, likely pathogenic; VUS, variant of unknown significance.
Carriers: number of individuals harboring the variant.

14 carriers in the Ivanovo population (**Supplementary Table 3**). The 24 participants with novel or rare PTVs gave an extra 1.4% of the total sample of 1,685 participants. Taking in view our gene sequencing target included approximately one-half of the current list of 56 dominant ACMG59 genes (excluding *ATP7B*, *MUTYH*, and *PMS2*), one may approximate the prevalence of dominant KP variants and novel or rare PTVs in the Ivanovo population at 1.4 and 2.8%, respectively.

All Genes

Evaluating the complete set of 242 sequenced genes, we identified 85 pathogenic or likely pathogenic ClinVar variants in 44 genes: 36 protein-truncating, 47 missense with 36 of them predicted to be strictly damaging, and two intron variants (**Supplementary Table 3**). This set of KP variants included eight variants in the ACMG59 genes described above. All observed genotypes involving known pathogenic or novel variants discussed in this section were heterozygous. Among the 77 known KP in non-ACMG59 genes, 9 and 63 were associated with dominant (AD) and recessive (AR) diseases, respectively (**Table 4**).

Known variants in dominant non-ACMG genes were mostly present in one participant, except for the missense Pro279Leu variant in the *MEF2A* gene (rs121918529) observed in three participants and associated with CAD/myocardial infarction. As expected, known disease variants in recessive genes were more prevalent, with 12 recessive variants observed in three or more participants. Among them were two variants in *GAA* (rs375470378) and *PMM2* (rs28936415), found in eight participants each. One protein-truncating variant in *APOC3* without any assigned specific inheritance mode (rs138326449) was classified as pathogenic and associated with coronary artery disease and hyperalphalipoproteinemia by a single submitter in ClinVar (VCV000139560.3) and found in nine individuals.

The full set of 242 targeted genes also contained 69 novel PTVs in 48 genes (**Table 4** and **Supplementary Table 3**), all being singletons, except for the two doubleton cases: the frameshift variant ENSP00000347507.3:p.Lys1173ArgfsTer41 in the ACMG59 gene *MYH7* and the stop gain variant ENSP00000364979.4:p.Arg1063Ter in the non-ACMG gene *COL4A1*. Unlike known pathogenic variants, novel PTVs did not exhibit any detectable enrichment or depletion in the dominant or recessive genes, suggesting that their functional role may require thorough evaluation.

We used VarSome (Kopanos et al., 2019) to classify KP and novel PTVs according to the ACMG criteria (Richards et al., 2015). All variants were evaluated as pathogenic or likely pathogenic with few exceptions: five KP variants and six novel PTVs were classified by VarSome as VUS, and one KP variant as likely benign (**Supplementary Table 3**). The latter is the missense Ala296Thr variant rs80356462 in the *SIX5* gene, classified by ClinVar as pathogenic, with no assertion criteria provided (ClinVar id: VCV000008599.1) and predicted benign by multiple computational tools, including SIFT and PolyPhen-2.

With novel PTVs observed in 48 genes, every fifth of the targeted genes in the Ivanovo sample contained a novel heterozygous PTV, although only in one or two individuals from the studied population. These genes included eight genes which are confidently depleted for PTV variation in gnomAD (Karczewski et al., 2020): *DSP, RYR2, COL4A1, EYA1, FBN2, MYH11, NNT,* and *SVEP1*. The most PTV-depleted gene in this list is the collagen type IV alpha 1 chain *COL4A1* with only

TABLE 3 | Known pathogenic or likely pathogenic and novel protein-truncating variants in the 28 genes from the ACMG59 list.

Gene	Variant	HGVS	VarSome	Carriers
Hypercholesterolemia, OMIM:144010 (AD); Hypobetalipoproteinemia, OMIM: 615558 (AR)				
APOB	chr2:21232683_G/A	ENSP00000233242.1:p.Gln2353Ter	P	1
APOB	chr2:21234967_GA/G	ENSP00000233242.1:p.Phe1591SerfsTer19	P	1
APOB	chr2:21260870_AC/A	ENSP00000233242.1:p.Val166PhefsTer66	P	1
Dilated cardiomyopathy, OMIM: 615821 (AD)				
DSP	chr6:7580077_GACCA/G	ENSP00000369129.3:p.Thr1219ProfsTer30	P	1
DSP	rs121912997	ENSP00000369129.3:p.Arg1267Ter	P	1
Long QT syndrome, OMIM: 192500 (AD)				
KCNQ1	chr11:2466373_G/A	ENSP00000155840.2:p.Trp15Ter	P	1
KCNQ1	rs1337409061	ENSP00000155840.2:p.Thr96Arg	VUS	3
KCNQ1	rs199472814	ENSP00000155840.2:p.Arg591Leu	LP	1
KCNQ1	rs199473411	ENSP00000155840.2:p.Arg366Trp	P	1
Hypertrophic cardiomyopathy, OMIM: 115197 (AD, AR), OMIM: 192600 (AD)				
MYBPC3	rs376395543	ENST00000545968.1:c.26-2A > G	P	3
MYH7	chr14:23889261_CT/C	ENSP00000347507.3:p.Lys1173ArgfsTer41	P	2
MYH7	rs121913650	ENSP00000347507.3:p.Arg1712Trp	P	1
Aortic aneurysm, OMIM: 132900 (AD)				
MYH11	chr16:15917230_A/T	ENSP00000379616.3:p.Tyr128Ter	P	1
Hypercholesterolemia; low level of LDL-C, OMIM: 603776 (AD)				
PCSK9	chr1:55529040_A/T	ENST00000302118.5:c.1864-2A > T	VUS	1
Arrhythmogenic right ventricular dysplasia, OMIM: 600996 (AD), 604400 (AD)				
RYR2	chr1:237433817_G/GT	ENSP00000355533.2:p.Cys24LeufsTer61	P	1
RYR2	rs753850982	ENSP00000355533.2:p.Gly4095Ser	LP	1
TMEM43	rs63750743	ENSP00000303992.4:p.Ser358Leu	P	1

Variant: dbSNP rsID for known variants or chr:pos_ref/alt identifier for novel PTVs.
HGVS: variant description.
VarSome: variant classification by VarSome according to the ACMG criteria.
P, pathogenic; LP, likely pathogenic; VUS, variant of unknown significance.
Carriers: number of Ivanovo individuals harboring the variant.

six PTVs observed in gnomAD. The novel protein-truncating variant Arg1063Ter in this gene is observed in two participants.

We also detected 82 protein-truncating variants with worldwide maximal allele frequency below 0.1% as reported by VEP and refer to them as rare PTVs (**Supplementary Table 3**). Twenty-three of them (28%) are reported in ClinVar as VUS or variants with conflicting interpretation. Most of the rare PTV alleles are rare in the Ivanovo population as well, with only six seen in 3–6 individuals and one (rs201068740) in 14 individuals, all heterozygous.

Overrepresented Known Pathogenic and Likely Pathogenic Variants

For the 16 KP variants harbored by three or more individuals, we compared the allele frequencies in our population with gnomAD v.2.1.1 non-Finnish Europeans (NFE). Since three or more copies of an allele correspond to population frequencies roughly equal or exceeding 0.1% in our sample with 1,685 individuals, we expected that KP variants observed in the Ivanovo population more frequently than in NFE would be present among such variants. Indeed, we found two dominant and eight recessive pathogenic variants with allele frequencies 6.7–67.2 times greater than in the NFE aggregated genomes and exomes allele counts (**Table 5**). In each case, the significance of the observed frequency difference was validated with the Fisher's test on direct allele counts. The variants were accepted if the corresponding *P*-values did not exceed 0.05 after Benjamini–Hochberg correction (**Supplementary Table 4**).

Missense substitution Thr96Arg in *KCNQ1* (dbSNP: rs1337409061) was recently submitted to ClinVar as likely pathogenic and associated with long QT syndrome (ClinVar id: VCV000983021.1). This allele was observed in four copies only in non-Finnish Europeans with the frequency of 3.4×10^{-5} in that population. With three harboring participants, the frequency in the Ivanovo population equals 8.9×10^{-4} with lower and upper 95% confidence interval (CI) margins 2.0×10^{-4} and 2.6×10^{-3}, respectively. The Thr96Arg substitution, however, was classified by VarSome as VUS, had no supporting evidence for pathogenicity in ClinVar, and was not observed among the variants discovered in 9 and 53 unrelated Russian families with the long QT syndrome (Polyak et al., 2016; Maltese et al., 2017). The presence of this variant in the general Russian population without any prevalence in patients may suggest that its pathogenicity needs to be confirmed by future independent submissions.

Pathogenic splice acceptor variant c.26-2A > G in *MYBPC3* (rs376395543) is associated with hypertrophic cardiomyopathy

TABLE 4 | Known pathogenic or likely pathogenic and novel protein-truncating variants in the full set of 242 targeted genes.

	KP variants			Novel PTVs		
Genes	**Variants**	**Genes**	**Carriers**	**Variants**	**Genes**	**Carriers**
ACMG (28)	8	6	12	9	7	10
Non-ACMG AD (78)	9	8	11	18	11	19
Non-ACMG AR (68)	63	26	124	20	16	20
Other (68)	5	4	13	22	15	22
Total	85	44		69	49	

KP, known pathogenic or likely pathogenic variants from ClinVar.
PTV, protein-truncating variants.
The full set of targeted genes included 28 genes from the ACMG59 list; among the non-ACMG genes, 78 are dominant, 68 recessive, and 68 with other types of inheritance.

(VCV000042644.10) and is observed in six copies in non-Finnish Europeans (frequency 5.2×10^{-5}) and one allele copy in Latino Americans. This variant was observed in three individuals in our set, with population frequency estimated as 8.9×10^{-4} (CI from 2.0×10^{-4} to 2.6×10^{-3}), not observed in NWR exomes and classified as pathogenic by VarSome.

Table 5 also contains eight pathogenic recessive variants which were significantly overrepresented in the Ivanovo population compared to gnomAD NFE. The most frequent one harbored by eight participants in our study was the acid alpha-glucosidase *GAA* intronic variant c.1552-3C > G (rs375470378) associated with glycogen storage disease type II, also known as Pompe disease. This intronic variant is reported in ClinVar as pathogenic or likely pathogenic by multiple submitters (VCV000419722.9) and likely pathogenic by VarSome. The observed frequency in the Ivanovo population (0.24%) significantly exceeds the NFE (0.03%) and other population frequencies, with the exception of the Estonian population where the observed frequency is 3/4480 (0.07%) with marginally significant difference from the Ivanovo population (Fisher's test $P = 0.045$) likely due to modest study sample sizes. The northwest Russia frequency is estimated as 0.14%, suggesting that this variant is mostly prevalent in Northern Europe. Glycogen storage disease type II (Pompe disease) is an autosomal recessive metabolic disorder which damages muscle and nerve cells and results from the accumulation of glycogen in the lysosome due to the deficiency of the alpha-glucosidase enzyme. The worldwide prevalence of this disease is in the range 1:14,000 to 1:300,000 (Van der Ploeg and Reuser, 2008), with no estimates for Russia available (Semyachkina et al., 2014). It was also hypothesized that the overwhelming majority of Pompe disease cases in Russia may not be diagnosed (Nikitin, 2016). By combining allele counts in our study and the NWR, we calculated the frequency of homozygous carriers of the disease-causing *GAA* variant rs375470378 as 1:226,000 which is the lower estimate of Pompe disease prevalence in Russia due to the existence of other pathogenic alleles.

The largest frequency difference between NFE and our sample was observed for the alpha-2 laminin *LAMA2* frameshift deletion c.7536delC (rs398123387) associated with merosin-deficient congenital muscular dystrophy type 1A (MDC1A) and segregating in the Ivanovo population in four copies (0.11%), which is approximately 67 times greater than the NFE frequency of 0.0017%. This variant was reported earlier in a single Russian patient (Dadali et al., 2010) and later in 21% of all affected chromosomes in the sample of 29 unrelated MDC1A patients, suggesting that this is the most abundant disease-causing *LAMA2* allele in the Russian population (Milovidova et al., 2018). Based on the analysis of four microsatellite markers in the *LAMA2* region of chromosome 6, the authors suggested that this mutation belongs to a founder haplotype spanning at least 3.2 cM. Our data seem to confirm this assumption and suggest the frequency range (0.11% with CI from 0.03 to 0.30%) for this variant in the unaffected population.

The frameshift variant c.845_846del in the Cytochrome C oxidase assembly Factor *SURF1* (rs782316919) is associated with Leigh syndrome due to mitochondrial complex IV deficiency (VCV000012770.12) and has been reported earlier as the most prevalent *SURF1* disease allele among Russian, Ukrainian, and Polish patients (Piekutowska-Abramczuk et al., 2009;

TABLE 5 | KP variants with frequencies in the Ivanovo population significantly exceeding the NFE.

Gene, phenotype	Variant	HGVS	Carriers	Ivanovo/gnomAD ratio
KCNQ1, Long QT syndrome (AD, 192500)	rs1337409061	ENSP00000155840.2:p.Thr96Arg	3	25.7
MYBPC3, Hypertrophic cardiomyopathy (AD, 115197)	rs376395543	ENST00000545968.1:c.26-2A > G	3	17.2
GAA, Glycogen storage disease (Pompe disease) (AR, 232300)	rs375470378	ENST00000302262.3:c.1552-3C > G	8	8.8
GLB1, GM1-gangliosidosis (AR, 253010, 230600)	rs376663785	ENSP00000306920.4:p.Tyr270Asp	4	25.4
LAMA2, Merosin-deficient congenital muscular dystrophy type 1A (AR, 607855)	rs398123387	ENST00000421865.2:c.7536del	4	67.2
MTO1, Combined oxidative phosphorylation deficiency (AR, 614702)	rs201544686	ENSP00000402038.2:p.Arg517His	6	7.7
SURF1, Mitochondrial complex IV deficiency, Leigh syndrome (AR, 220110)	rs782316919	ENST00000371974.3:c.845_846del	4	8.0
ALMS1, Alstrom syndrome (AR, 203800)	rs797045228	ENST00000264448.6:c.4150dup	3	19.0
ALMS1, Alstrom syndrome (AR, 203800)	rs747272625	ENST00000264448.6: c.11310_11313del	3	16.7
SCO2, Cardioencephalo-myopathy (AR, 604377)	rs74315511	ENSP00000444433.1: p.Glu140Lys	4	6.7

Carriers: number of Ivanovo individuals harboring the variant.
Ivanovo/gnomAD is the ratio of allele frequencies in our sample and gnomAD NFE genomes and exomes considered together.

Tsygankova et al., 2010). Our frequency estimate of 0.11% is comparable with those given earlier for the NWR and Poland (0.22 and 0.28%, respectively) and is by order of magnitude higher than that in the NFE (0.014%), confirming the hypothesis of the Eastern European origin of this variant (Piekutowska-Abramczuk et al., 2009).

The missense variant Arg517His in the *MTO1* gene associated with combined oxidative phosphorylation deficiency (rs201544686, VCV000089037.4) and c.4150dup duplication in the *ALMS1* gene associated with the Alstrom syndrome (rs797045228, VCV000210127.8) were reported earlier among 14 known pathogenic variants most prevalent in NWR (Barbitoff et al., 2019). It is noteworthy that we discovered another pathogenic variant in *ALMS1* (rs747272625, VCV000550797.4) detected in three other participants and overrepresented in the Ivanovo population compared to NFE (0.09% vs. 0.005%). The missense Tyr270Asp variant in *GLB1* is associated with GM1-gangliosidosis (rs376663785, VCV000284172.5) and was observed in three copies in the Ivanovo population. Finally, the missense substitution Glu140Lys in *SCO2* is associated with cardioencephalomyopathy, also known as mitochondrial complex IV deficiency (rs74315511, VCV000005681.11). The observed Ivanovo frequency for this allele (0.0012) is 6.7 greater than the overall NFE frequency (0.00017) and most close to that of the Estonian subset of NFE (0.0008) and NWR (0.0007), suggesting the putative geographical origin of this variant.

Phenotype Evaluation

We analyzed the available clinical data of participants with detected pathogenic/likely pathogenic variants or novel protein-truncating variants in 28 of ACMG59 genes and other genes with autosomal dominant inheritance. For the ACMG59 genes, we also evaluated phenotypes in the carriers of rare protein-truncating variants (VEP MAX_AF < 0.1%). Variants subject to carrier evaluation are marked in **Supplementary Table 3**; the results are presented in **Table 6**. For the first group that contained 17 KP variants in 23 carriers, we found five individuals with matching phenotypes: prolonged QTc interval (Arg366Trp in *KCNQ1* and Arg858His in *CACNA1C*), hypertriglyceridemia (Gln313Ter in *APOA5*), CAD (Pro279Leu in *MEF2A*), and cerebral infarction (Arg153Cys in *NOTCH3*).

The second group, novel PTVs, comprised 27 variants in 27 carriers and yielded four individuals: three with hypobetalipoproteinemia, all with various novel truncating mutations in *APOB*, and one individual with hypertrophic cardiomyopathy and frameshift deletion in *MYH7* (**Table 6**). Various truncated forms of *APOB* have been found earlier to segregate with the familial hypobetalipoproteinemia phenotype (OMIM:615558). Missense mutations are the most prevalent cardiomyopathy-related pathogenic mechanism in *MYH7*; however, ClinVar reports 20 cases of pathogenic and likely pathogenic truncating variants with various degrees of confidence.

Finally, nine rare PTVs in the ACMG59 genes were harbored by 14 carriers, two of which displayed expected phenotypes: myopathy (frameshift deletion in *RYR1*) and prolonged QTc interval (frameshift insertion in *KCNQ1*). We also checked 17 carriers of 18 novel truncating variants in non-ACMG59 genes and did not discover any remarkable genotype–phenotype associations.

Out of the total 11 genotype–phenotype associations described above, 7 are in the ACMG59 genes: 1 among KP variants, 4 among novel PTVs, and 2 among the rare PTVs. Taking in view that the total numbers of carriers checked in these three groups were 12, 10, and 14, respectively (**Supplementary Table 3**), we estimated the overall success rate of phenotype evaluation in the ACMG59 genes as 19% (=7/36). It is worth noting that the highest phenotype confirmation rate (40%) was observed in the subset of novel PTVs, thus emphasizing the importance of studying this class of variation by deeper sequencing the individuals from the Russian population.

In the course of the analysis of the study endpoints, we found a female participant who died of heart failure due to dilated cardiomyopathy. The manual review of rare candidate variants harbored by this individual discovered the singleton missense variant Thr891Met in the *FLNC* gene (rs766023596). This substitution was submitted to ClinVar (VCV000855393.2) by a single submitter as VUS associated with this type of cardiomyopathy (Cui et al., 2018). Our observation provides evidence suggesting the pathogenicity of this variant.

DISCUSSION

The overwhelming majority of variants detected in our study in three or more copies were known and shared with other populations. The alternative allele frequencies of these variants were highly correlated with non-Finnish European population. This is in agreement with the earlier observation that in the case of a smaller sample set of 694 individuals from the northwest region of Russia, the majority of disease alleles were shared between Russian and European populations (Barbitoff et al., 2019). However, with the sample size of 1,685 individuals, we detected a considerable fraction (11.4%) of novel variants not present in the dbSNP and not reported earlier in the Russian population. Most such variants were singletons and doubletons in our sample, with the exceptions discussed above: only 26 novel alleles were observed in three-to-seven copies, with no protein-truncating variants among them. Our results therefore suggest that virtually no “private” alleles specific for the Russian population reach the frequencies above 0.1–0.2%

The fraction of novel variants was highest among the protein-truncating ones: out of 218 PTVs discovered in our dataset, 70 are novel and, with one exception, not reported before in the Russian population. This effect can be explained by the purifying selection acting on specific classes of variants enriched with deleterious alleles. Since rare and novel protein-truncating variants are of special interest in the context of clinical genome interpretation (DeBoever et al., 2018), our study emphasizes the importance of deeper sequencing, taking in view the size and ethnic diversity of the Russian population.

Approximately 18% of 242 targeted genes harbored known pathogenic or likely pathogenic variants reported earlier

TABLE 6 | Variants with confirmed phenotypes.

Gene	ACMG	Variant	HGVS	Phenotype (Source)
I. Known pathogenic or likely pathogenic: 17 variants, 23 carriers				
KCNQ1	Yes	rs199473411 VCV000052955.8	ENSP00000155840.2: p.Arg366Trp	Prolonged QTc interval, QT = 453 mc for male (ECG)
CACNA1C	No	rs786205753 VCV000190653.57	ENSP00000266376.6: p.Arg858His	Prolonged QTc interval, QT = 453 mc for female (ECG)
APOA5	No	rs147528707 VCV000420172.2	ENSP00000445002.1: p.Gln313Ter	Hypertriglyceridemia, TG = 14.3 mmol/l (Biochemical assay)
MEF2A	No	rs121918529 VCV000008949.1	ENSP00000346389.5: p.Pro279Leu	CAD (CAD validation)
NOTCH3	No	rs797045014 VCV000208501.2	ENSP00000263388.1: p.Arg153Cys	Cerebral infarction (Endpoint)
II. Novel protein truncating: 27 variants, 27 carriers				
APOB	Yes	chr2:21232683_G/A	ENSP00000233242.1: p.Gln2353Ter	Hypobetalipoproteinemia, LDL-Ñ = 1.47 mmol/l (Biochemical assay)
APOB	Yes	chr2:21234967_GA/G	ENSP00000233242.1: p.Phe1591SerfsTer19	Hypobetalipoproteinemia, LDL-Ñ = 0.95 mmol/l (Biochemical assay)
APOB	Yes	chr2:21260870_AC/A	ENSP00000233242.1: p.Val166PhefsTer66	Hypobetalipoproteinemia, LDL-Ñ = 0.72 mmol/l (Biochemical assay)
MYH7	Yes	chr14:23889261_CT/C	ENSP00000347507.3: p.Lys1173ArgfsTer41	Hypertrophic cardiomyopathy (Medical record)
III. Rare protein truncating: 9 variants, 14 carriers (ACMG59 genes only)				
RYR1	Yes	rs797045932 VCV000212099.3	ENSP00000352608.2: p.Pro836LeufsTer48	Myopathy (Medical record)
KCNQ1	Yes	rs397508104 VCV000053027.10	ENSP00000155840.2: p.Arg632GlnfsTer20	Prolonged QTc interval, QT = 458 mc for female (ECG)

Variant: dbSNP rsID for known variants or chr:pos_ref/alt identifier for novel PTVs.
HGVS: variant description.
Phenotype: disease phenotype confirmed by evaluation of clinical data; source of clinical data is specified in the parentheses.

in ClinVar. Similar to the novel variants, most of disease-causing alleles were singletons or doubletons in our sample. We report nine variants with disease alleles significantly more frequent in the Ivanovo population compared to the non-Finnish Europeans. The largest frequency difference between NFE and our sample (>67x) was observed for the alpha-2 laminin *LAMA2* frameshift deletion c.7536delC (rs398123387) associated with merosin-deficient congenital muscular dystrophy type 1A (MDC1A). It was suggested earlier that this deletion is the most abundant disease-causing *LAMA2* mutation in the Russian population residing on the founder haplotype spanning at least 3.2 cM (Milovidova et al., 2018). We estimated the frequency range for this variant in the unaffected population as 0.11 with 95% CI range from 0.03 to 0.30%.

To the best of our knowledge, this is the first report of the spectrum of known and putative pathogenic variants in the ACMG59 genes in the Russian population. We approximated the prevalence of known dominant pathogenic variants as 1.4% which is in agreement with the range of 1.3−2.7% suggested recently for other European populations (Haer-Wigman et al., 2019; Kuo et al., 2020; Lacaze et al., 2020; Van Hout et al., 2020). Since our target included only 28 genes of the ACMG59 list, further studies may provide more precise estimates of KP variant prevalence in these genes in the Russian population. Our results also emphasize the importance of sequencing representative cohorts (>1,000 individuals) to uncover the population-specific genetic variation of clinical relevance.

We observed the incomplete penetrance of known pathogenic variants in the ACMG59 genes: only 1 individual out of 12 with KP variants in these genes had the phenotype most likely related to the variant. When known pathogenic and novel or rare protein-truncating variants were considered together, the overall rate of confirmed phenotypes was about 19%, with maximum in the subset of novel PTVs. This is in overall agreement with the other studies: in the population cohort from the Rotterdam Study, there were 13% of the carriers of known pathogenic variants in the ACMG59 genes who had phenotype correlated with genotype (Van Rooij et al., 2020). The Rotterdam study was based on a larger sample size of 2,628 individuals and involved an in-depth analysis of the phenotypes. The Rotterdam study included persons of age 55 and above, while in our study, the age of participants was above 25, which enabled the identification of individuals with earlier disease onset.

Hou et al. (2020) reported results of the whole-genome sequencing of 1,190 self-referred volunteers with a median age of 54 years (range 20–89+ years, 70.6% European) and showed that deeper phenotyping (metabolomics, advanced imaging, and clinical laboratory tests in addition to family/medical history) significantly increased the concordance between genetic and phenotypic data. Genotype–phenotype associations were identified in 66.5% of the participants with pathogenic or likely pathogenic variants, not restricted to the ACMG59 genes. Overall, 44.5% of the cases were revealed through genomics and metabolomics analysis and had phenotype manifestations affecting serum metabolite levels. We hypothesize that the volume of the available clinical information is the limiting factor for our ability to observe better concordance between genotype and phenotype. In particular, results of imaging tests, which are essential components of cardiomyopathy diagnostics, were not available for all Ivanovo participants. The medical records, which usually contain more detailed clinical information about a patient, were very useful in our analysis and enabled identification of two genotype–phenotype associations in 10 variant carriers with available medical records. In order to develop a procedure for the return of secondary findings in the ACMG59 genes, we are designing an in-depth clinical examination targeted at all study participants with detected known and expected pathogenic variants in these genes and the genetic cascade screening of their families.

This work presents the analysis of targeted sequencing of the largest sample of the general Russian population to date. The observed spectrum of genetic variants in a region with predominantly Russian population is close in genetic composition to European populations, with the exception of a rare variation. We believe that our results provide a valuable reference for the clinical interpretation of genome sequencing and are the first step toward creating a comprehensive reference of genetic variability observed in the Russian Federation.

DATA AVAILABILITY STATEMENT

Aggregated variant frequency information can be requested from the authors. Individual genotype information cannot be made available in order to protect participant privacy.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the independent Ethics Committee of the National Medical Research Center for Therapy and Preventive Medicine (Protocol numbers 07-03/12 from 03.07.2012). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

OD, AM, AE, SS, and VR conceived and designed the study. VR, AE, AZ, MZ, AK, and YV analyzed the data and wrote the manuscript. IE, GM, OB, SR, and MP carried out participant and clinical data management. AK, ES, OK, OS, and MD performed the target design and sequencing. All authors contributed to the article, revised the manuscript, and approved the submitted version.

FUNDING

This study was supported by the State assignment no. 121021100127-8. VR acknowledges support by the RFBR and DFG joint research project no 20-54-12008.

ACKNOWLEDGMENTS

The authors are grateful to the participants from Ivanovo for their contributions and support of this research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.709419/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res.* 25, 305–315. doi: 10.1101/gr.183483.114

Amendola, L. M., Jarvik, G. P., Leo, M. C., McLaughlin, H. M., Akkari, Y., Amaral, M. D., et al. (2016). Performance of ACMG-AMP variant-interpretation guidelines among nine laboratories in the clinical sequencing exploratory research consortium. *Am. J. Hum. Genet.* 98, 1067–1076. doi: 10.1016/j.ajhg.2016.03.024

Anisimov, S. V., Meshkov, A. N., Glotov, A. S., Borisova, A. L., Balanovsky, O. P., Belyaev, V. E., et al. (2021). National association of biobanks and biobanking specialists: new community for promoting biobanking ideas and projects in Russia. *Biopreserv. Biobank.* 19, 73–82. doi: 10.1089/bio.2020.0049

Barbitoff, Y. A., Skitchenko, R. K., Poleshchuk, O. I., Shikov, A. E., Serebryakova, E. A., Nasykhova, Y. A., et al. (2019). Whole-exome sequencing provides insights into monogenic disease prevalence in Northwest Russia. *Mol. Genet. Genomic Med.* 7:e964. doi: 10.1002/mgg3.964

Boitsov, S. A., Chazov, E. I., Shlyakhto, E. V., Shalnova, S. A., Konradi, A. O., Karpov, Y. A., et al. (2013). Epidemiology of cardiovascular diseases in different regions of Russia (ESSE-RF). The rationale for and design of the study. *Profilakticheskaya Med.* 16, 25–34.

Brovkina, O. I., Shigapova, L., Chudakova, D. A., Gordiev, M. G., Enikeev, R. F., Druzhkov, M. O., et al. (2018). The ethnic-specific spectrum of germline nucleotide variants in DNA damage response and repair genes in hereditary breast and ovarian cancer patients of tatar descent. *Front. Oncol.* 8:421. doi: 10.3389/fonc.2018.00421

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

Chen, R., Shi, L., Hakenberg, J., Naughton, B., Sklar, P., Zhang, J., et al. (2016). Analysis of 589,306 genomes identifies individuals resilient to severe Mendelian childhood diseases. *Nat. Biotechnol.* 34, 531–538. doi: 10.1038/nbt.3514

Cohen, J., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K., and Hobbs, H. H. (2005). Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* 37, 161–165. doi: 10.1038/ng1509

Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C., and Kehrer-Sawatzki, H. (2013). Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum. Genet.* 132, 1077–1130. doi: 10.1007/s00439-013-1331-2

Cui, H., Wang, J., Zhang, C., Wu, G., Zhu, C., Tang, B., et al. (2018). Mutation profile of FLNC gene and its prognostic relevance in patients with hypertrophic cardiomyopathy. *Mol. Genet. Genomic Med.* 6, 1104–1113. doi: 10.1002/mgg3.488

Dadali, E. L., Rudenskaia, G. E., Shchagina, O. A., Tiburkova, T. B., Sukhorukov, V. S., Kharlamov, D. A., et al. (2010). [Merosin-deficient congenital muscular dystrophy]. *Z. Nevrol. Psikhiatrii* 110, 83–89.

DeBoever, C., Tanigawa, Y., Lindholm, M. E., McInnes, G., Lavertu, A., Ingelsson, E., et al. (2018). Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* 9:1612. doi: 10.1038/s41467-018-03910-9

Dewey, F. E., Grove, M. E., Pan, C., Goldstein, B. A., Bernstein, J. A., Chaib, H., et al. (2014). Clinical interpretation and implications of whole-genome sequencing. *JAMA* 311:1035. doi: 10.1001/jama.2014.1717

Dewey, F. E., Murray, M. F., Overton, J. D., Habegger, L., Leader, J. B., Fetterolf, S. N., et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* 354:aaf6814. doi: 10.1126/science.aaf6814

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., Gallego, C. J., et al. (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* 93, 631–640. doi: 10.1016/j.ajhg.2013.08.006

Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., et al. (2013). ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet. Med.* 15, 565–574. doi: 10.1038/gim.2013.73

Haer-Wigman, L., van der Schoot, V., Feenstra, I., Vulto-van Silfhout, A. T., Gilissen, C., Brunner, H. G., et al. (2019). 1 in 38 individuals at risk of a dominant medically actionable disease. *Eur. J. Hum. Genet.* 27, 325–330. doi: 10.1038/s41431-018-0284-2

Hou, Y.-C. C., Yu, H.-C., Martin, R., Cirulli, E. T., Schenker-Ahmed, N. M., Hicks, M., et al. (2020). Precision medicine integrating whole-genome sequencing, comprehensive metabolomics, and advanced imaging. *Proc. Natl. Acad. Sci. U.S.A.* 117, 3053–3062. doi: 10.1073/pnas.1909378117

Jain, A., Gandhi, S., Koshy, R., and Scaria, V. (2018). Incidental and clinically actionable genetic variants in 1005 whole exomes and genomes from Qatar. *Mol. Genet. Genomics* 293, 919–929. doi: 10.1007/s00438-018-1431-8

Kalia, S. S., Adelman, K., Bale, S. J., Chung, W. K., Eng, C., Evans, J. P., et al. (2017). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet. Med.* 19, 249–255. doi: 10.1038/gim.2016.190

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Kiseleva, A. V., Klimushina, M. V., Sotnikova, E. A., Divashuk, M. G., Ershova, A. I., Skirko, O. P., et al. (2020). A data-driven approach to carrier screening for common recessive diseases. *J. Pers. Med.* 10:140. doi: 10.3390/jpm10030140

Kopanos, C., Tsiolkas, V., Kouris, A., Chapple, C. E., Albarca Aguilera, M., Meyer, R., et al. (2019). VarSome: the human genomic variant search engine. *Bioinformatics* 35, 1978–1980. doi: 10.1093/bioinformatics/bty897

Kulikova, O., Brodehl, A., Kiseleva, A., Myasnikov, R., Meshkov, A., Stanasiuk, C., et al. (2021). The desmin (DES) mutation p.A337P is associated with left-ventricular non-compaction cardiomyopathy. *Genes* 12:121. doi: 10.3390/genes12010121

Kuo, C., Hwu, W., Chien, Y., Hsu, C., Hung, M., Lin, I., et al. (2020). Frequency and spectrum of actionable pathogenic secondary findings in Taiwanese exomes. *Mol. Genet. Genomic Med.* 8:e1455. doi: 10.1002/mgg3.1455

Kwak, S. H., Chae, J., Choi, S., Kim, M. J., Choi, M., Chae, J.-H., et al. (2017). Findings of a 1303 Korean whole-exome sequencing study. *Exp. Mol. Med.* 49:e356. doi: 10.1038/emm.2017.142

Lacaze, P., Sebra, R., Riaz, M., Tiller, J., Revote, J., Phung, J., et al. (2020). Medically actionable pathogenic variants in a population of 13,131 healthy elderly individuals. *Genet. Med.* 22, 1883–1886. doi: 10.1038/s41436-020-0881-7

Landrum, M. J., Lee, J. M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* 44, D862–D868. doi: 10.1093/nar/gkv1222

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., et al. (2016). The simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* 538, 201–206. doi: 10.1038/nature18964

Maltese, P. E., Orlova, N., Krasikova, E., Emelyanchik, E., Cheremisina, A., Kuscaeva, A., et al. (2017). Gene-targeted analysis of clinically diagnosed long QT Russian families. *Int. Heart J.* 58, 81–87. doi: 10.1536/ihj.16-133

Marakhonov, A. V., Brodehl, A., Myasnikov, R. P., Sparber, P. A., Kiseleva, A. V., Kulikova, O. V., et al. (2019). Noncompaction cardiomyopathy is caused by a novel in-frame desmin (DES) deletion mutation within the 1A coiled-coil rod segment leading to a severe filament assembly defect. *Hum. Mutat.* 40, 734–741. doi: 10.1002/humu.23747

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

Meshkov, A., Ershova, A., Kiseleva, A., Zotova, E., Sotnikova, E., Petukhova, A., et al. (2021). The LDLR, APOB, and PCSK9 variants of index patients with familial hypercholesterolemia in Russia. *Genes* 12:66. doi: 10.3390/genes12010066

Milovidova, T. B., Bulach, M. V., Schagina, O. A., and Polyakov, A. V. (2018). Molecular genetic analysis of congenital merozin-negative muscular dystrophy in Russia. *Med. Genet.* 17, 38–45.

Miroshnikova, V., Romanova, O., Ivanova, O., Fedyakov, M., Panteleeva, A., Barbitoff, Y., et al. (2021). Identification of novel variants in the LDLR gene in Russian patients with familial hypercholesterolemia using targeted sequencing. *Biomed. Rep.* 14:15. doi: 10.3892/br.2020.1391

Ng, P. C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* 7, 61–80. doi: 10.1146/annurev.genom.7.080505.115630

Nikitin, S. S. (2016). Resolution of the 3rd Russian panel of experts in diagnostics and treatment of Pompe disease. *Neuromuscul. Dis.* 6, 89–90.

Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y.-Y., et al. (2017). Sherloc: a comprehensive refinement of the ACMG–AMP variant classification criteria. *Genet. Med.* 19, 1105–1117. doi: 10.1038/gim.2017.37

Petrova, N. V., Kashirskaya, N. Y., Vasilyeva, T. A., Kondratyeva, E. I., Zhekaite, E. K., Voronkova, A. Y., et al. (2020). Analysis of CFTR mutation spectrum in ethnic Russian cystic fibrosis patients. *Genes* 11:554. doi: 10.3390/genes11050554

Piekutowska-Abramczuk, D., Popowska, E., Pronicki, M., Karczmarewicz, E., Tylek-Lemanska, D., Sykut-Cegielska, J., et al. (2009). High prevalence of SURF1 c.845_846delCT mutation in Polish Leigh patients. *Eur. J. Paediatr. Neurol.* 13, 146–153. doi: 10.1016/j.ejpn.2008.03.009

Pokrovskaya, M. S., Sivakova, O. V., Efimova, I. A., Meshkov, A. N., Metelskaya, V. A., Shalnova, S. A., et al. (2019). Biobanking as a necessary tool for research in the field of personalized medicine in the scientific medical center. *Per. Med.* 16, 501–509. doi: 10.2217/pme-2019-0049

Polyak, M. E., Ivanova, E. A., Polyakov, A. V., and Zaklyazminskaya, E. V. (2016). Mutation spectrum of the gene KCNQ1 in russian patients with long QT syndrome. *Russ. J. Cardiol.* 10, 15–20. doi: 10.15829/1560-4071-2016-10-15-20

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* 17, 405–423. doi: 10.1038/gim.2015.30

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Semyachkina, A. N., Sukhorukov, V. S., Bukina, T. M., Yablonskaya, M. I., Merkuryeva, E. S., Kharabadze, M. N., et al. (2014). Glycogen storage disease type II (Pompe disease) in children. *Ross. Vestn. Perinatol. Pediatr. Russian Bull. Perinatol. Pediatr.* 59, 48–55.

Shah, N., Hou, Y.-C. C., Yu, H.-C., Sainger, R., Caskey, C. T., Venter, J. C., et al. (2018). Identification of misclassified clinvar variants via disease population prevalence. *Am. J. Hum. Genet.* 102, 609–619. doi: 10.1016/j.ajhg.2018.02.019

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shestak, A. G., Bukaeva, A. A., Saber, S., and Zaklyazminskaya, E. V. (2021). Allelic dropout is a common phenomenon that reduces the diagnostic yield of PCR-based sequencing of targeted gene panels. *Front. Genet.* 12:337. doi: 10.3389/fgene.2021.620337

Solodskikh, S. A., Panevina, A. V., Gryaznova, M. V., Gureev, A. P., Serzhantova, O. V., Mikhailov, A. A., et al. (2019). Targeted sequencing to discover germline variants in the BRCA1 and BRCA2 genes in a Russian population and their association with breast cancer risk. *Mutat. Res. Mol. Mech. Mutagen.* 813, 51–57. doi: 10.1016/j.mrfmmm.2018.12.005

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69. doi: 10.1126/science.1219240

Tsygankova, P. G., Mikhaĭlova, S. V., Zakharova, E. I., Pichkur, N. A., Il'ina, E. S., Nikolaeva, E. A., et al. (2010). Syndrome Leigh caused by mutations in the SURF1 gene: clinical and molecular-genetic characteristics. *Z. Nevrol. Psikhiatrii* 110, 25–32.

Van der Ploeg, A. T., and Reuser, A. J. (2008). Pompe's disease. *Lancet* 372, 1342–1353. doi: 10.1016/S0140-6736(08)61555-X

Van Hout, C. V., Tachmazidou, I., Backman, J. D., Hoffman, J. D., Liu, D., Pandey, A. K., et al. (2020). Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* 586, 749–756. doi: 10.1038/s41586-020-2853-0

Van Rooij, J., Arp, P., Broer, L., Verlouw, J., van Rooij, F., Kraaij, R., et al. (2020). Reduced penetrance of pathogenic ACMG variants in a deeply phenotyped cohort study and evaluation of ClinVar classification over time. *Genet. Med.* 22, 1812–1820. doi: 10.1038/s41436-020-0900-8

Wong, E. H. M., Khrunin, A., Nichols, L., Pushkarev, D., Khokhrin, D., Verbenko, D., et al. (2017). Reconstructing genetic history of Siberian and Northeastern European populations. *Genome Res.* 27, 1–14. doi: 10.1101/gr.202945.115

Wright, C. F., West, B., Tuke, M., Jones, S. E., Patel, K., Laver, T. W., et al. (2019). Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am. J. Hum. Genet.* 104, 275–286. doi: 10.1016/j.ajhg.2018.12.015

Zaklyazminskaya, E., Mikhailov, V., Bukaeva, A., Kotlukova, N., Povolotskaya, I., Kaimonov, V., et al. (2019). Low mutation rate in the TTN gene in paediatric patients with dilated cardiomyopathy – a pilot study. *Sci. Rep.* 9:16409. doi: 10.1038/s41598-019-52911-1

Zhernakova, D. V., Brukhin, V., Malov, S., Oleksyk, T. K., Koepfli, K. P., Zhuk, A., et al. (2020). Genome-wide sequence analyses of ethnic populations across Russia. *Genomics* 112, 442–458. doi: 10.1016/j.ygeno.2019.03.007

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.

ORIGINAL RESEARCH
published: 15 October 2021
doi: 10.3389/fgene.2021.716558

Local Ancestry Adjusted Allelic Association Analysis Robustly Captures Tuberculosis Susceptibility Loci

Yolandi Swart[1], *Caitlin Uren*[1,2], *Paul D. van Helden*[1], *Eileen G. Hoal*[1] *and Marlo Möller*[1,2]*

[1] *DSI-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa,* [2] *Centre for Bioinformatics and Computational Biology, Stellenbosch University, Stellenbosch, South Africa*

Pulmonary tuberculosis (TB), caused by *Mycobacterium tuberculosis*, is a complex disease. The risk of developing active TB is in part determined by host genetic factors. Most genetic studies investigating TB susceptibility fail to replicate association signals particularly across diverse populations. South African populations arose because of multi-wave genetic admixture from the indigenous KhoeSan, Bantu-speaking Africans, Europeans, Southeast Asian-and East Asian populations. This has led to complex genetic admixture with heterogenous patterns of linkage disequilibrium and associated traits. As a result, precise estimation of both global and local ancestry is required to prevent both false positive and false-negative associations. Here, 820 individuals from South Africa were genotyped on the SNP-dense Illumina Multi-Ethnic Genotyping Array (~1.7M SNPs) followed by local and global ancestry inference using RFMix. Local ancestry adjusted allelic association (LAAA) models were utilized owing to the extensive genetic heterogeneity present in this population. Hence, an interaction term, comprising the identification of the minor allele that corresponds to the ancestry present at the specific locus under investigation, was included as a covariate. One SNP (rs28647531) located on chromosome 4q22 was significantly associated with TB susceptibility and displayed a SNP minor allelic effect (G allele, frequency = 0.204) whilst correcting for local ancestry for Bantu-speaking African ancestry (*p*-value = 5.518×10^{-7}; OR = 3.065; SE = 0.224). Although no other variants passed the significant threshold, clear differences were observed between the lead variants identified for each ancestry. Furthermore, the LAAA model robustly captured the source of association signals in multi-way admixed individuals from South Africa and allowed the identification of ancestry-specific disease risk alleles associated with TB susceptibility that have previously been missed.

Keywords: South Africa, admixture mapping, TB susceptibility, ancestry-specific risk alleles, local ancestry adjustments, population genetics, host genetics

OPEN ACCESS

Edited by:
Ranajit Das,
Yenepoya University, India

Reviewed by:
Tesfaye B. Mersha,
Cincinnati Children's Hospital Medical Center, United States
Mohammed S. Mustak,
Mangalore University, India

****Correspondence:***
Marlo Möller
marlom@sun.ac.za

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *28 May 2021*
Accepted: *01 October 2021*
Published: *15 October 2021*

Citation:
Swart Y, Uren C, van Helden PD, Hoal EG and Möller M (2021) Local Ancestry Adjusted Allelic Association Analysis Robustly Captures Tuberculosis Susceptibility Loci. Front. Genet. 12:716558. doi: 10.3389/fgene.2021.716558

INTRODUCTION

Pulmonary tuberculosis (TB), caused by the *bacillus Mycobacterium tuberculosis (M.tb)*, is a complex disease which affects populations disproportionately and results from a multifactorial interaction between host and pathogen (Yim and Selvaraj, 2010). It is often said that approximately 5–10% of infected individuals (±3 billion people worldwide) will go on to develop active TB whilst the majority will remain asymptomatic (Bañuls et al., 2015; El Kamel et al., 2015; Chaw et al., 2020). According to the World Health Organization (WHO), an estimated 10 million TB cases and 1.5 million deaths were reported in 2019 (WHO, 2019). TB therefore remains a global health burden and is of particular concern in low- to middle-income countries where a generally higher incidence rate (615 per 100 000 in South Africa) occurs, together with the limitations of currently available therapies and vaccines (Bao et al., 2016; WHO | Global tuberculosis report 2019, 2019). Numerous genetic and heritability studies have established the role of host genetic factors in susceptibility to TB (Rudko et al., 2016; Kinnear et al., 2017; Cai et al., 2019; Luo et al., 2019), but with minimal overlap between populations from various geographical regions (Thye et al., 2010; Oki et al., 2011; Mahasirimongkol et al., 2012; Png et al., 2012; Thye et al., 2012; Chimusa et al., 2014, 2014; Curtis et al., 2015; Schurz et al., 2015; Grant et al., 2016; Sobota et al., 2016; Uren et al., 2017a; Omae et al., 2017; Qi et al., 2017; Zheng et al., 2018). The variation observed between populations from diverse geographic regions indicates possible ancestry-specific differences that contribute to the host genetic variability observed in TB genome-wide association studies (GWAS) (van Helden et al., 2006; Chimusa et al., 2014; Schurz et al., 2019a; Cai et al., 2019).

Previous investigations into southern African history and population structure elucidated indigenous KhoeSan ancestry in the region, in addition to populations being multi-way admixed due to multiple inter-and intra-continental migrations (de Wit et al., 2010; Quintana-Murci et al., 2010; Uren et al., 2017b). This population history has resulted in admixture from indigenous KhoeSan, Bantu-speaking African, European, Southeast Asian and East Asian populations (de Wit et al., 2010; Quintana-Murci et al., 2010; Uren et al., 2016). Ancestral populations contributed linked alleles (haplotype blocks) resulting in a mosaic of phenotypic consequences. This admixture can be leveraged to identify associations between various TB phenotypes and genomic regions harbouring variants with highly differentiated allele frequencies among ancestral populations, known as admixture mapping (Wang et al., 2020). Hence, the unique and complex admixed individuals from southern Africa, harbouring genomic contributions from ancestral populations with differing historical disease burden, present an opportunity to investigate ancestry-specific disease risk alleles associated with TB susceptibility (Shriner, 2013; Wang et al., 2020).

Previous admixture mapping and association studies investigating TB susceptibility loci in South Africa were restricted by a low number of controls, small reference population sample size and low SNP density (de Wit et al., 2010; Chimusa et al., 2014; Daya et al., 2014b, 2014a). With the recent adaption of computational algorithms to better suit multi-way admixed populations, a more suitable, high-density genotyping platform and the availability of large scale, population-specific datasets, we aimed to perform an updated scan for variants associated with TB using local ancestry adjusted allelic (LAAA) association models.

MATERIALS AND METHODS

Study Population and Ethics Approval

A total of 413 pulmonary TB cases and 407 healthy controls were recruited from the metropolitan area of Cape Town in the Western Cape Province, South Africa. The population from this area was elected due to the high incidence of TB as well as the equal socio-economic status and low prevalence of HIV at the time of sampling (Rossouw et al., 2003; Möller et al., 2009; Gallant et al., 2010). Furthermore, TB cases and controls were sampled from the same area, therefore socio-economic status is unlikely to be a confounding factor as previously determined by Chimusa et al. (2014). TB cases were distinguished through bacteriological confirmation (culture positive and/or smear positive). Healthy controls had no previous history of TB. However, 80% of individuals above 15 years of age in this area were estimated to have been exposed to *M.tb*, and could therefore be regarded as latently infected (Gallant et al., 2010). If study participants were under the age of 18 or were HIV-positive, they were excluded from the analysis.

Written informed consent was obtained from all study participants before recruitment and blood collection. Sample collection (protocol number 95/072) and this study (S20/02/041) were both approved by the Health Research Ethics Committee of the Faculty of Health Sciences (HREC), Stellenbosch University. The research was conducted according to the principles expressed in the Declaration of Helsinki (2013).

Genotyping, Data Merging and Quality Control

Genotype data on the case-control cohort was generated using the Illumina (Illumina, CA, United States) multi-ethnic genotyping array (MEGA) comprising ~1.7 million markers (Schurz et al., 2019b). The Sanger Imputation Server (SIS) (https://imputation.sanger.ac.uk) and the African Genome Resource (AGR) reference panel (Gurdasani et al., 2015) was utilised for the imputation of missing genotypes. The imputed data was subjected to iterative quality control as previously described by Schurz et al. (2019b). Thereafter, the data from the admixed individuals were merged with the respective appropriate source populations (summarised in **Table 1**) using PLINK v2.0 (https://www.cog-genomics.org/plink/2.0/) (Purcell et al., 2007) in order to generate input files required for global and local ancestry inference.

After merging of admixed and source ancestral populations, all individuals missing more than 10% genotypes were removed, SNPs with more than 3% missing data were excluded and a Hardy-Weinberg equilibrium (HWE) filter was used in controls

TABLE 1 | Ancestral populations included in analysis.

Population	*n*	Source
European (British)	60	1000G phase 3
African (Luhya)	50	1000G phase 3
East Asian (Chinese)	36	1000G phase 3
KhoeSan (Nama)	44	European Genome-Phenome archive- https://ega-archive.org/
South East Asian (Malay)	40	Wong et al. (2013)

(threshold < 0.01). The data was screened for relatedness using the software KING (Manichaikul et al., 2010) and individuals up to second degree relatedness were subsequently removed. Variants with a minor allele frequency (MAF) below 1% were removed. The final dataset after quality control and data filtering consisted of 392 TB cases and 346 controls in addition to 289 ancestral individuals. A total of 4,249,442 variants passed quality control and filtering parameters.

Global Ancestry Inference

ADMIXTURE was used to investigate the population substructure amongst our cohort, as well as to determine the correct number of contributing ancestries (Alexander and Lange, 2011; Zhou et al., 2011). This is a model-based approach to estimate individual ancestry coefficients of an individual's genome from *k* ancestral populations and corresponding ancestral genotype frequencies through cross validation. For the purpose of computational efficiency, redundant single-nucleotide polymorphisms (SNPs) were removed and only tagging SNPs representative of the genetic haplotype blocks remained. Therefore, each SNP that has a linkage disequilibrium (LD) r^2 of >0.1 within a 50-SNP sliding window (advanced by 10 SNPs at a time) was removed. A total of 261,694 autosomal markers after LD pruning and 820 individuals (413 cases and 407 controls) were used to infer ancestry in an unsupervised manner for k = 3–10 (5 iterations). All 820 individuals were grouped into running groups of equal size together with 289 ancestral populations whilst inferring global ancestry proportions. Related individuals were included in separate running groups. Running groups were created to ensure an equal number of reference populations and admixed populations whilst removing relatedness as a confounding factor during global ancestry assignment. After determining the correct *k* number of contributing ancestries through cross validation, the software RFMix was used to infer global ancestry proportions for downstream statistical analysis, since ADMIXTURE is not as accurate as haplotype-based analyses (Uren et al., 2020). The software PONG was used for visualisation of global ancestry proportions and amalgamation of multiple iterations into the major mode (Behr et al., 2016).

Local Ancestry Inference

Local ancestry inference requires phasing of haplotypes prior to inferring local ancestry. The software program SHAPEIT2 (Delaneau et al., 2013; Delaneau and Marchini, 2014) (utilizing the HapMap Genetic map – GRCh37) was used to phase the merged dataset before inferring local ancestry for each position in the genome using RFMix (Maples et al., 2013). RFMix is 30X faster than other local ancestry inference software and is accurate in multi-way admixture scenarios (Maples et al., 2013; Uren et al., 2020). Default parameters were used, except for the number of generations since admixture, which was set to 15, consistent with previous studies (Uren et al., 2016). Both global and local ancestry was inferred for 1,027 individuals (392 TB cases, 346 controls and 289 ancestral individuals) and 4,249,442 autosomal SNPs.

Statistical Analysis

A Local Ancestry Adjusted Allelic (LAAA) model, first described by Duan et al. (2018), was used to investigate if there are allelic, ancestry-specific or ancestry-specific allelic associations with TB susceptibility in an admixed South African population (Duan et al., 2018). Dosage files were compiled at each locus as a biallelic state and were calculated as 0, 1 or 2 copies of a specific ancestry at any locus along the genome. Separate regression models for each ancestral group were fitted to investigate which ancestral population(s) drive the association between TB status and local ancestry at each locus. Genome-wide admixture proportions obtained from RFMix were included in all regression models to account for population structure. The smallest ancestry proportion (East Asian) was excluded as covariate to avoid complete separation of data. Therefore, four ancestral components (KhoeSan, African, European, and Southeast Asian) were included as covariates in association testing, together with age and gender. The number of alternate alleles (not the reference alleles) were counted, as these are more likely to be ancestry-specific. A total of 738 unrelated individuals (392 TB cases and 346 controls) and 4,249,442 autosomal markers were included in this analysis. The *glm()* function in R was used for logistic regression association testing.

The following four regression models were tested simultaneously to detect the source (allelic, ancestry or both ancestry-allelic effect) of the association signals observed:

1. Global ancestry proportions were included as covariates and thus represents the null model. This test is regularly used in GWAS to investigate whether an additive allelic dose affect exists on the phenotype, not considering local ancestry (Homozygous for the reference allele = 0; Heterozygous = 1; Homozygous for the alternate allele = 2).
2. Local ancestry expressed in terms of the number of copies of a specific ancestry (Ancestry of interest = 1; Other ancestries = 0) at a locus were included as covariates. This model is often

utilised to conduct admixture mapping studies to elucidate ancestry effects of variants which showcases frequency disparities across ancestral populations (Homozygous for other ancestry = 0; Heterozygous = 1; Homozygous for ancestry of interest = 2).

3. Minor allelic effects were used in an additive manner and were included as covariates whilst still adjusting for local ancestry. Therefore, jointly testing for model 1 + 2.
4. This model utilises the ancestry-specific minor alleles at a locus, thus the minor alleles together with the corresponding ancestry of the minor allele were included as covariates (Minor allele and ancestry not on the same haplotypes = 0; Minor allele and ancestry are on the same haplotype = 1). This model is an extension to the allelic (3) and local ancestry (2) model by modelling the combination of the minor allele present at a specific locus and the ancestry of the specific allele at that genomic locus. (Both minor allele and ancestry not on the same haplotype = 0; Heterozygote (only one haplotype has both minor allele and ancestry on the same haplotype = 1; Both minor allele and ancestry on the same haplotype = 2).

Since the true underlying causal variants as well as the LD between the marker under study are unknown, modelling all three terms simultaneously is the most effective approach to elucidate causal variants in an admixed cohort with minimal power loss (Duan et al., 2018). Therefore, we can determine if a specific minor allele, ancestry or both a minor allelic and ancestry co-occurs with TB status more often than would be expected by chance.

The development of power and sample size analysis tools for mapping ancestry-specific effects are lacking. The power to detect significant associations depends greatly on the proportion of admixture, differences in effect sizes between diverse ancestries and differences in the allele risk frequencies among ancestral populations. It is noteworthy to highlight that this information will vary for each admixture scenario. Nonetheless, it remains critical to conduct some sort of power calculation to ensure the reliability of elucidating ancestry-specific genomic regions amongst admixed individuals. Hence, we conducted a priori power analysis in order to ensure the reliability of results given our samples size using G*Power (Faul et al., 2007, 2009).

To account for the multiple testing burden, the R package *STEAM* (Significance Threshold Estimation for Admixture Mapping) (Grinde et al., 2019) was used to estimate the genome-wide significance threshold. *STEAM* is specifically designed to estimate genome-wide significance thresholds for admixture mapping studies given the admixture proportions and number of generations since admixture. We quantified the degree of inflation by generating a Quantile-Quantile plot of the residuals.

RESULTS

Global Ancestry Inference

After close inspection of global ancestry proportions generated using ADMIXTURE, the *k* number of contributing ancestries was determined to be k = 5, since this was the lowest k-value through cross validation (**Supplementary Table S1**). Since haplotype-based admixture software is more accurate at global ancestry inference, ancestry proportions (genome-wide ancestral contributions) were inferred for all individuals using RFMix (Uren et al., 2020). **Figure 1** represents the global ancestry proportions plotted vertically for each admixed individual and contributing ancestral populations using RFMix (k = 5). It is evident from the global ancestry inference that the cohort is a complex five-way admixed group, with ancestral contributions from the indigenous KhoeSan (~35–40%), Bantu-speaking Africans (~27–30%), Europeans (~20%), Southeast Asians (~7–8%) and East Asians (~5%). Furthermore, extensive genetic heterogeneity can be observed, since genome-wide proportions differ vastly between individuals.

Local Ancestry Inference

Local ancestry was estimated for all individuals and visually observed with karyograms. As shown in **Figure 2**, admixture between geographically distinct populations creates complicated ancestral-and admixture induced LD blocks. **Figure 2** represents a single five-way admixed individual. Since not all individuals will harbour the same number and length of ancestry segments, it is necessary to accurately infer local ancestry in every individual at each genomic locus.

Local Ancestry Allelic Adjusted Association Analysis

A total of 4,249,442 autosomal markers and 738 unrelated individuals (392 TB cases and 346 controls) were included in logistic regression models to assess whether any loci were significantly associated with TB status (adjusting for gender, age, and global ancestry proportions inferred by RFMix). More information regarding the distribution of age, gender and ancestry proportions of the cohort can be found in the **Supplementary Figures S1–S3** and **Supplementary Table S2**. LAAA models were successfully conducted for all five ancestries present in this highly complex admixed cohort.

One variant (rs28647531) was significantly associated with TB status (p-value $< 1.078 \times 10^{-6}$) due to an allelic SNP effect (G allele; 0.204 frequency) whilst adjusting for Bantu-speaking African local ancestry on chromosome 4 (OR = 3.065, p-value $= 5.518 \times 10^{-7}$) (**Figure 3**). This variant is an intronic variant with a gene consequence on Follistatin-related protein (*FSTL5*), which is a protein coding gene involved in calcium ion binding. No restrictions on the analysis or inflation of results were observed as indicated by the Quantile-Quantile plot (**Supplementary Figure S4**). Although no other variants passed the significance threshold, multiple lead variants (p-value $< 1 \times 10^{-5}$) were identified. Furthermore, it is clear from our results that multiple distinct lead variants were identified for each ancestry.

The lead variants identified using only the global ancestry as covariates (model 1), are summarised in **Supplementary Table S3**. One lead variant (rs38672118) is near the protein

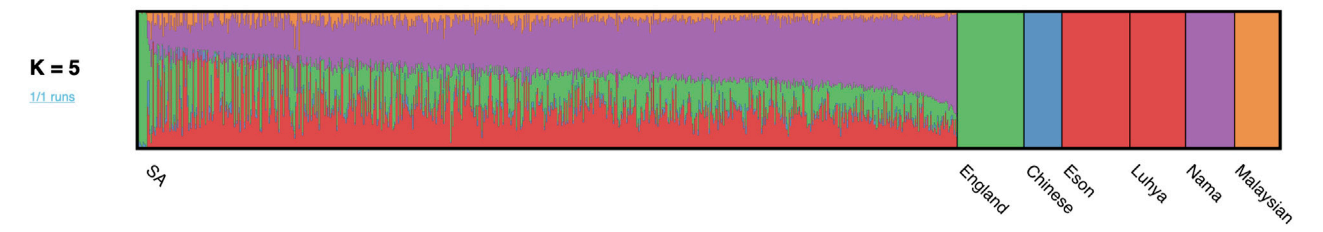


FIGURE 1 | Genome-wide ancestral proportions of all SA individuals, with the ancestry proportion of each individual plotted vertically.

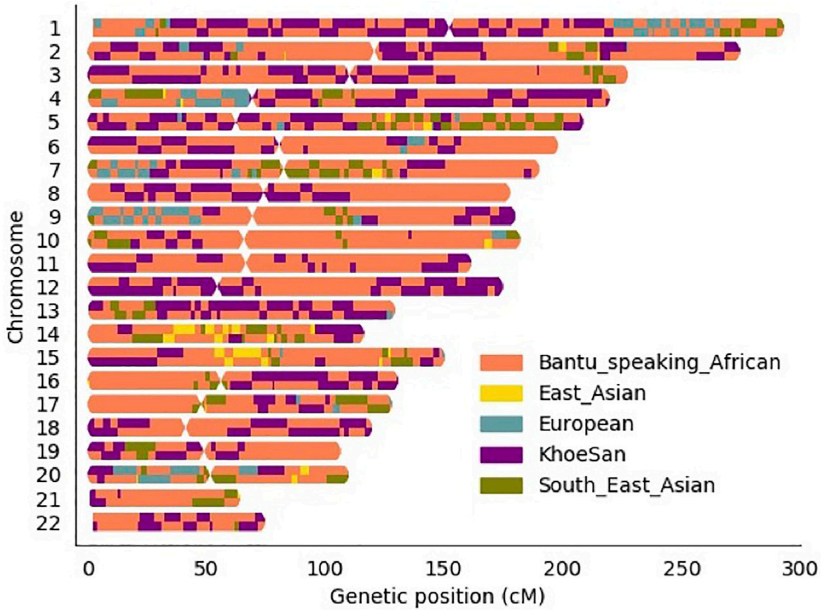


FIGURE 2 | Karyogram of one admixed SA individual.

coding gene, *CUL2* (Cullin-2), located on chromosome 10. The lead variants identified by conducting admixture mapping (model 2), are summarised in **Supplementary Table S4**. Only one ancestry (European) identified a local ancestry peak on chromosome 15 (**Supplementary Figure S5**). The lead variants identified utilising the allelic model adjusting for local ancestry (model 3), are summarised in **Table 2**. The lead variants identified by the LAAA model (model 4) are summarised in **Table 3**. It is noteworthy that both the allelic model adjusting for local ancestry (model 3), and the LAAA model (model 4) captured association signals not previously observed for this cohort.

DISCUSSION

We conducted local ancestry allelic adjusted association analysis in a multi-way admixed South African (SA) population to investigate whether ancestry-specific genetic regions are associated with TB susceptibility. Multi-way admixed populations allow the opportunity to simultaneously assess the association of TB status in multiple continental populations and elucidate possible ancestry-specific effects on TB susceptibility. Previous studies were confounded by the limited number of representative reference populations available to infer local ancestry and the use of the low-density Affymetrix gene chip array (~500k markers) in the analyses. New, more representative ancestral populations and an increase in accuracy of several software tools facilitated the novel findings presented here.

Global ancestry deconvolution suggested a five-way admixed scenario for the study cohort. This is in accordance with previous studies (de Wit et al., 2010; Chimusa et al., 2014; Uren et al., 2016). This diverse admixture and associated regional heterogeneity are reflected in the karyograms generated via

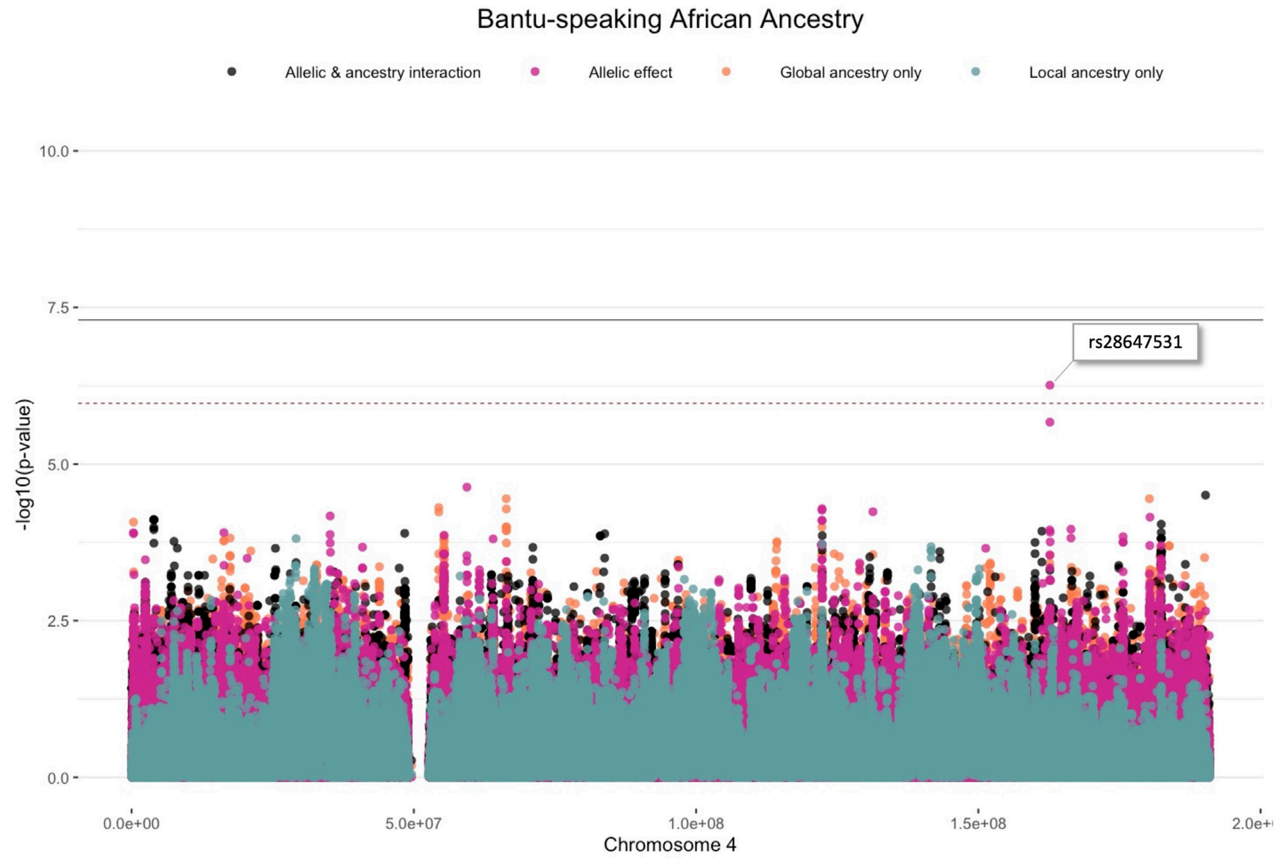


FIGURE 3 | Log transformation of association signals (p-value < 1.078 × 10^{-6}) obtained for Bantu-speaking African ancestry whilst using the allelic model whilst adjusting for local ancestry on chromosome 4. The dashed red line represents the significant threshold for admixture mapping calculated with the software *STEAM* and the black solid line represents the genome-wide significant threshold of 5 × 10^{-8}. The four different models are represented in orange (global ancestry only), blue (local ancestry effect), pink (minor allelic effect only) and black (both minor allelic and ancestry effects).

local ancestry inference (**Figure 2**). This scale of genetic heterogeneity suggests that no two individuals will harbour the same DNA segment from the same ancestral population, i.e., there is a high degree of locus-specific ancestry (Duan et al., 2018). The results presented here highlight that only including global ancestry proportions in the analysis is not sufficient to identify which ancestry is located on distinct chromosomal segments. The only lead variant (rs38672118) identified using the global ancestry-only model is near the protein coding gene, *CUL2*. Although the function of *CUL2* on *M.tb* clearance is still uncertain, *CUL2* forms an important part of the cullin-RING-based E3 ubiquitin-protein ligase complex and subsequently targets the ubiquitination of target proteins (Nguyen et al., 2017). The model used for admixture mapping (only utilising local ancestry) seems overconservative for complex multi-way admixed individuals, since only one admixture peak was close to the significance threshold for European ancestry (located on chromosome 15). This highlights the phenomenon of genetic heterogeneity where the presence of both admixture-induced LD blocks and haplotype LD blocks often results in missed association signals due to tagging SNPs being possibly located in different ancestral LD blocks (Duan et al., 2018).

One example of missing relevant associated variants in complex admixed populations, is the association signal obtained on chromosome 11q13 while adjusting for Bantu-speaking African- and European local ancestry. This lead variant indicated an association with the TIR Domain Containing Adaptor Protein (*TIRAP*) gene (**Figure 4**) and is involved in the toll-like receptor (TLR) 4 signalling pathway of the immune system via the TIR adaptor protein it codes for. *TIRAP* is a protein which identifies microbial pathogens trough TLRs as part of the initial innate immune response (Selvaraj et al., 2010). This acts via *IRAK2* and *TRAF-6*, leading to the activation of NF-kappa-B, MAPK1, MAPK3 and JNK, which is essential for cytokine secretion in order to mount an inflammatory response (Capparelli et al., 2013). Polymorphisms in the *TIRAP* gene were previously identified to be associated with TB susceptibility in a South Indian population (Selvaraj et al., 2010), as well as a Chinese population (Zhang et al., 2011). This suggests a possible role of the *TIRAP* gene in TB susceptibility via activation of TLRs in order to recognize several components of *M.tb* during active TB disease. The T allele of TLR4 (rs4986791) was found to be associated with an increased risk for an Asian subgroup in a meta-analysis

TABLE 2 | Summary statistics of the top results (*p*-value < 1 × 10^{-5}) whilst utilising the Additive allelic model whilst adjusting for local ancestry.

Chr	Position	rsID	Ref	Alt	Altfreq	OR	SE	*p*-value	Ancestry	Location	Gene
4	153960368	rs1024148	T	G	0.463	1.589	0.126	9.948e-06	European	None	None
13	24752695	rs7325698	T	C	0.310	1.363	0.150	7.721e-06	European	Intronic	*SPATA13*
13	24753449	rs2862243	A	G	0.263	1.300	0.174	3.811e-06	European	Intronic	*SPATA13*
13	50909771	rs67217502	T	C	0.103	1.108	0.225	6.474e-06	European	Intronic	*DLEU1*
13	50913874	rs12853498	A	T	0.103	1.108	0.225	6.474e-06	European	Intronic	*DLEU1*
13	50915280	rs17074141	C	T	0.103	1.108	0.225	6.474e-06	European	Intronic	*DLEU1*
13	50920890	rs17363026	T	C	0.104	1.110	0.225	4.129e-06	European	Intronic	*DLEU1*
13	50922773	rs17074143	T	A	0.099	1.104	0.230	5.059e-06	European	Intronic	*DLEU1*
13	50925317	rs34712361	T	C	0.098	1.103	0.230	7.398e-06	European	Intronic	*DLEU1*
13	50925565	rs67964536	C	T	0.098	1.103	0.230	7.398e-06	European	Intronic	*DLEU1*
13	50926076	rs79714483	A	G	0.098	1.103	0.230	7.398e-06	European	Intronic	*DLEU1*
14	48325261	rs447600	T	A	0.459	1.582	0.123	4.350e-06	European	None	None
2	52241352	rs2883609	C	G	0.318	1.374	0.122	9.361e-06	East Asian	nRNA_intronic	*AC007682.1*
2	180940603	rs13411512	T	C	0.274	1.315	0.130	7.428e-06	East Asian	Intronic	*CWC22*
12	9388842	ss1388098326	C	T	0.124	1.132	0.176	9.961e-06	East Asian	nRNA_intronic	*A2MP1*
14	48325261	rs447600	T	A	0.459	1.582	0.107	5.988e-06	East Asian	Intergenic	*RP11-476J6.1*
22	46046477	rs134850	A	G	0.223	1.250	0.131	7.042e-06	East Asian	Intergenic	*ATXN10*
1	151185502	rs4971014	A	G	0.187	1.206	0.150	3.279e-06	SouthEast Asian	Intronic	*PIP5K1A*
2	52241352	rs2883609	C	G	0.318	1.374	0.126	9.022e-06	SouthEast Asian	ncRNA_intronic	*AC007682.1*
2	180940603	rs13411512	T	C	0.274	1.315	0.132	5.645e-06	SouthEast Asian	Intergenic	*CWC22*
8	126754436	rs12547413	T	C	0.135	1.144	0.170	8.440e-06	SouthEast Asian	Intronic	*CLU2*
10	35527543	rs3867218	C	T	0.513	1.670	0.116	3.194e-06	SouthEast Asian	Intergenic	*RNU6794P*
14	90278083	rs10137384	T	C	0.121	1.129	0.192	8.522e-06	SouthEast Asian	ncRNA_intronic	*RP1133N16.3*
21	43759441	rs692544	C	T	0.508	1.662	0.115	4.414e-06	SouthEast Asian	Intergenic	*TFF2*
22	46036079	rs1894617	G	C	0.235	1.265	0.129	7.610e-06	SouthEast Asian	Intergenic	*RNU6794P*
22	46046477	rs134850	A	G	0.223	1.250	0.131	6.868e-06	SouthEast Asian	Intergenic	*ATXN10*
2	143729878	rs10928161	C	T	0.129	1.137	0.204	7.369e-06	KhoeSan	Intronic	*KYNU*
2	143730019	rs16855223	G	A	0.130	1.139	0.204	6.268e-06	KhoeSan	Intronic	*KYNU*
2	143731496	rs35991933	A	T	0.129	1.137	0.204	7.369e-06	KhoeSan	Intronic	*KYNU*
2	143731661	rs34891373	T	A	0.129	1.137	0.204	7.369e-06	KhoeSan	Intronic	*KYNU*
2	143737201	rs11904225	G	A	0.146	1.157	0.194	1.374e-06	KhoeSan	Intronic	*KYNU*
2	143742532	rs10496933	G	A	0.129	1.138	0.204	5.676e-06	KhoeSan	Intronic	*KYNU*
2	143743246	rs12463750	G	A	0.146	1.157	0.194	1.374e-06	KhoeSan	Intronic	*KYNU*
2	180940603	rs13411512	T	C	0.274	1.315	0.164	8.617e-06	KhoeSan	Intronic	*KYNU*
4	54413304	rs4864469	T	C	0.067	1.070	0.247	8.725e-06	KhoeSan	ncRNA_intronic	*FIP1L1*
4	114309839	rs6533681	C	T	0.356	1.427	0.157	4.856e-06	KhoeSan	Intergenic	*ANK2*
5	81172726	rs62368165	G	A	0.401	1.494	0.180	9.898e-06	KhoeSan	Intergenic	*SHFM1P1*
6	7328023	rs145663084	T	C	0.113	1.120	0.211	4.122e-06	KhoeSan	ncRNA_exonic	*PRSS23*
11	86632570	rs612410	T	C	0.337	1.400	0.158	8.791e-06	KhoeSan	Intronic	*PRSS23*
11	86641079	rs10792884	A	G	0.361	1.435	0.160	9.751e-06	KhoeSan	Intronic	*PRSS23*
11	86641484	rs7940935	C	T	0.363	1.437	0.160	9.775e-06	KhoeSan	Intronic	*PRSS23*
11	86642522	rs10792886	G	A	0.363	1.437	0.160	8.430e-06	KhoeSan	Intronic	*PRSS23*
11	86643022	rs7948323	C	A	0.364	1.439	0.160	8.311e-06	KhoeSan	Intronic	*PRSS23*
11	86643351	rs10751145	A	G	0.364	1.439	0.160	8.311e-06	KhoeSan	Intronic	*PRSS23*
11	86644159	rs10792887	G	A	0.364	1.439	0.160	8.311e-06	KhoeSan	Intronic	*PRSS23*
11	86644244	rs10792888	A	C	0.364	1.439	0.160	8.311e-06	KhoeSan	Intronic	*PRSS23*
11	86644300	rs7484279	C	T	0.364	1.439	0.160	8.311e-06	KhoeSan	Intronic	*PRSS23*
11	86644938	rs3740665	C	T	0.365	1.440	0.160	8.889e-06	KhoeSan	Intronic	*PRSS23*
11	86645157	rs10898560	A	G	0.366	1.442	0.160	9.506e-06	KhoeSan	Intronic	*PRSS23*
11	86645214	rs1902425	C	T	0.365	1.440	0.160	8.889e-06	KhoeSan	Intronic	*PRSS23*
11	124196782	rs676720	C	A	0.239	1.271	0.162	8.973e-06	KhoeSan	Intergenic	*OR8B7P*
11	124199570	rs7119360	A	G	0.239	1.271	0.162	8.973e-06	KhoeSan	Intergenic	*OR8B7P*
12	18844727	rs10841067	T	C	0.050	1.052	0.292	4.802e-06	KhoeSan	Intronic variant	*PLCZ1*
12	18845754	rs1973289	C	T	0.054	1.056	0.282	1.710e-06	KhoeSan	ncRNA_exonic	*PLCZ1*
12	18846108	rs2900416	A	G	0.054	1.056	0.282	1.710e-06	KhoeSan	ncRNA_exonic	*PLCZ1*
15	39513293	rs7176317	C	G	0.516	1.675	0.127	2.405e-06	KhoeSan	ncRNA_exonic	*RP11-624L4.1*
4	162663106	rs10517752	G	A	0.199	1.221	0.222	2.142e-06	African	Intronic	*FSTL5*
4	162663775	rs28647531	A	G	0.204	1.226	0.224	5.518e-07	African	Intronic	*FSTL5*
17	75030582	rs11077888	G	A	0.470	1.600	0.142	3.400e-06	African	Intergenic	*AC015815.5*
18	41342728	rs11659620	T	C	0.081	1.084	0.251	5.876e-06	African	Intergenic	*RNU6443P*
18	41351686	rs1822027	T	G	0.081	1.084	0.251	5.876e-06	African	Intergenic	*RNU6443P*
18	41352221	rs35810759	A	G	0.081	1.084	0.251	5.876e-06	African	Intergenic	*RNU6443P*
21	43759441	rs692544	C	T	0.508	1.662	0.136	1.094e-06	African	Intergenic	*TFF2*

TABLE 3 | Summary statistics of the top results (p-value < 1 × 10^{-5}) whilst utilising the Local Ancestry Adjusted Allelic (LAAA) model.

Chr	Position	rsID	Ref	Alt	Altfreq	OR	SE	*p*-value	Ancestry	Location	Gene
1	208027696	rs61821315	C	T	0.132	1.141	0.356	8.804e-06	African	None	None
1	208029947	rs7550821	C	T	0.132	1.141	0.358	4.771e-06	African	None	None
1	208030856	rs7551724	C	T	0.131	1.140	0.359	8.065e-06	African	None	None
2	59343477	rs17049931	C	T	0.148	1.160	0.318	5.919e-06	African	None	None
2	172987232	rs7583008	T	C	0.230	1.259	0.318	9.628e-06	African	None	None
2	172987357	rs7569224	G	C	0.230	1.259	0.318	9.628e-06	African	None	None
3	22648301	rs1449916	T	C	0.394	1.483	0.307	3.665e-06	African	None	None
3	104923287	rs13061116	G	A	0.216	1.241	0.357	5.134e-06	African	None	None
3	104923579	rs1525840	T	C	0.216	1.241	0.357	6.076e-06	African	None	None
3	104924774	rs11923672	A	T	0.216	1.241	0.357	6.076e-06	African	None	None
3	104924866	rs11926446	G	A	0.216	1.241	0.357	5.134e-06	African	None	None
3	104929569	rs9834777	T	C	0.217	1.242	0.357	5.161e-06	African	None	None
5	31584670	rs10940959	C	A	0.123	1.131	0.380	5.596e-06	African	None	None
9	73895875	rs7037178	T	G	0.459	1.582	0.235	8.903e-06	African	Intronic	*TRPM3*
9	73899145	rs1504387	T	C	0.461	1.586	0.234	3.558e-06	African	Intronic	*TRPM3*
11	126163124	rs609634	T	C	0.261	1.298	0.255	7.570e-06	African	Intronic	*TIRAP*
18	65322790	rs1444107	T	A	0.082	1.085	0.446	6.644e-06	African	Intronic	*DSEL-AS1*
18	65323846	rs2448767	A	G	0.082	1.085	0.446	6.644e-06	African	Intronic	*DSEL-AS1*
18	65324070	rs2448766	A	G	0.082	1.085	0.446	6.644e-06	African	Intronic	*DSEL-AS1*
2	183351225	rs1594304	T	C	0.421	1.523	0.253	3.557e-06	Khoesan	Intronic	*PDE1A*
5	26027283	rs12659706	C	T	0.337	1.401	0.340	9.503e-06	Khoesan	None	None
6	9576203	rs4715321	G	T	0.406	1.501	0.288	7.051e-06	Khoesan	None	None
5	142454386	rs13340374	C	T	0.061	1.063	0.532	3.421e-06	European	Intronic	*ARHGAP26*
9	119475712	rs72763937	C	T	0.184	1.202	0.320	8.548e-06	European	Intronic	*ASTN2*
11	126163124	rs609634	T	C	0.261	1.298	0.311	9.909e-06	European	Intronic	*TIRAP*
20	36998495	rs11698149	T	C	0.064	1.066	0.520	7.203e-06	European	Intronic	*LBP*
9	73099454	AS	T	C	0.324	1.383	0.531	9.295e-06	SouthEast Asian	Intronic	*KLF9-DT*

investigating TLR variants and susceptibility to TB (Schurz et al., 2015). Additionally, chromosome 11p13 was also previously associated with African ancestry in a previous GWAS (Thye et al., 2012; Chimusa et al., 2014). If the allelic model was not used while adjusting for local ancestry, this lead variant located near the *TIRAP* gene would have been missed due to the tagging SNP being located on a different ancestral haplotype LD block. This underlines the importance of including the LAAA models in association studies investigating complex multi-way admixed individuals.

One variant (rs28647531) passed the significance threshold and is located on chromosome 4q22 using the allelic model adjusting for Bantu-speaking African local ancestry (**Figure 3**). This variant is an intronic variant and located near the *FSTL5* gene, which has not been associated with TB susceptibility previously. This gene is a coding protein and was previously associated with colorectal cancer and acute myeloid leukaemia (Lv et al., 2017). Previous investigations of TB susceptibility in a southern African cohort identified African-and KhoeSan ancestry to be associated with an increased risk for TB (Chimusa et al., 2014, 2014; Daya et al., 2014b). Likewise, previous association signals for TB susceptibility in Africans included the *WT1* gene located on chromosome 11p13 and locus 18q12 and polymorphisms in the *TLR8* genes (Thye et al., 2010, 2012; Chimusa et al., 2014). Although we did not validate these genes in our study, we did however elucidate a lead variant located on chromosome 18q12 for Bantu-speaking African ancestry whilst utilising the LAAA model, meaning both the minor allele and ancestry co-occurs in this region. A previously unmapped protein coding gene (*DSEL-AS1*) was identified to be in LD with a leading SNP located on chromosome 18q12 for Bantu-speaking African ancestry (**Supplementary Figure S6**). *DSEL-AS1* is a lncRNA gene and was previously associated with unipolar depression, asparagine levels, bipolar disorder, body mass index and gut microbiome levels (Shi et al., 2011; Rhee et al., 2013; Winham et al., 2014; Ishida et al., 2020), but no biological pathways or interactions were reported for this lncRNA.

Moreover, another lead variant was identified for Bantu-speaking African ancestry. Transient receptor potential cation channel subfamily Melastatin member 3 (*TRPM3*), located on chromosome 9, is a protein coding gene which belongs to the family of transient receptor potential (TRP) channels. *TRPM3* is a permeable non-selective cation gene channel (Zhao et al., 2020, 3). Therefore, this gene is essential for cellular calcium signalling and homeostasis. Previous GWAS indicated the potential role of *TRPM3* in the measurement of mean platelet volume and were previously discovered in mostly European individuals (Astle et al., 2016; Vuckovic et al., 2020). Another protein coding gene, Phosphodiesterase 1A (*PDE1A*), is involved in calcium signalling and was amongst the lead variants identified for KhoeSan ancestry located on chromosome 2q14 by the LAAA model. This gene forms part of the cyclic nucleotide phosphodiesterases, which plays a role in signal transduction by regulating intracellular cyclic nucleotide concentrations through hydrolysis of cAMP and/or cGMP to their respective nucleoside 5-prime monophosphates. Therefore, this gene is

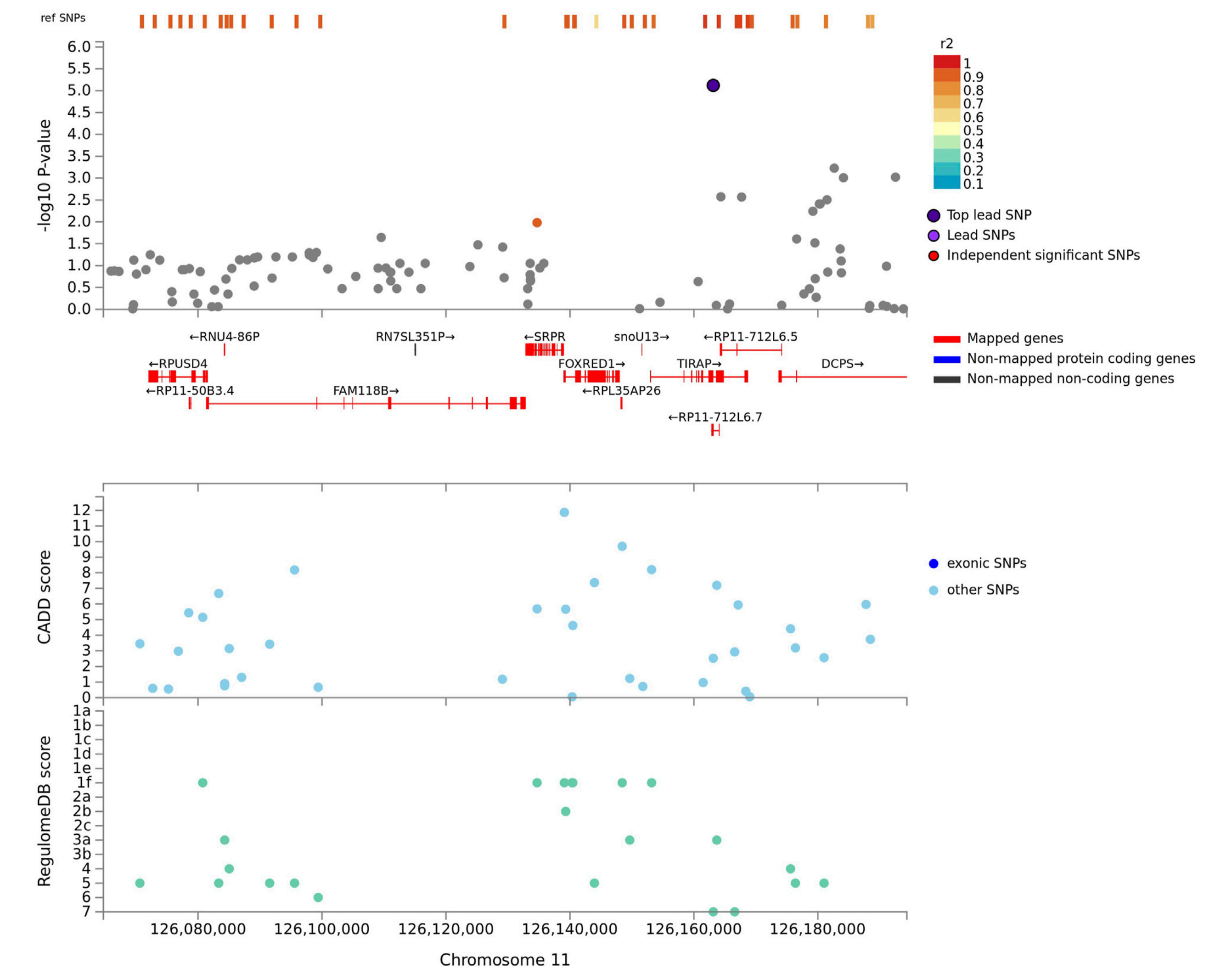


FIGURE 4 | Regional plot indicating the nearest genes in linkage disequilibrium for the lead variant observed for Bantu-speaking African ancestry on chromosome 11, whilst utilising the LAAA model. SNPs not in linkage disequilibrium are coloured grey and the lead variant is indicated in purple. Mapped genes are coded in red.

important for calmodulin binding and cGMP binding, as well as associated with urate measurement and glomerular filtration rate (Hellwege et al., 2019; Gill et al., 2021). Hence, there is evidence of the role of calcium ion channel activity in TB susceptibility, which includes the *FSTL5* gene and *TRPM3* gene for African ancestry, and the *PDE1A* gene for KhoeSan ancestry. *M.tb* modulates the levels and activity of key intracellular second messengers, such as calcium, to evade protective immune responses. Furthermore, calcium plays a crucial role in *M.tb* pathogenesis by activating differential transcription factors or mediating of the phagosome-lysosome fusion and cell survival (Sharma et al., 2016).

Our results demonstrate the benefit of simultaneously modelling allele, local ancestry, and ancestry-specific minor allelic effects when the admixed population under study exhibits extreme heterogeneity, since multiple distinct ancestry-specific genetic variants were identified for TB susceptibility that were previously missed by standard analyses. Thus, including an interaction term between the minor allele present and the corresponding ancestry of that minor allele can robustly identify ancestry-specific effects on disease phenotypes in a complex admixed population. It is important to mention that only variants that met certain quality control criteria during the imputation procedure were included in our analysis. Furthermore, minor alleles might have become evident after populations diverged, or have occurred in recent human history, and they are more likely to be ancestry-specific (Qin et al., 2019). The LAAA model first described by Duan et al. (2018) counts the number of reference alleles, whereas we counted the number of copies of the alternate alleles. Minor alleles might have become evident after populations diverged, or have occurred in recent human history, and they are more likely to be ancestry-specific (Qin et al., 2019). Therefore, allowing the detection of minor ancestry-specific allelic effects.

Currently there is no clear best practise for deriving the significance cut-off threshold for admixture mapping studies.

Every admixture scenario is unique in terms of contributing ancestral source populations, density markers analysed and particularly generations since admixture occurred. Moreover, in the presence of correlated tests the Bonferroni correction for multiple testing burden is overconservative for admixture mapping studies and does not necessarily control for family-wise error rate control in association analysis (Grinde et al., 2019). For this reason, we used the method described by Grinde et al. (2019), which entails a test statistic simulation directly from the asymptotic distribution implemented in the R software package *STEAM*. It considers the number of contributing ancestral populations, number of generations since admixture occurred and the distribution of admixture proportions in the cohort of interest and permutes these factors 1,000 times to get a new cut-off for significance (Grinde et al., 2019).

A limitation of the current study is the small sample size and findings should be validated in additional larger cohorts from various ethnic groups. Given our sample size of 735 participants (392 TB cases and 346 controls), we have 95% chance to correctly rejecting the null hypothesis for large (>0.5) and medium effect sizes (>0.3). We do however lose power if the effect size is small (0,1–0,3) and any reported associations with a smaller effect size should therefore be interpreted with caution (**Supplementary Figure S7**). Furthermore, there is a possibility that the true effect could be smaller than 0.1 for ancestry-specific effects in five different continental populations, confounding the study power (Skotte et al., 2019). Since literature suggests that TB susceptibility is governed by numerous SNPs with small effect sizes, we may have missed true local ancestry effects (type 2 errors) due to our small sample size. To report on ancestry-specific susceptibility to TB in a multi-way admixed southern African population, we estimate that at least 5,568 participants are required to confidently identify markers with smaller effect sizes (0.1–0.3).

Future studies should also include *in silico* and *in vitro* validation. Moreover, progression to active TB might be explained by numerous variants having a small effect on disease outcome, or exceptionally rare variants (Schurz et al., 2015). Variants that are unique to different populations and at low frequency should also be interrogated in well-powered studies. In addition, the information on the infecting *M.tb* strain should also be included in association analysis, if possible, since it appears that *M.tb* co-evolved with humans (Brites and Gagneux, 2015) and that the interaction between host genes and *M.tb* lineage affects TB severity (Müller et al., 2021). The combination of the ancestral allele and older *M.tb* lineages, i.e., the genotype and lineage that co-existed historically, had the lowest average TB score (McHenry et al., 2020). According to the TB score system, individuals are ranked according to their relative risk of being infected with TB given certain diagnostic information. A TB score of more than 40 indicates that a TB diagnosis is highly likely, a score of 30–35 indicates a possible TB diagnosis and a score below 25 indicates an unlikely diagnosis (dos Santos et al., 2017). Thus, the host populations that were historically exposed to a specific lineage have a lower chance of disease. Similarly, the average TB score for the combinations of genotype and lineage that have not historically co-existed, were the highest (McHenry et al., 2020). Thus, the evolutionary history of both species should be considered together.

In conclusion, this is the first study to apply the LAAA model to a complex five-way admixed population from South Africa which exhibits extensive genetic heterogeneity. This was enabled by newly developed algorithms for local ancestry inference, updated reference panels to represent contributing ancestral populations and a more suitable genotyping platform for diverse populations worldwide. We have demonstrated that the LAAA model robustly captured the source of association signals in highly complex admixed individuals. The true underlying architecture at each locus is unknown for most southern African populations, indicating that careful consideration of both global-and local ancestry is required for successful complex-trait mapping. Furthermore, local ancestry information across the genome is likely to become relevant to determine whether a genetic variant is expected to be useful in precision medicine, specifically in admixed populations.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: No new genetic data was generated for this study however, summary statistics for the quality and accuracy assessment of the genetic data will be made available to researchers who meet the criteria for access after application to the Health Research Ethics Committee of Stellenbosch University. Requests to access these datasets should be directed to MM, marlom@sun.ac.za.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Health Research Ethics Committee of the Faculty of Health Sciences, Stellenbosch University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YS performed all computational analyses. MM and CU conceived of the presented idea and supervised the work. EH and PvH contributed to the implementation of the study and established the sample bank. All authors contributed to the writing and proofreading of the final manuscript.

FUNDING

This research was partially funded by the South African government through the South African Medical Research Council and the National Research Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the South African Medical Research Council or the National Research Foundation. YS was supported by a Stellenbosch University Postgraduate Bursary.

ACKNOWLEDGMENTS

We would like to acknowledge and thank the study participants for their contribution and participation in projects from which genetic data was used.

REFERENCES

Alexander, D. H., and Lange, K. (2011). Enhancements to the ADMIXTURE Algorithm for Individual Ancestry Estimation. *BMC Bioinformatics* 12, 246. doi:10.1186/1471-2105-12-246

Astle, W. J., Elding, H., Jiang, T., Allen, D., Ruklisa, D., Mann, A. L., et al. (2016). The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 167, 1415–1429. doi:10.1016/j.cell.2016.10.042

Bañuls, A.-L., Sanou, A., Van Anh, N. T., and Godreuil, S. (2015). *Mycobacterium tuberculosis*: Ecology and Evolution of a Human Bacterium. *J. Med. Microbiol.* 64, 1261–1269. doi:10.1099/jmm.0.000171

Bao, Z., Chen, R., Zhang, P., Lu, S., Chen, X., Yao, Y., et al. (2016). A Potential Target Gene for the Host-Directed Therapy of Mycobacterial Infection in Murine Macrophages. *Int. J. Mol. Med.* 38, 823–833. doi:10.3892/ijmm.2016.2675

Behr, A. A., Liu, K. Z., Liu-Fang, G., Nakka, P., and Ramachandran, S. (2016). Pong: Fast Analysis and Visualization of Latent Clusters in Population Genetic Data. *Bioinformatics* 32, 2817–2823. doi:10.1093/bioinformatics/btw327

Brites, D., and Gagneux, S. (2015). Co-evolution of M Ycobacterium Tuberculosis and H Omo Sapiens. *Immunol. Rev.* 264, 6–24. doi:10.1111/imr.12264

Cai, L., Li, Z., Guan, X., Cai, K., Wang, L., Liu, J., et al. (2019). The Research Progress of Host Genes and Tuberculosis Susceptibility. *Oxidative Med. Cell Longevity* 2019, 1–8. doi:10.1155/2019/9273056

Capparelli, R., De Chiara, F., Di Matteo, A., Medaglia, C., and Iannelli, D. (2013). The MyD88 Rs6853 and TIRAP Rs8177374 Polymorphic Sites Are Associated with Resistance to Human Pulmonary Tuberculosis. *Genes Immun.* 14, 504–511. doi:10.1038/gene.2013.48

Chaw, L., Chien, L.-C., Wong, J., Takahashi, K., Koh, D., and Lin, R.-T. (2020). Global Trends and Gaps in Research Related to Latent Tuberculosis Infection. *BMC Public Health* 20, 352. doi:10.1186/s12889-020-8419-0

Chimusa, E. R., Zaitlen, N., Daya, M., Möller, M., van Helden, P. D., Mulder, N. J., et al. (2014). Genome-wide Association Study of Ancestry-specific TB Risk in the South African Coloured Population. *Hum. Mol. Genet.* 23, 796–809. doi:10.1093/hmg/ddt462

Curtis, J., Luo, Y., Zenner, H. L., Cuchet-Lourenço, D., Wu, C., Lo, K., et al. (2015). Susceptibility to Tuberculosis Is Associated with Variants in the ASAP1 Gene Encoding a Regulator of Dendritic Cell Migration. *Nat. Genet.* 47, 523–527. doi:10.1038/ng.3248

Daya, M., van der Merwe, L., Gignoux, C. R., van Helden, P. D., Möller, M., and Hoal, E. G. (2014a). Using Multi-Way Admixture Mapping to Elucidate TB Susceptibility in the South African Coloured Population. *BMC Genomics* 15. doi:10.1186/1471-2164-15-1021

Daya, M., van der Merwe, L., van Helden, P. D., Möller, M., and Hoal, E. G. (2014b). The Role of Ancestry in TB Susceptibility of an Admixed South African Population. *Tuberculosis* 94, 413–420. doi:10.1016/j.tube.2014.03.012

de Wit, E., Delport, W., Rugamika, C. E., Meintjes, A., Möller, M., van Helden, P. D., et al. (2010). Genome-wide Analysis of the Structure of the South African Coloured Population in the Western Cape. *Hum. Genet.* 128, 145–153. doi:10.1007/s00439-010-0836-1

Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F., and Marchini, J. (2013). Haplotype Estimation Using Sequencing Reads. *Am. J. Hum. Genet.* 93, 687–696. doi:10.1016/j.ajhg.2013.09.002

Delaneau, O., Marchini, J., and Marchini, J. (2014). Integrating Sequence and Array Data to Create an Improved 1000 Genomes Project Haplotype Reference Panel. *Nat. Commun.* 5, 1–9. doi:10.1038/ncomms4934

dos Santos, I. C. C., Genre, J., Marques, D., da Silva, A. M. G., dos Santos, J. C., de Araújo, J. N. G., et al. (2017). A New Panel of SNPs to Assess Thyroid Carcinoma Risk: a Pilot Study in a Brazilian Admixture Population. *BMC Med. Genet.* 18, 140. doi:10.1186/s12881-017-0502-8

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.716558/full#supplementary-material

Duan, Q., Xu, Z., Raffield, L. M., Chang, S., Wu, D., Lange, E. M., et al. (2018). A Robust and Powerful Two-step Testing Procedure for Local Ancestry Adjusted Allelic Association Analysis in Admixed Populations. *Genet. Epidemiol.* 42, 288–302. doi:10.1002/gepi.22104

El Kamel, A., Joobeur, S., Skhiri, N., Cheikh Mhamed, S., Mribah, H., and Rouatbi, N. (2015). La lutte antituberculeuse dans le monde. *Revue de Pneumologie Clinique* 71, 181–187. doi:10.1016/j.pneumo.2014.03.004

Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (2009). Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behav. Res. Methods* 41, 1149–1160. doi:10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behav. Res. Methods* 39, 175–191. doi:10.3758/BF03193146

Gallant, C. J., Cobat, A., Simkin, L., Black, G. F., Stanley, K., Hughes, J., et al. (2010). Impact of Age and Sex on Mycobacterial Immunity in an Area of High Tuberculosis Incidence. *Int. J. Tuberc. Lung Dis.* 14, 952–959.

Gill, D., Cameron, A. C., Burgess, S., Li, X., Doherty, D. J., Karhunen, V., et al. (2021). Urate, Blood Pressure, and Cardiovascular Disease. *Hypertension* 77, 383–392. doi:10.1161/HYPERTENSIONAHA.120.16547

Grant, A. V., Sabri, A., Abid, A., Abderrahmani Rhorfi, I., Benkirane, M., Souhi, H., et al. (2016). A Genome-wide Association Study of Pulmonary Tuberculosis in Morocco. *Hum. Genet.* 135, 299–307. doi:10.1007/s00439-016-1633-2

Grinde, K. E., Brown, L. A., Reiner, A. P., Thornton, T. A., and Browning, S. R. (2019). Genome-wide Significance Thresholds for Admixture Mapping Studies. *Am. J. Hum. Genet.* 104, 454–465. doi:10.1016/j.ajhg.2019.01.008

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., et al. (2015). The African Genome Variation Project Shapes Medical Genetics in Africa. *Nature* 517, 327–332. doi:10.1038/nature13997

Hellwege, J. N., Velez Edwards, D. R., Giri, A., Qiu, C., Park, J., Torstenson, E. S., et al. (2019). Mapping eGFR Loci to the Renal Transcriptome and Phenome in the VA Million Veteran Program. *Nat. Commun.* 10, 3842. doi:10.1038/s41467-019-11704-w

Ishida, S., Kato, K., Tanaka, M., Odamaki, T., Kubo, R., Mitsuyama, E., et al. (2020). Genome-wide Association Studies and Heritability Analysis Reveal the Involvement of Host Genetics in the Japanese Gut Microbiota. *Commun. Biol.* 3, 1–10. doi:10.1038/s42003-020-01416-z

Kinnear, C., Hoal, E. G., Schurz, H., van Helden, P. D., and Möller, M. (2017). The Role of Human Host Genetics in Tuberculosis Resistance. *Expert Rev. Respir. Med.* 11, 721–737. doi:10.1080/17476348.2017.1354700

Luo, Y., Suliman, S., Asgari, S., Amariuta, T., Baglaenko, Y., Martínez-Bonet, M., et al. (2019). Early Progression to Active Tuberculosis Is a Highly Heritable Trait Driven by 3q23 in Peruvians. *Nat. Commun.* 10. doi:10.1038/s41467-019-11664-1

Lv, H., Zhang, M., Shang, Z., Li, J., Zhang, S., Lian, D., et al. (2017). Genome-wide Haplotype Association Study Identify the FGFR2 Gene as a Risk Gene for Acute Myeloid Leukemia. *Oncotarget* 8, 7891–7899. doi:10.18632/oncotarget.13631

Mahasirimongkol, S., Yanai, H., Mushiroda, T., Promphittayarat, W., Wattanapokayakit, S., Phromjai, J., et al. (2012). Genome-wide Association Studies of Tuberculosis in Asians Identify Distinct At-Risk Locus for Young Tuberculosis. *J. Hum. Genet.* 57, 363–367. doi:10.1038/jhg.2012.35

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust Relationship Inference in Genome-wide Association Studies. *Bioinformatics* 26, 2867–2873. doi:10.1093/bioinformatics/btq559

Maples, B. K., Gravel, S., Kenny, E. E., and Bustamante, C. D. (2013). RFMix: a Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am. J. Hum. Genet.* 93, 278–288. doi:10.1016/j.ajhg.2013.06.020

McHenry, M. L., Williams, S. M., and Stein, C. M. (2020). Genetics and Evolution of Tuberculosis Pathogenesis: New Perspectives and Approaches. *Infect. Genet. Evol.* 81, 104204. doi:10.1016/j.meegid.2020.104204

Möller, M., Nebel, A., Valentonyte, R., van Helden, P. D., Schreiber, S., and Hoal, E. G. (2009). Investigation of Chromosome 17 Candidate Genes in Susceptibility to TB in a South African Population. *Tuberculosis* 89, 189–194. doi:10.1016/j.tube.2008.10.001

Müller, S. J., Schurz, H., Tromp, G., van der Spuy, G. D., Hoal, E. G., van Helden, P. D., et al. (2021). A Multi-Phenotype Genome-wide Association Study of Clades Causing Tuberculosis in a Ghanaian- and South African Cohort. *Genomics* 113, 1802–1815. doi:10.1016/j.ygeno.2021.04.024

Nguyen, H. C., Wang, W., and Xiong, Y. (2017). Cullin-RING E3 Ubiquitin Ligases: Bridges to Destruction. *Subcell Biochem.* 83, 323–347. doi:10.1007/978-3-319-46503-6_12

Oki, N. O., Motsinger-Reif, A. A., Antas, P. R., Levy, S., Holland, S. M., and Sterling, T. R. (2011). Novel Human Genetic Variants Associated with Extrapulmonary Tuberculosis: a Pilot Genome Wide Association Study. *BMC Res. Notes* 4, 28. doi:10.1186/1756-0500-4-28

Omae, Y., Toyo-oka, L., Yanai, H., Nedsuwan, S., Wattanapokayakit, S., Satproedprai, N., et al. (2017). Pathogen Lineage-Based Genome-wide Association Study Identified CD53 as Susceptible Locus in Tuberculosis. *J. Hum. Genet.* 62, 1015–1022. doi:10.1038/jhg.2017.82

Png, E., Alisjahbana, B., Sahiratmadja, E., Marzuki, S., Nelwan, R., Balabanova, Y., et al. (2012). A Genome Wide Association Study of Pulmonary Tuberculosis Susceptibility in Indonesians. *BMC Med. Genet.* 13, 5. doi:10.1186/1471-2350-13-5

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

Qi, H., Zhang, Y.-B., Sun, L., Chen, C., Xu, B., Xu, F., et al. (2017). Discovery of Susceptibility Loci Associated with Tuberculosis in Han Chinese. *Hum. Mol. Genet.* 26, 4752–4763. doi:10.1093/hmg/ddx365

Qin, H., Zhao, J., and Zhu, X. (2019). Identifying Rare Variant Associations in Admixed Populations. *Sci. Rep.* 9, 5458. doi:10.1038/s41598-019-41845-3

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., et al. (2010). Strong Maternal Khoisan Contribution to the South African Coloured Population: A Case of Gender-Biased Admixture. *Am. J. Hum. Genet.* 86, 654. doi:10.1016/j.ajhg.2010.03.015

Rhee, E. P., Ho, J. E., Chen, M.-H., Shen, D., Cheng, S., Larson, M. G., et al. (2013). A Genome-wide Association Study of the Human Metabolome in a Community-Based Cohort. *Cel Metab.* 18, 130–143. doi:10.1016/j.cmet.2013.06.013

Rossouw, M., Nel, H. J., Cooke, G. S., van Helden, P. D., and Hoal, E. G. (2003). Association between Tuberculosis and a Polymorphic NFκB Binding Site in the Interferon γ Gene. *The Lancet* 361, 1871–1872. doi:10.1016/S0140-6736(03)13491-5

Rudko, A. A., Bragina, E. Y., Puzyrev, V. P., and Freidin, M. B. (2016). The Genetics of Susceptibility to Tuberculosis: Progress and Challenges. *Asian Pac. J. Trop. Dis.* 6, 680–684. doi:10.1016/S2222-1808(16)61109-X

Schurz, H., Daya, M., Möller, M., Hoal, E. G., and Salie, M. (2015). TLR1, 2, 4, 6 and 9 Variants Associated with Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis. *PLOS ONE* 10, e0139711. doi:10.1371/journal.pone.0139711

Schurz, H., Kinnear, C. J., Gignoux, C., Wojcik, G., van Helden, P. D., Tromp, G., et al. (2019a). A Sex-Stratified Genome-wide Association Study of Tuberculosis Using a Multi-Ethnic Genotyping Array. *Front. Genet.* 9. doi:10.3389/fgene.2018.00678

Schurz, H., Müller, S. J., van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., et al. (2019b). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Front. Genet.* 10. doi:10.3389/fgene.2019.00034

Selvaraj, P., Harishankar, M., Singh, B., Jawahar, M. S., and Banurekha, V. V. (2010). Toll-like Receptor and TIRAP Gene Polymorphisms in Pulmonary Tuberculosis Patients of South India. *Tuberculosis* 90, 306–310. doi:10.1016/j.tube.2010.08.001

Sharma, D., Tiwari, B. K., Mehto, S., Antony, C., Kak, G., Singh, Y., et al. (2016). Suppression of Protective Responses upon Activation of L-type Voltage Gated Calcium Channel in Macrophages during Mycobacterium Bovis BCG Infection. *PLOS ONE* 11, e0163845. doi:10.1371/journal.pone.0163845

Shi, J., Potash, J. B., Knowles, J. A., Weissman, M. M., Coryell, W., Scheftner, W. A., et al. (2011). Genome-wide Association Study of Recurrent Early-Onset Major Depressive Disorder. *Mol. Psychiatry* 16, 193–201. doi:10.1038/mp.2009.124

Shriner, D. (2013). Overview of Admixture Mapping. *Curr. Protoc. Hum. Genet.* 76. , 2013 Unit 1.23. doi:10.1002/0471142905.hg0123s76

Skotte, L., Jørsboe, E., Korneliussen, T. S., Moltke, I., and Albrechtsen, A. (2019). Ancestry-specific Association Mapping in Admixed Populations. *Genet. Epidemiol.* 43, 506–521. doi:10.1002/gepi.22200

Sobota, R. S., Stein, C. M., Kodaman, N., Scheinfeldt, L. B., Maro, I., Wieland-Alter, W., et al. (2016). A Locus at 5q33.3 Confers Resistance to Tuberculosis in Highly Susceptible Individuals. *Am. J. Hum. Genet.* 98, 514–524. doi:10.1016/j.ajhg.2016.01.015

Thye, T., Owusu-Dabo, E., Vannberg, F. O., van Crevel, R., Curtis, J., Sahiratmadja, E., et al. (2012). Common Variants at 11p13 Are Associated with Susceptibility to Tuberculosis. *Nat. Genet.* 44, 257–259. doi:10.1038/ng.1080

Thye, T., Vannberg, F. O., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., et al. (2010). Genome-wide Association Analyses Identifies a Susceptibility Locus for Tuberculosis on Chromosome 18q11.2. *Nat. Genet.* 42, 739–741. doi:10.1038/ng.639

Uren, C., Henn, B. M., Franke, A., Wittig, M., van Helden, P. D., Hoal, E. G., et al. (2017a). A post-GWAS Analysis of Predicted Regulatory Variants and Tuberculosis Susceptibility. *PLoS ONE* 12, e0174738. doi:10.1371/journal.pone.0174738

Uren, C., Hoal, E. G., and Möller, M. (2020). Putting RFMix and ADMIXTURE to the Test in a Complex Admixed Population. *BMC Genet.* 21, 40. doi:10.1186/s12863-020-00845-3

Uren, C., Kim, M., Martin, A. R., Bobo, D., Gignoux, C. R., van Helden, P. D., et al. (2016). Fine-scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. *Genetics* 204, 303–314. doi:10.1534/genetics.116.187369

Uren, C., Möller, M., van Helden, P. D., Henn, B. M., and Hoal, E. G. (2017b). Population Structure and Infectious Disease Risk in Southern Africa. *Mol. Genet. Genomics* 292, 499–509. doi:10.1007/s00438-017-1296-2

van Helden, P. D., Möller, M., Babb, C., Warren, R., Walzl, G., Uys, P., et al. (2006). TB Epidemiology and Human Genetics. *Novartis Found. Symp.* 279, 17–19. discussion 31-41, 216–219.

Vuckovic, D., Bao, E. L., Akbari, P., Lareau, C. A., Mousas, A., Jiang, T., et al. (2020). The Polygenic and Monogenic Basis of Blood Traits and Diseases. *Cell* 182, 1214–e11. doi:10.1016/j.cell.2020.08.008

Wang, K., Goldstein, S., Bleasdale, M., Clist, B., Bostoen, K., Bakwa-Lufu, P., et al. (2020). Ancient Genomes Reveal Complex Patterns of Population Movement, Interaction, and Replacement in Sub-saharan Africa. *Sci. Adv.* 6, eaaz0183. doi:10.1126/sciadv.aaz0183

WHO (2019). Global Tuberculosis Report 2019. *WHO*. Available at: http://www.who.int/tb/publications/global_report/en/ (Accessed October 24, 2019).

Winham, S. J., Cuellar-Barboza, A. B., Oliveros, A., McElroy, S. L., Crow, S., Colby, C., et al. (2014). Genome-wide Association Study of Bipolar Disorder Accounting for Effect of Body Mass index Identifies a New Risk Allele in TCF7L2. *Mol. Psychiatry* 19, 1010–1016. doi:10.1038/mp.2013.159

Yim, J.-J., and Selvaraj, P. (2010). Genetic Susceptibility in Tuberculosis. *Respirology* 15, 241–256. doi:10.1111/j.1440-1843.2009.01690.x

Zhang, Y. X., Xue, Y., Liu, J. Y., Zhao, M. Y., Li, F. J., Zhou, J. M., et al. (2011). Association of TIRAP (MAL) Gene Polymorhisms with Susceptibility to Tuberculosis in a Chinese Population. *Genet. Mol. Res.* 10, 7–15. doi:10.4238/vol10-1gmr980

Zhao, S., Yudin, Y., and Rohacs, T. (2020). Disease-associated Mutations in the Human TRPM3 Render the Channel Overactive via Two Distinct Mechanisms. *Elife* 9, e55634. doi:10.7554/eLife.55634

Zheng, R., Li, Z., He, F., Liu, H., Chen, J., Chen, J., et al. (2018). Genome-wide Association Study Identifies Two Risk Loci for Tuberculosis in Han Chinese. *Nat. Commun.* 9, 4072. doi:10.1038/s41467-018-06539-w

Zhou, H., Alexander, D., and Lange, K. (2011). A Quasi-Newton Acceleration for High-Dimensional Optimization Algorithms. *Stat. Comput.* 21, 261–273. doi:10.1007/s11222-009-9166-3

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.

Identification of Bovine miRNAs with the Potential to Affect Human Gene Expression

Moldir Myrzabekova[1], *Siegfried Labeit*[2,3], *Raigul Niyazova*[1], *Aigul Akimniyazova*[1] *and Anatoliy Ivashchenko*[1]*

[1]Faculty of Biology and Biotechnology, Al-Farabi Kazakh National University, Almaty, Kazakhstan, [2]Medical Faculty Mannheim, University of Heidelberg, Heidelberg, Germany, [3]Myomedix GmbH, Neckargemuend, Germany

OPEN ACCESS

Edited by:
Elvira Galieva,
Novosibirsk State University, Russia

Reviewed by:
Tengfei Zhu,
Chinese Academy of Fishery Sciences (CAFS), China
Paola Lorena Smaldini,
CONICET Instituto de Estudios Inmunológicos y Fisiopatalógicos (IIFP), Argentina
Diana Resendez,
Universidad Autónoma de Nuevo León, Mexico

****Correspondence:***
Anatoliy Ivashchenko
a.iavashchenko@gmail.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *05 May 2021*
Accepted: *02 December 2021*
Published: *11 January 2022*

Citation:
Myrzabekova M, Labeit S, Niyazova R, Akimniyazova A and Ivashchenko A (2022) Identification of Bovine miRNAs with the Potential to Affect Human Gene Expression.
Front. Genet. 12:705350.
doi: 10.3389/fgene.2021.705350

Milk and other products from large mammals have emerged during human evolution as an important source of nutrition. Recently, it has been recognized that exogenous miRNAs (mRNA inhibited RNA) contained in milk and other tissues of the mammalian body can enter the human body, which in turn have the ability to potentially regulate human metabolism by affecting gene expression. We studied for exogenous miRNAs from *Bos taurus* that are potentially contain miRNAs from milk and that could act postprandially as regulators of human gene expression. The interaction of 17,508 human genes with 1025 bta-miRNAs, including 245 raw milk miRNAs was studied. The milk bta-miR-151-5p, bta-miR-151-3p, bta-miRNA-320 each have 11 BSs (binding sites), and bta-miRNA-345-5p, bta-miRNA-614, bta-miRNA-1296b and bta-miRNA-149 has 12, 14, 15 and 26 BSs, respectively. The bta-miR-574-5p from cow's milk had 209 human genes in mRNAs from one to 25 repeating BSs. We found 15 bta-miRNAs that have 100% complementarity to the mRNA of 13 human target genes. Another 12 miRNAs have BSs in the mRNA of 19 human genes with 98% complementarity. The bta-miR-11975, bta-miR-11976, and bta-miR-2885 BSs are located with the overlap of nucleotide sequences in the mRNA of human genes. Nucleotide sequences of BSs of these miRNAs in 5′UTR mRNA of human genes consisted of GCC repeats with a total length of 18 nucleotides (nt) in 18 genes, 21 nt in 11 genes, 24 nt in 14 genes, and 27–48 nt in nine genes. Nucleotide sequences of BSs of bta-miR-11975, bta-miR-11976, and bta-miR-2885 in CDS mRNA of human genes consisted of GCC repeats with a total length of 18 nt in 33 genes, 21 nt in 13 genes, 24 nt in nine genes, and 27–36 nt in 11 genes. These BSs encoded polyA or polyP peptides. In only one case, the polyR (*SLC24A3* gene) was encoded. The possibility of regulating the expression of human genes by exogenous bovine miRNAs is discussed.

Keywords: miRNA, exogenous miRNA, mRNA, gene, Bos taurus, human, disease

INTRODUCTION

The miRNAs (mRNA inhibited RNA) are 18–24-nucleotide-long RNA nanoscale molecules that are highly conserved among species. They regulate post-transcriptional gene expression either by inhibiting mRNA translation or by degrading through exonuclease action (Huntzinger and Izaurralde, 2011; Fabian and Sonenberg, 2012; Ipsaro and Joshua-Tor, 2015; Jonas and

Izaurrallde, 2015). In recent years, numerous studies have shown that the milk of humans and cows is enriched with miRNAs (*Chen et al., 2010*; *Weber et al., 2010*), most of which are packed in extracellular vesicles that are 30–120 nm in diameter, namely exosomes, which are derived from all types of cells and released into all biological fluids, such as blood plasma, serum, urine, breast milk, colostrum, and more (*Kosaka et al., 2010*; *Gu et al., 2012*; *Xiao et al., 2018*; *Yun et al., 2021*; *Zeng et al., 2020*; *Zeng et al.,2021*). Zhang et al. suggested that exogenous miRNAs (ex-miRs), specifically from plants, can withstand the digestion process, enter the animal's bloodstream through the gastrointestinal tract, and regulate gene expression (*Zhang L. et al., 2012*, *Zhang et al., 2012 Y.*; *Dickinson et al., 2013*; *Laubier* et *al., 2015*; *Title et al., 2015*; *Auerbach et al., 2016*; *Rakhmetullina et al., 2020*). It was shown that when piglets were fed pig or cow's milk, miRNAs could be absorbed both *in vivo* and *in vitro*, which creates the basis for understanding the participation of miRNAs in physiological functions *(Lin et al., 2020)*. miRNAs are key effectors in physiology and development of infants (*Wang L. et al., 2021*; *Zhou Q. et al., 2021*; *Leroux et al., 2021*; *Miao et al., 2021*; *Shah et al. 2021)*. In earlier works, it was shown that ex-miRs from food are bioactive, and both humans and animals can absorb miRNAs from a diet of plant or animal source (*Zhang L. et al., 2012*; *Wang et al., 2012*; *Baier et al., 2014*; *Zhou et al., 2015*). Baier et al. provided evidence that the amounts of miRNAs absorbed from milk are sufficient to alter human gene expression, i.e., miRNAs from one mammalian species can affect gene networks in another species (*Zhou et al., 2012*; *Baier et al., 2014*). Milk exosomes increase the stability of miRNAs, which facilitates their absorption through the digestive tract *(Aarts et al., 2021*; *Diomaiuto et al., 2021*; *Gao et al., 2021*; *Marsh et al., 2021*; *Wehbe & Kreydiyyeh, 2021)*. Animal and plant miRNAs are detected in all foods irrespective of processing (*Dever et al., 2015*; Benmoussa and Provost, 2019; *Mar-Aguilar et al., 2020*; *Melnik et al., 2021*). The study on the miRNA content of milk was undertaken by Izumi et al. (2012); they found that colostrum contained twice the amount of miRNAs found in mature milk, and that immune- and development-related miRNAs had significantly higher levels of expression (*Izumi et al., 2012; **Link et al., 2019**)*. The authors demonstrated that miRNAs and messenger RNAs that exist naturally in milk were resistant to acidic conditions and RNAses, as well as to industrial processing conditions. Humans absorb biologically meaningful amounts of miRNAs from nutritionally relevant doses of cow's milk; these miRNAs enter peripheral blood mononuclear cells and presumably other peripheral tissues, and physiological concentrations of milk miRNAs may affect human gene expression *in vivo* and in cell cultures (*Baier et al., 2014*; Lukasik and Zielenkiewicz, 2014; *Shu et al., 2015*; *Yim et al., 2016*).

Research in the field of dietary miRNAs has shifted away from miRNAs in plant-borne foods (*Rakhmetullina et al., 2020*) to miRNAs in foods of animal origin, particularly cow's milk, after observations that a large proportion of milk miRNAs is encapsulated in extracellular vesicles such as exosomes. The miRNAs encapsulated in milk exosomes are stable under harsh conditions, such as low pH and exposure to ribonucleases (*Kosaka et al., 2010*; *Izumi et al, 2012*).

Furthermore, it has been established that human exosomes can pass the gastrointestinal mucosa and deliver their miRNA to various peripheral tissues, possibly regulating their target genes (*Gu et al., 2012*; *Kusuma et al., 2016*; *Munagala et al., 2016*; *Bayraktar et al., 2017*; *Bahrami et al., 2018*; Rajagopal and Harikumar, 2018; *Xiao et al., 2018*; *Komine-Aizawa et al., 2020*).

Breast milk is the main source of nutrition and supply of the child, containing the biologically active substances that regulate the development of the body (*Kim et al., 2020*; McNeill and Hirschi, 2020). A significant part of regulatory molecules, including miRNAs, is transported in exosomes that are resistant to the conditions of the gastrointestinal tract, enter the bloodstream and spread throughout the recipient's body (*Reif et al., 2020*; *Akimniyazova et al., 2021*; *Melnik B., 2021*; *Billa et al., 2021*; *Melnik B. C., 2021*; *Wang X. et al., 2021*; *Jiang et al., 2021*; *Leroux et al., 2021*; *Lowry et al., 2021*; *Xia et al., 2021*).

Since miRNAs have long been considered as exclusively endogenously acting molecules, miRNA exocytosis, secretion, and possible cross-species signaling roles have not been a focus of research during the past decade (*Wang et al., 2012*; *Baier et al., 2014*; Lukasik and Zielenkiewicz, 2014; *Bryniarski et al., 2015*; *Munagala et al., 2016*; *Yim et al., 2016*; *Golan-Gerstl et al., 2017*; *Manca et al., 2018*; *van Herwijnen et al., 2018*; *Zempleni et al., 2018*). So far, only a few studies have explored whether humans can absorb a meaningful amount of certain exosomal miRNAs from cow's milk. The findings indicated that milk-derived miRNAs in pasteurized milk are absorbed by adults in meaningful amounts, and moreover, that endogenous miRNA synthesis cannot compensate for dietary deficiency (*Baier et al., 2014*; *Zhang et al., 2021*).

The high degree of sequence conservation of miRNAs between different mammalian species also suggests conserved functional roles of the miRNA-mRNA signal networks of orthologous genes. This is evidence of the similarity of the systems regulating the expression of mammalian genomes. The conservation of exogenous miRNAs, firstly, indicates the similarity of the systems for regulating the expression of genes and mammalian genomes and, secondly, allows manipulating miRNA changes as biocompatible regulators of biological processes. The aim of this work was to predict possible bovine miRNAs - human gene expression regulatory networks *in silico*. The obtained results will help to use exogenous miRNAs to purposefully change the expression of human genes.

MATERIALS AND METHODS

The nucleotide sequences of the 17 thousand mRNAs of targeted genes were downloaded from NCBI GenBank (http://www.ncbi.nlm.nih.gov accessed on 5 January 2020). The nucleotide sequences of the miRNAs were taken from miRBase v.22 (http://www.mirbase.org/ accessed on 5 January 2020). 1025 miRNAs encoded by the bovine genome are available in the miRBase database (https://www.mirbase.org/summary.shtml?org=bta). The miRNA BSs (binding sites) in the mRNAs of several genes were predicted using the MirTarget program (*Ivashchenko et al., 2014*; *Ivashchenko et al., 2016*). This program defines the following features of miRNA binding to mRNA: *1*) the initiation of the miRNA binding to the mRNAs from the first nucleotide of the mRNAs; *2*) the localization of the miRNA BSs in the 5′-untranslated region (5′UTR), coding domain sequence (CDS), and 3′-untranslated region (3′UTR) of the mRNAs; *3*) the schemes of nucleotide interactions between miRNAs and mRNAs *4*) the free energy of the interaction between miRNA and the mRNA (ΔG, kJ/mole); and the ratio ΔG/ΔGm (%) is determined for each site (ΔGm equals the free energy of the miRNA binding with its fully complementary nucleotide sequence). The MirTarget program finds hydrogen bonds between adenine (A) and uracil (U), guanine (G) and cytosine (C), G and U, and A and C. The free energy of interactions (ΔG) a pair of G and C is equal to 6. 37 kJ/mol, a pair of A and U is equal to 4.25 kJ/mol, G and U, A and C equal to 2.12 kJ/mol (Friedman and Honig, 1995). The distances between the bound A and C (1.04 nm) and G and U (1. 02 nm) are similar to those between bound G and C, A and U, which are equal to 1.03 nm (*Kool, 2001*; *Leontis et al., 2002*; *Garg and Heinemann, 2018*). The numbers of hydrogen bonds in the G–C, A–U, G–U, and A–C interactions were 3, 2, 1, and 1, respectively. By comparison, MirTarget differs from other programs in terms of finding the BSs of miRNA on the mRNAs of plant genes (*Dai et al., 2011*) in that *1*) it takes into account the interaction of the miRNA with mRNA over the entire miRNA sequence; *2*) it takes into account non-canonical pairs G–U and A–C; and *3*) it calculates the free energy of the interaction of the miRNA with mRNA, and when two or more miRNAs are bound with one mRNA or, if the BSs of two different miRNAs coincide in part, the preferred miRNA binding site is considered to be the one for which the free binding energy is greater. The adequacy of the program in terms of finding BSs has been confirmed in several publications (*Davis et al., 2005*; *Wang J. et al., 2016*, *Wang et al., 2016 Y.*; *Yurikova et al., 2019*). The MirTarget program predicts the BSs of plant and animal miRNAs equally well (*Bari et al., 2013, 2014*). To construct WebLogo schemes, Create Sequence Logos were used (https://weblogo.berkeley.edu/logo.cgi). A better confirmation of the obtained results than "wet" experiments is provided by the schemes of interaction of nucleotides along the entire length of the miRNAs and BSs. There are no "wet" experiments to find BSs for all miRNAs nucleotides with BSs and to determine the free energy of their interaction. The existing programs for determining BSs based only on "seed," in principle, cannot give adequate predictions of target genes and, moreover, the free energy of interaction of all nucleotides of miRNAs and BSs. In addition, widely used programs do not take into account the interaction of non-canonical nucleotide pairs, which significantly distorts the value of the free energy of interaction between miRNAs and BSs. Consideration of any of the above schemes shows which nucleotides of non-canonical pairs and in which position decrease the maximum possible energy of interaction between miRNAs and BSs. The schemes can be verified manually by finding the predicted miRNA BSs in the mRNA nucleotide sequence in the NCBI.

RESULTS

For mammalian predators, exogenous miRNAs can be ingested with raw food, and this pathway of miRNA transmission is preserved in nature during evolution. Some human food is prepared from meat without heat treatment, which does not lead to the destruction of miRNA.

The Interaction of Milk Bta-miRNAs with mRNAs of Human Genes

The results of the possible influence of miRNAs from milk presented in **Supplementary Table S1** shows, that out of 245 milk bta-miRNAs (*Chen et al., 2010*), 103 miRNAs can affect human protein synthesis. Consequently, the obtained data made it possible to select for future studies those miRNAs for which their target effect on human genes is known. Note that some miRNAs have more than 10 target genes. Interestingly, bta-miR-151-5p and bta-miR-151-3p, originating from the same pre-miRNA, each have 11 target genes, which may indicate the optimization of the energy costs of the synthesis of these miRNAs and the dependence of their target genes on a single source. The bta-miRNA-320, bta-miRNA-345-5p, bta-miRNA-614, bta-miRNA-1296b and bta-miRNA-149 have 11, 12, 14, 15 and 26 target genes, respectively. Consequently, these miRNAs have an increased effect on the expression of the human genome. The large numbers of human genes that can be targets of dairy bta-miRNAs indicate the possibility of using cow's milk for baby food.

We identified bta-miR-574-5p from cow's milk which had from one to 14 repeating BSs in mRNAs of 209 human genes (**Supplementary Table S2**). The bta-miR-574-5p and human miR-574-5p have identical nucleotide sequences, which, firstly, indicates that this miRNA is necessary for fetal development in the prenatal and postnatal periods, and, secondly, it controls the expression of a significant number of genes in the genome of cow and human. Since miR-574-5p is expressed in 17 genomes of other mammals (miRBase), its biological role is great in these organisms. The high value of ΔG from −115 kJ/mol to −123 kJ/mol indicates a significant interaction between miRNAs and BSs. The value of ΔG/ΔGm for associations of bta-miR-574-5p and *GABRB2, LRTM2, UBN2, SLITRK3* genes reaches 97%. The mRNA of the *SLITRK3* gene contains 12 BSs, which indicates a high probability of its interaction with bta-miR-574-5p.

We have selected 28 human target genes for bta-miR-574-5p, having the clusters of 14 or more BSs (**Table 1**). The bta-miR-574-5p and mRNA associations of these target genes had similar characteristics of the free energy of interaction. The increased number of BSs allows all of these mRNAs to interact with two bta-miR-574-5p at once, since the length of a cluster of 14 repeats is 50 nt with a bta-miR-574-5p length of 24 nt. These 28 bovine genes, like human target genes, also have BSs for miR-574-5p (**Table 1**). The obtained results indicate a strong dependence of the expression of many human genes on bta-miR-574-5p. One of the putative functions of miRNAs acting on many genes is to maintain their consistent expression. For example, increased expression of a gene will cause miR-574-5p to bind and decrease its concentration, which

TABLE 1 | Characteristics of interactions of bta-miR-574-5p with human mRNA 28 genes containing clusters of 14 and more repetitive BSs.

ID of human genes	Gene	ID of bovine genes	Start of first and last sites, nt	ΔG, kJ/mole	ΔG/ΔGm, %
ID:90416	*CCDC32*	ID: 506935	861–903 (22)	−115÷−119	90–93
ID: 959	*CD40LG*	ID: 282387	1550–1578 (15)	−119	93
ID: 2033	*EP300*	ID: 112446776	8556–8587 (16)	−115÷−119	90–93
ID: 1112	*FOXN3*	ID: 505469	2419–2445 (14)	−115÷−119	90–93
ID: 2674	*GFRA1*	ID: 534801	8452–8480 (15)	−119	93
ID: 2736	*GLI2*	ID: 510255	6118–6147 (15)	−115÷−119	90–93
ID: 2742	*GLRA2*	ID: 537660	2525–2567 (20)	−115÷−119	90–93
ID: 22801	*ITGA11*	ID: 523755	4599–4635 (22)	−115÷−119	90–93
ID: 84056	*KATNAL1*	ID: 537739	4197–4238 (20)	−115÷−119	90–93
ID: 56479	*KCNQ5*	ID: 613605	5367–5413 (24)	−115÷−119	90–93
ID: 653319	*KIAA0895L*	ID: 512420	2878–2927 (25)	−115÷−119	90–93
ID: 11155	*LDB3*	ID: 536781	4420–450 (15)	−115÷−119	90–93
ID:10186	*LHFP*	ID: 532944	1396–1425 (14)	−115÷−119	90–93
ID: 108927	*NCDN*	ID: 505994	3556–3586 (16)	−115÷−119	90–93
ID: 7101	*NR2E1*	ID: 528156	2954–2980 (14)	−119	93
ID: 5579	*PRKCB*	ID: 282325	7060–7088 (15)	−115÷−121	90–95
ID: 862	*RUNX1T1*	ID: 538628	3268–3300 (17)	−115÷−119	90–93
ID: 388228	*SBK1*	ID: 614815	2320–2361 (20)	−117÷−121	92–95
ID: 23157	*SEPT6*	ID: 540783	4494–4522 (15)	−115÷−119	90–93
ID: 9342	*SNAP29*	ID: 532261	1360–1386 (14)	−115÷−119	90–93
ID: 54558	*SPATA6*	ID: 534169	3206–3232 (14)	−115÷−119	90–93
ID: 727837	*SSX2B*	ID: 534692	1276–1307 (16)	−115÷−121	90–95
ID: 10214	*SSX3*	ID: 6757	1085–1130 (18)	−115÷−121	90–95
ID: 11346	*SYNPO*	ID: 533531	3598–3624 (14)	−115÷−119	90–93
ID: 202500	*TCTE1*	ID: 523600	2239–2267 (15)	−115÷−119	90–93
ID: 84951	*TNS4*	ID: 532898	3472–3512 (21)	−115÷−119	90–93
ID: 79865	*TREML2*	ID: 515548	2346–2379 (17)	−115÷−119	90–93
ID: 134510	*UBLCP1*	ID: 508163	1544–1573 (15)	−115÷−119	90–93

will lead to increased expression of other genes from the miR-574-5p target genes sample.

The Interaction of Bta-miRNAs and mRNA of Human Genes with High Complementarity

The total number of target genes for bta-miRNAs with 98–100% complementarity is 32 (**Table 2**). The nucleotide sequences of bta-miR-2881, bta-miR-2444, bta-miR-11975, bta-miR-135a, bta-miR-151-5p, bta-miR-1777b, bta-miR-1777a, bta-miR-2478, bta-miR-136, bta-miR-432, bta-miR-127, bta-miR-433, bta-miR-431 bta-miR-1282, and bta-miR-11976 BSs are full complementary to 13 mRNAs of human genes. The eight human miRNAs were identical to bovine miRNAs in name and nucleotide sequence. For these eight human miRNAs, the target genes indicated in **Table 2** were experimentally verified (*Davis et al., 2005*; *Wang J. et al., 2016*; *Yurikova et al., 2019*).

The mRNAs of 19 genes have BSs for 12 miRNAs with 98% complementarity. 15 mRNAs have BSs in 3′UTR, 10 mRNAs in CDS, and seven in 5′UTR and the free energy of interaction of miRNAs with mRNAs of these genes ranges from −93 to −127 kJ/mol. The bta-miR-2444 has five (*CXorf38, PTP4A2, ATP2B2, CELF2,* and *HDX)* human target genes, bta-miR-1777b (*MEX3A, RHOB,* and *HCN2*), bta-miR-6528 (*BCAM, TMEM164,* and *TNKS1BP1*), and bta-miR-1584-5p (*ARID1A, MCRS1,* and *SLIT3*) have three human target genes and bta-miR-151-5p (*LPPR5* and *LYPD3*), bta-miR-1777a (*MEX3A* and *RHOB*), and bta-miR-11975 (*EGFR* and *ZIC5*) has two human target genes. Other genes are targeted by one miRNA. Note that hsa-miR-619-5p binds fully complementary to the mRNA of more than 200 human genes (*Atambayeva et al., 2017*). The biological significance of the fully complementary binding of miRNAs to mRNAs remains a mystery since siRNAs typically destroy the mRNAs of the target gene.

The mRNA of the *MEX3A* gene has BSs for bta-miR-1777b and bta-miR-1777a. The mRNAs of *EGFR* and *RTL1* genes are targeted by bta-miR-11976 and bta-miR-431 with a high free energy of −127 kJ/mol. The gene targeted by the greatest number of miRNAs is *RTL1*, specifically five miRNAs: bta-miR-136, bta-miR-432, bta-miR-127, bta-miR-433, and bta-miR-431.

Despite the divergence of mammalian species, associations of miRNAs and their target genes persist for many millions of years, which indicate the importance of these associations in the regulation of genome expression. In recent years, quantitative characteristics of miRNA interactions with the mRNA of target genes have been established and the features of the organization of miRNA BS in mRNA of target genes have been revealed. This opens up new possibilities for regulating gene expression using miRNA (*Aisina et al., 2019*; *Mukushkina et al., 2020*; *Akimniyazova et al., 2021*; *Kamenova et al., 2021*). In particular, data on fully complementary miRNA and mRNA interactions raise

TABLE 2 | Characteristics of interactions of bta-miRNAs with human mRNA genes with high complementarity.

Gene	bta-miRNA	Start of site, nt	Region of miRNA	ΔG, kJ/mole	ΔG/ΔGm, %	Length, nt
AR	bta-miR-2881	416	5′UTR	–112	100	18
CXorf38	bta-miR-2444	1546	3′UTR	–93	100	20
EGFR	bta-miR-11975	87	5′UTR	–127	100	20
GLYCTK	bta-miR-135a[a,b]	2812	3′UTR	–113	100	23
LPPR5	bta-miR-151-5p[a]	1328	3′UTR	–113	100	21
LYPD3	bta-miR-151-5p[a]	1608	3′UTR	–113	100	21
MEX3A	bta-miR-1777b	301	CDS	–125	100	20
MEX3A	bta-miR-1777a	302	CDS	–123	100	20
PTP4A2	bta-miR-2444	2110	3′UTR	–93	100	20
RBM43	bta-miR-2478	2911	3′UTR	–106	100	20
RHOB	bta-miR-1777b	206	5′UTR	–125	100	20
RTL1	bta-miR-136[a,b]	110	CDS	–115	100	23
RTL1	bta-miR-432[a,b]	330	CDS	–123	100	23
RTL1	bta-miR-127[a,b]	1792	CDS	–121	100	22
RTL1	bta-miR-433[a,b]	2878	CDS	–119	100	22
RTL1	bta-miR-431[a,b]	3800	CDS	–127	100	23
SERF2	bta-miR-1282[a]	1072	CDS	–102	100	20
ZIC5	bta-miR-11976	1316	CDS	–134	100	21
ZIC5	bta-miR-11975	1317	CDS	–127	100	20
ARID1A	bta-miR-1584-5p	4587	CDS	–115	98	20
ATP2B2	bta-miR-2444	7003	3′UTR	–91	98	20
BCAM	bta-miR-6528	3012	3′UTR	–108	98	20
CELF2	bta-miR-2444	5313	3′UTR	–91	98	20
CTSH	bta-miR-2333	540	CDS	–117	98	21
HCN2	bta-miR-1777b	2374	CDS	–123	98	20
HDX	bta-miR-2444	5052	3′UTR	–91	98	20
HOXB8	bta-miR-196a[b]	1378	3′UTR	–110	98	22
KLF9	bta-miR-2897	799	5′UTR	–115	98	20
MCRS1	bta-miR-1584-5p	353	CDS	–115	98	20
RFNG	bta-miR-2412	1238	3′UTR	–125	98	22
RGL2	bta-miR-10173-5p	310	5′UTR	–117	98	22
RHOB	bta-miR-1777a	207	5′UTR	–121	98	20
SEPT8	bta-miR-151-3p[a]	2767	3′UTR	–110	98	21
SLIT3	bta-miR-1584-5p	57	5′UTR	–115	98	20
SPAM1	bta-miR-2285k	1320	CDS	–102	98	21
TAF4	bta-miR-1777b	763	CDS	–123	98	20
TMEM164	bta-miR-6528	2163	3′UTR	–108	98	20
TNKS1BP1	bta-miR-6528	5595	3′UTR	–108	98	20

[a] *Identical miRNAs, with human.*

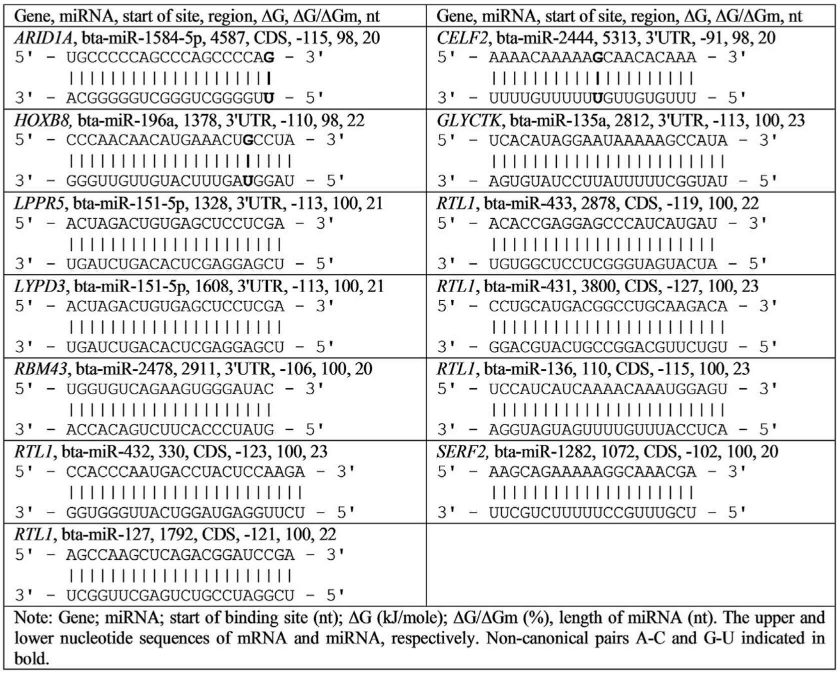
[b] *Milk miRNAs.*

questions about the function of such associations as they are similar to siRNA (**Table 2**). The complete and high complementarity of miRNA with the mRNA of many genes indicates a strong dependence of the participants in the miRNA and target gene associations.

The interaction of all nucleotides of miRNAs and mRNA of target genes shows how effectively these molecules bind. **Figure 1** shows the construction of hydrogen bonds between all nucleotides of bta-miR-1584-5p, bta-miR-2444, and bta-miR-196a and their BSs in mRNA. Because the MirTarget program considers the interaction of the non-canonical pairs A-C and G-U, the interaction of miRNAs and mRNAs preserves the spiral structures of both molecules, and therefore, stacking interactions are found between all nucleotides of miRNA and mRNA, which stabilize the duplex (*Garg and Heinemann, 2018*).

Characteristics of Bta-miR-11975, Bta-miR-11976, and Bta-miR-2885 Interactions with the CDS of mRNAs of Human Genes

As a result of the analysis of the interaction of 1025 bta-miRNAs with 17,508 human genes, we identified bta-miR-11975, bta-miR-11976, and bta-miR-2885, which had 118 human target genes. We predicted the cluster of BSs of bta-miR-11975, bta-miR-11976, and bta-miR-2885 in the mRNA of 66 human genes in CDSs. The BSs of these miRNAs (**Table 3**) consisted of repeating GCC triplets in CDS of mRNAs that encoded oligopeptides: polyP with a length of six amino acids (a.a.) in 16 genes, polyA six a.a. long in 17 genes, polyA seven a.a. long in eight genes, polyP seven a.a. long in four genes and polyR seven a.a. long in one gene, polyA eight a.a. long in five genes, polyP eight



<table>
<tr><th>Gene, miRNA, start of site, region, ΔG, ΔG/ΔGm, nt</th><th>Gene, miRNA, start of site, region, ΔG, ΔG/ΔGm, nt</th></tr>
<tr><td>*ARID1A*, bta-miR-1584-5p, 4587, CDS, -115, 98, 20
5' - UGCCCCCAGCCCAGCCCCA**G** - 3'
||||||||||||||||||||
3' - ACGGGGGUCGGGUCGGGGU**U** - 5'</td><td>*CELF2*, bta-miR-2444, 5313, 3'UTR, -91, 98, 20
5' - AAAACAAAAA**G**CAACACAAA - 3'
||||||||||||||||||||
3' - UUUUGUUUUU**U**GUUGUGUUU - 5'</td></tr>
<tr><td>*HOXB8,* bta-miR-196a, 1378, 3'UTR, -110, 98, 22
5' - CCCAACAACAUGAAACU**G**CCUA - 3'
||||||||||||||||||||||
3' - GGGUUGUUGUACUUUGA**U**GGAU - 5'</td><td>*GLYCTK*, bta-miR-135a, 2812, 3'UTR, -113, 100, 23
5' - UCACAUAGGAAUAAAAAGCCAUA - 3'
|||||||||||||||||||||||
3' - AGUGUAUCCUUAUUUUUCGGUAU - 5'</td></tr>
<tr><td>*LPPR5*, bta-miR-151-5p, 1328, 3'UTR, -113, 100, 21
5' - ACUAGACUGUGAGCUCCUCGA - 3'
|||||||||||||||||||||
3' - UGAUCUGACACUCGAGGAGCU - 5'</td><td>*RTL1*, bta-miR-433, 2878, CDS, -119, 100, 22
5' - ACACCGAGGAGCCCAUCAUGAU - 3'
||||||||||||||||||||||
3' - UGUGGCUCCUCGGGUAGUACUA - 5'</td></tr>
<tr><td>*LYPD3*, bta-miR-151-5p, 1608, 3'UTR, -113, 100, 21
5' - ACUAGACUGUGAGCUCCUCGA - 3'
|||||||||||||||||||||
3' - UGAUCUGACACUCGAGGAGCU - 5'</td><td>*RTL1*, bta-miR-431, 3800, CDS, -127, 100, 23
5' - CCUGCAUGACGGCCUGCAAGACA - 3'
|||||||||||||||||||||||
3' - GGACGUACUGCCGGACGUUCUGU - 5'</td></tr>
<tr><td>*RBM43*, bta-miR-2478, 2911, 3'UTR, -106, 100, 20
5' - UGGUGUCAGAAGUGGGAUAC - 3'
||||||||||||||||||||
3' - ACCACAGUCUUCACCCUAUG - 5'</td><td>*RTL1*, bta-miR-136, 110, CDS, -115, 100, 23
5' - UCCAUCAUCAAAACAAAUGGAGU - 3'
|||||||||||||||||||||||
3' - AGGUAGUAGUUUUGUUUACCUCA - 5'</td></tr>
<tr><td>*RTL1*, bta-miR-432, 330, CDS, -123, 100, 23
5' - CCACCCAAUGACCUACUCCAAGA - 3'
|||||||||||||||||||||||
3' - GGUGGGUUACUGGAUGAGGUUCU - 5'</td><td>*SERF2,* bta-miR-1282, 1072, CDS, -102, 100, 20
5' - AAGCAGAAAAAGGCAAACGA - 3'
||||||||||||||||||||
3' - UUCGUCUUUUUCCGUUUGCU - 5'</td></tr>
<tr><td>*RTL1*, bta-miR-127, 1792, CDS, -121, 100, 22
5' - AGCCAAGCUCAGACGGAUCCGA - 3'
||||||||||||||||||||||
3' - UCGGUUCGAGUCUGCCUAGGCU - 5'</td><td></td></tr>
<tr><td colspan="2">Note: Gene; miRNA; start of binding site (nt); ΔG (kJ/mole); ΔG/ΔGm (%), length of miRNA (nt). The upper and lower nucleotide sequences of mRNA and miRNA, respectively. Non-canonical pairs A-C and G-U indicated in bold.</td></tr>
</table>

FIGURE 1 | Schemes of the interaction of nucleotide sequences of bta-miRNAs and mRNA human genes.

a.a. long in four genes, polyP nine a.a. long in one gene, polyP nine a.a. long in six genes, polyA ten a.a. long in two genes, polyA 11 a.a. long in one gene and polyP 12 a.a. long in one gene. There is a cluster 18 nt long of BSs of bta-miR-11975 and bta-miR-11976 in the 3′UTR of mRNA of only one human gene.

Figure 2 shows the high conservatism of bovine miRNAs BSs in CDS and 5′UTR mRNAs of human genes. To confirm the interaction of cluster of bta-miR-11975, bta-miR-11976, and bta-miR-2885 BSs with human genes, we plotted the WebLogo for mRNA sections containing conservative BSs in CDS and 5′UTR with various lengths, specifically 18, 21, 24, 27–36, and 27–45 nt.

Supplementary Tables S3–S6 show characteristics of the interaction of bta-miR-11975, bta-miR-11976, and bta-miR-2885 BSs with the mRNA of human genes in a cluster. The mRNAs of *CASZ1, FOXK1, NANOS1, POU3F3,* and *TSPYL2* genes have clusters of BSs for three miRNAs, and other genes can bind two miRNAs. The mRNAs of *GPR88, LOXL1, LTBP1, MECP2,* and *TMEM121* genes can bind single bta-miR-11975, with a free energy −115 to −117 kJ/mol. BSs of bta-miR-11976 and bta-miR-11975 in mRNAs of *ATOH8, GPR150, FAM117B, FOXD1, HOXA2, TPRN, ZNF367,* and *ZNF839* genes are located through one nucleotide. In mRNAs of *CASZ1, NANOS1, POU3F3, TSPYL2,* and *UNCX* genes BSs of bta-miR-11976 and bta-miR-2885 are located with overlapping, that is, they form a cluster. The mRNAs of *FOXG1, LCORL,* and *SOX12* genes have multiple BSs for bta-miR-11975 located through three nucleotides. Therefore, of these three miRNAs, only one can bind to the mRNA of target genes. At equal concentrations of miRNAs, it will preferentially bind with bta-miR-11976 since it has more free energy of interaction with the mRNA. However, at significantly higher concentrations of bta-miR-11975 and bta-miR-2885, they will preferentially bind to the mRNA of the target gene.

The mRNA of the *GABBR2* gene is characterized by the presence of two clusters: the first cluster starts at 276 nt and ends at 306 nt, with a length of 30 nt. The second cluster is 32 nt long, starting at 487 nt and ending at 519 nt. The bta-miR-11976 interacts with *HOXA13* mRNA with the free energy more than −129 kJ/mol and has two BSs: the first starts at 399 nt and ends at 428 nt, and the second starts at 600 nt and ends at 641 nt. The mRNA of the *FBXL17* gene contains four BSs for bta-miR-11975 and two BSs for bta-miR-11976 located through three nucleotides. The mRNAs of *IRX2* and *IRX4* genes have three BSs for bta-miR-11975 and two BSs for miR-11976 located through three nucleotides.

The mRNA of the *POU3F3* gene contains multiple BSs for three miRNAs which present in two clusters. The cluster size is 41 nt. The first cluster starts at position 304 nt and ends at position 345 nt. The mir-11975 has eight BSs and miR-11976 has six BSs located through three nucleotides. The free energy of interactions of the three miRNAs with the mRNAs ranges from −110 to −123 kJ/mol. The second cluster starts at position 583 to 612 nt, and the cluster size is 29 nt. The free energy of interactions ranges from −110 to −121 kJ/mol. It is noteworthy that *POU3F3* is characterized by the presence of two clusters and many BSs when compared to other target genes. This suggests that this gene is more susceptible to regulation by miRNA.

The mRNA of the *SP8* gene contains five BSs for bta-miR-11976 and six BSs bta-miR-11975 located through three

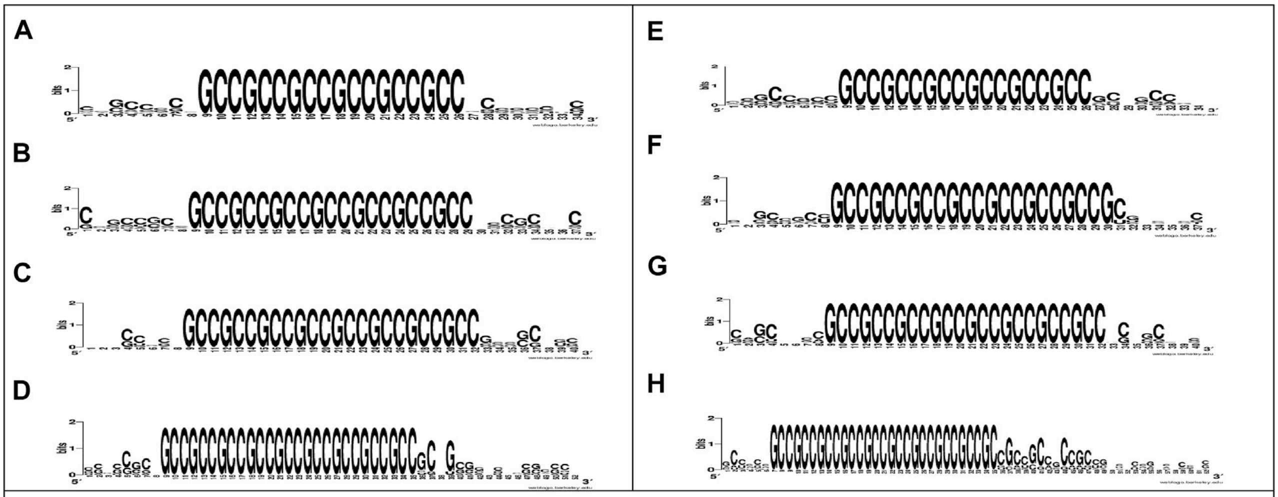
TABLE 3 | Nucleotide sequences of BSs of bta-miR-11975, bta-miR-11976, and bta-miR-2885 in CDS mRNA of human genes.

Gene	18 nt length cluster	
ATOH8	5′-CUCCCCACGCCGCCGCCGCCGCCUCCUGCGC-3′	PPPPPP
GABBR2	5′-CAGCCCGGGCCGCCGCCGCCGCCACCGCCGC-3′	PPPPPP
FAM117B	5′-CCCCCACGGCCGCCGCCGCCGCCGCUGCUGG-3′	PPPPPP
FOXG1	5′-CAGCAGCAGCCGCCGCCGCCGCCCCCGGCAC-3′	PPPPPP
FOXK1	5′-CCGCCCGGGCCGCCGCCGCCGCCACCGCCGC-3′	PPPPPP
GPR150	5′-UCGCCGCUGCCGCCGCCGCCGCCAACGUCCC-3′	PPPPPP
IRX5	5′-ACCCCGCGGCCGCCGCCGCCGCCUUCUCCUC-3′	PPPPPP
LOXL1	5′-CCCUACGUGCCGCCGCCGCCGCCCCCCGACG-3′	PPPPPP
LTBP1	5′-CUCAGACCGCCGCCGCCGCCGCCGGAGCCUG-3′	RPPPPP
MECP2	5′-GGAAAATGGCCGCCGCCGCCGCCGCGCCGAG-3′	PPPPPP
TMEM121	5′-AACUCGGUGCCGCCGCCGCCGCCGCUGCACG-3′	PPPPPP
TSPYL2	5′-CCGCCCCCGCCGCCGCCGCCGCCGCUCCUCC-3′	PPPPPP
TPRN	5′-CGCCCCCCGCCGCCGCCGCCGCCCGCGCCGC-3′	PPPPPP
ZCCHC2	5′-CGCCCCCCGCCGCCGCCGCCGCCCGCGGGCC-3′	PPPPPP
ZNF367	5′-GAGAACCCGCCGCCGCCGCCGCCCGUCAUCU-3′	PPPPPP
ZNF839	5′-AAGGCGCAGCCGCCGCCGCCGCCCCCCUUCG-3′	PPPPPP
FOXD1	5′-CGCAGCGCGCCGCCGCCGCCGCCUUCCACCC-3′	AAAAAA
CASZ1	5′-GCGAGGGCGCCGCCGCCGCCGCCGCAGCUGG-3′	AAAAAA
GPR88	5′-CCGGCUGCGCCGCCGCCGCCGCCUUCCCGGG-3′	AAAAAA
FBXL17	5′-UAUCCUCGGCCGCCGCCGCCGCCGCUGCCGC-3′	AAAAAA
HOXA2	5′-UUCUGCCGGCCGCCGCCGCCGCCACCGCCGC-3′	AAAAAA
HOXA13	5′-CCGCUGCAGCCGCCGCCGCCGCCGCGUCGUC-3′	AAAAAA
IRX2	5′-CGGCCGACGCCGCCGCCGCCGCCGGCUUCCC-3′	AAAAAA
IRX3	5′-UCUCUCCGGCCGCCGCCGCCGCCGCUCACAG-3′	AAAAAA
IRX4	5′-CAGCCACCGCCGCCGCCGCCGCCACCUCCCU-3′	AAAAAA
LCORL	5′-CCGCUGCUGCCGCCGCCGCCGCCGCUCAGUG-3′	AAAAAA
LHFPL3	5′-CCGCCGCUGCCGCCGCCGCCGCCGCGAUGCU-3′	AAAAAA
NANOS1	5′-GCGCGCCCGCCGCCGCCGCCGCCACCACCAC-3′	AAAAAA
POU3F3	5′-UGCCCCACGCCGCCGCCGCCGCCGCUGCCGC-3′	AAAAAA
SOX12	5′-AGGGGGCGGCCGCCGCCGCCGCCUCCCCGAC-3′	AAAAAA
SOX21	5′-CCGCCGCUGCCGCCGCCGCCGCCGCGGGCAG-3′	AAAAAA
SP8	5′-GCGCCGCAGCCGCCGCCGCCGCCGCAGCCGC-3′	AAAAAA
UNCX	5′-CUUCCAACGCCGCCGCCGCCGCCGCGGGGCU-3′	AAAAAA
	21 nt length cluster	
ARX	5′-CGGCCGCUGCCGCCGCCGCCGCCGCCUUCCCGAG-3′	AAAAAAA
DGKI	5′-CUCCUGCAGCCGCCGCCGCCGCCGCCAGCCCGCC-3′	AAAAAAA
GSG1L	5′-CCGCCCCCGCCGCCGCCGCCGCCGCCACCGCCUC-3′	AAAAAAA
JUND	5′-CGGCCGCUGCCGCCGCCGCCGCCGCCGGGGGCC-3′	AAAAAAA
SKOR2	5′-CCGGCCCCGCCGCCGCCGCCGCCGCCCCCGCCGC-3′	PPPPPPP
CEBPA	5′-CCUUACCAGCCGCCGCCGCCGCCGCCCUCGCACC-3′	PPPPPPP
CHD3	5′-CUCUUCCCGCCGCCGCCGCCGCCGCCACCGCUGC-3′	PPPPPPP
CTNND2	5′-GAGCCCGCGCCGCCGCCGCCGCCGCCGCGGAGC-3′	PPPPPPP
HCN2	5′-GCGCCGGGGCCGCCGCCGCCGCCGCCCGCGCCCC-3′	PPPPPPP
HTT	5′-CCGCCACCGCCGCCGCCGCCGCCGCCUCCUCAGC-3′	PPPPPPP
SOBP	5′-CCCGAGCAGCCGCCGCCGCCGCCGCCCGCGCCCC-3′	PPPPPPP
TGFBR3L	5′-CCUCUGACGCCGCCGCCGCCGCCGCCAUCGCGGU-3′	PPPPPPP
SLC24A3	5′-CGCGCGUCGCCGCCGCCGCCGCCGCCGGAGGGAC-3′	RRRRRRR
	24 nt length cluster	
CCDC177	5′-GCCCCGCGGCCGCCGCCGCCGCCGCCGCCGCGGCCUC-3′	AAAAAAAA
IRS2	5′-AGCCCAGGGCCGCCGCCGCCGCCGCCGCCGUGCCUUC-3′	AAAAAAAA
MEGF9	5′-UGUGCUGCGCCGCCGCCGCCGCCGCCGCCGUCGCCUC-3′	AAAAAAAA
SKIDA1	5′-ACCCGGCAGCCGCCGCCGCCGCCGCCGCCGCUGCUGC-3′	AAAAAAAA
ZIC3	5′-CAACCCACGCCGCCGCCGCCGCCGCCGCCGCUGCCUU-3′	AAAAAAAA
FMNL1	5′-GUGCCUCCGCCGCCGCCGCCGCCGCCGCCUCCCGGAG-3′	PPPPPPPP
GBX2	5′-GUAGUGCUGCCGCCGCCGCCGCCGCCGCCCGCGCUGC-3′	PPPPPPPP
MMP24	5′-GCGCCGGGGCCGCCGCCGCCGCCGCCGCCGGGCCAGG-3′	PPPPPPPP
TRIM67	5′-CUGGUGCAGCCGCCGCCGCCGCCGCCGCCCGCCGAGG-3′	PPPPPPPP
	27–36 nt length cluster	
DLX6	5′-CCTGCCCGGCCGCCGCCGCCGCCGCCGCCGCCGCAGCCGCCTCGCAGCA-3′	PPPPPPPPP
DMRTA2	5′-GCGTCGACGCCGCCGCCGCCGCCGCCGCCGCCGGGGGGCCTGGGCTGCC-3′	AAAAAAAAA
FOXF2	5′-CGCCGCCCGCCGCCGCCGCCGCCGCCGCCGCCGCCCCGGAGACCACCTCCTC-3′	AAAAAAAAA
IRF2BPL	5′-TAAGCGCTGCCGCCGCCGCCGCCGCCGCCGCCGCTGCGGTGGAACAGCG-3′	AAAAAAAAA
MNX1	5′-CGGCCGCTGCCGCCGCCGCCGCCGCCGCCGCCGCTGGGGGCCTGGCGCT-3′	AAAAAAAAA
NKX2-3	5′-CGGCCGCGGCCGCCGCCGCCGCCGCCGCCGCCGCAGCAGCGGCGGCCTA-3′	AAAAAAAAA

(Continued on following page)

TABLE 3 | (*Continued*) Nucleotide sequences of BSs of bta-miR-11975, bta-miR-11976, and bta-miR-2885 in CDS mRNA of human genes.

Gene	18 nt length cluster	
ZNF703	5′-TGGGCAGCGCCGCCGCCGCCGCCGCCGCCGCCGCCTCCTGCCATCTGCACCT-3′	AAAAAAAAA
ZSWIM6	5′-CCGCCGCTGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCGGGGGCCGGGGC-3′	AAAAAAAAAA
CASKIN1	5′-CCCCGCGAGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCGCCCCCCGCCCC-3′	AAAAAAAAAA
FOXE1	5′-GCTGCCCAGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCATCTTCCCAGG-3′	AAAAAAAAAAA
ZIC5	5′-GCCGGGCTGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCACCGCCCC-3′	PPPPPPPPPPP



Note: A, E - cluster with length 18 nt, B, F - cluster with length 21 nt, C, G – cluster with length 24 nt, D - cluster with length 27-36 nt, H - cluster with length 27-45 nt.

FIGURE 2 | WebLogo schemes of nucleotide sequences variability of CDS and 5′UTR mRNA having bta-miR-11976, bta-miR-11975, and bta-miR-2885 BSs clusters with lengths of 18, 21, 24, 27–36, and 27–45 nt.

nucleotides. The cluster starts at 553 nt and ends at 590 nt, with a length of 37 nt and $\Delta G/\Delta G_m$ of 92%. The free energy of interactions of two miRNAs with the mRNA varied from −117 to −123 kJ/mol. The mRNA of *LHFPL3* and *SOX21* genes have five BSs for bta-miR-11975 and three for bta-miR-11976 with a length of 32 and 37 nt, respectively.

Supplementary Table S4 shows interactions of mRNAs of 13 human genes with bta-miRNAs with a length of 21 nt. In all mRNAs, bta-miR-11975 and bta-miR-11976 are located through one nucleotide. The mRNAs of *ARX, CHD3, DGK1,* and *JUND* genes contain three BSs for bta-miR-11975 and two BSs for bta-miR-11976. In the mRNA of the *JUND* gene, bta-miR-11976 and bta-miR-2885 BSs are located through three nucleotides.

The mRNAs of *CTNND2* and *HTT* genes are characterized by clusters for three miRNAs BSs: bta-miR-11975, bta-miR-11976, and bta-miR-2885. Each of the mRNAs has four BSs for bta-miR-11975, three BSs for bta-miR-11976, and two BSs for bta-miR-2885 located through six nucleotides. The miR-2885 and bta-miR-11976 BSs are located with overlapping.

The mRNA of the *GSGL* gene has six BSs for bta-miR-11976 and bta-miR-11975. The mRNA of the *HCN2* gene has two BSs for bta-miR-11976 and bta-miR-11975 located from 109 to 136 nt, and the other from 152 to 175 nt, respectively. The cluster of BSs in mRNA of the *SKOR2* gene is one of the largest. The first cluster ranges from 851 to 882 nt, with a length of 30 nt. The second cluster starts at 2078 nt and ends at 3014 nt, with a length of 36 nt. In the second cluster, bta-miR-11976 and bta-miR-2885 BSs are located through six nucleotides. The mRNAs of *SLC24A3* and *TGFBR3L* genes have a cluster consisting of multiple BSs for bta-miR-11975 and bta-miR-11976. The mRNA of the *SOBP* gene has bta-miR-1975 BSs located through three nucleotides and overlapping with nucleotide sequences of bta-miR-11976 BSs. The bta-miRNA-11976 can interact with the mRNAs of *CHD3, CTNND2, GSG1L, HTT,* and *SKOR2* genes with a free energy of more than −120 kJ/mol.

Supplementary Table S5 shows characteristics of interactions of bta-miRNAs with mRNAs of nine human genes in the cluster with a length of 24 nt. The mRNAs of *CCDC177, GBX2,* and *TRIM67* genes have BSs for bta-miR-11975 and bta-miR-11976. The mRNA of the *MMP4* gene has two BSs for bta-miR-11975 and bta-miR-2885. The free energy of interaction value is equal to −127 kJ/mol.

The mRNAs of the *FNML1, IRS2, MEGF9,* and *ZIC3* genes have a cluster of multiple BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885, BSs are located through three nucleotides. The

TABLE 4 | Nucleotide sequences of BSs of bta-miR-11975, bta-miR-11976, and bta-miR-2885 in 5′UTR human mRNAs.

Gene	18 nt length cluster
ABCC1	5′-CUCCCUGCGCCGCCGCCGCCGCCGCCGCAGCGCU-3′
ASH1L	5′-CUGCUGCUGCCGCCGCCGCCGCCGCCGCUCCCGC-3′
BTF3L4	5′-CUGCUCCCGCCGCCGCCGCCGCCGCCGUCGUCUU-3′
C2CD4C	5′-ACUGCGCUGCCGCCGCCGCCGCCGCCCGCAUCGA-3′
CPT1A	5′-ACUCCACCGCCGCCGCCGCCGCCGCCGCUGCCGC-3′
EGLN1	5′-UCGCCGUCGCCGCCGCCGCCGCCGCCAUGGCCAA-3′
GRIN1	5′-UCCGCGGAGCCGCCGCCGCCGCCGCCGGGCCCUU-3′
GTF2E2	5′-CCGCCGCUGCCGCCGCCGCCGCCGCCACCGCCAG-3′
MAST1	5′-CUCCCCGCGCCGCCGCCGCCGCCGCCUCCGCCGC-3′
MEMO1	5′-CCGCUCCUGCCGCCGCCGCCGCCGCCUCCUCAUU-3′
MPRIP	5′-AGGCCUGCGCCGCCGCCGCCGCCGCCGUCGCCGC-3′
NOG	5′-GCGCGGACGCCGCCGCCGCCGCCGCCGCUGGAGU-3′
RIMS4	5′-AGCCGCCCGCCGCCGCCGCCGCCGCCGCGGCCGA-3′
RNF165	5′-CGCGCGCAGCCGCCGCCGCCGCCGCCGCGCGAGG-3′
RNF220	5′-CUGCCGCUGCCGCCGCCGCCGCCGCCGCUGCCUC-3′
SCAP	5′-CCCCCGUCGCCGCCGCCGCCGCCGCCGCAGCUUG-3′
SEPHS1	5′-GGGCCCCCGCCGCCGCCGCCGCCGCCGGGCGCGG-3′
SPEN	5′-CCGCCGCAGCCGCCGCCGCCGCCGCCCCGGCACC-3′
	21 nt length cluster
ABCD3	5′-GTAAGGUAGCCGCCGCCGCCGCCGCCGCCGCGUCCCC-3′
ANKH	5′-AACCUUCUGCCGCCGCCGCCGCCGCCGCCGUCCCTCC-3′
ANKRD13D	5′-GCCCCGCUGCCGCCGCCGCCGCCGCCGCCGCUACTGC-3′
C4orf19	5′-GGGACCCCGCCGCCGCCGCCGCCGCCGCCGUCUGGCC-3′
CA10	5′-UGGCUGCUGCCGCCGCCGCCGCCGCCGCCGCUGCTAG-3′
DISP2	5′-CCGCCACCGCCGCCGCCGCCGCCGCCGCCGCGGCTTC-3′
HS3ST4	5′-CGGGGGCUGCCGCCGCCGCCGCCGCCGCCGCGAGCCG-3′
JARID2	5′-GUGGUGCUGCCGCCGCCGCCGCCGCCGCCGCUGGAGT-3′
RGP1	5′-CAGCGGACGCCGCCGCCGCCGCCGCCGCCGCGUACCT-3′
UBE2R2	5′-GGCCCGGCGCCGCCGCCGCCGCCGCCGCCGCGAUGGC-3′
USP25	5′-GCGCCACCGCCGCCGCCGCCGCCGCCGCCGCGGGGGC-3′
	24 nt length cluster
AFF2	5′-CAGCCGCUGCCGCCGCCGCCGCCGCCGCCGCCGCGCCGCC-3′
CUL3	5′-GAGUCCGAGCCGCCGCCGCCGCCGCCGCCGCCCCCGCCGC-3′
FAM50A	5′-CGCCGCCCGCCGCCGCCGCCGCCGCCGCCGCCGCUGCCAU-3′
GSK3B	5′-GGGCUUGUGCCGCCGCCGCCGCCGCCGCCGCCCGGGCCAA-3′
MAP2K3	5′-CCGCAGUCGCCGCCGCCGCCGCCGCCGCCGCCGCUGCUCC-3′
MSI1	5′-CGCCGAGCGCCGCCGCCGCCGCCGCCGCCGCCGCUCCGCU-3′
MTHFD1L	5′-UCCUUCCCGCCGCCGCCGCCGCCGCCGCCGCCUGCUCCCC-3′
NCKAP1	5′-CCGGAGACGCCGCCGCCGCCGCCGCCGCCGCCACACCUAG-3′
RPRD2	5′-CCGCUCCCGCCGCCGCCGCCGCCGCCGCCGCCAGAGGAGC-3′
UBTF	5′-CAGCCACAGCCGCCGCCGCCGCCGCCGCCGCCACAGCAGC-3′
TCEA1	5′-GAGCCGGAGCCGCCGCCGCCGCCGCCGCCGCCGCGGGCUU-3′
THOC7	5′-CAGCUUGCGCCGCCGCCGCCGCCGCCGCCGCCGCGCACGC-3′
USP7	5′-GGCCGCCCGCCGCCGCCGCCGCCGCCGCCGCCCCGGCUCG-3′
ZNF219	5′-CGCCGCCCGCCGCCGCCGCCGCCGCCGCCGCCCGCUCCGC-3′
	27–48 nt length cluster
SBF1	5′-ACCUGGGCCGCCGCCGCCGCCGCCGCCGCCGCCGCGGAGCGAACCAGGGGUGUCCGGGGT-3′
SMAD9	5′-GCUGGGGCCGCCGCCGCCGCCGCCGCCGCCGCCGCUGCUGCAGCCGCUGUCUCGGUCCCC-3′
BCL11A	5′-CCGCCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCCGCCCCGCAGCCCACCAUGUCTCG-3′
WBP4	5′-GCUGCUGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCUGCUGCUGCCCACACGCUCCCG-3′
GNB2	5′-AUCCGCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCUCCGCCGCGGAGGAAGAC-3′
KIF3B	5′-GCCCCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCCGCUUUCGGCUCGGGCCT-3′
NDRG3	5′-CCUCUCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCUGCUGCUGCACTG-3′
BCL2L11	5′-GCCGCUGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCACUACCACCACT-3′
RHOT1	5′-GACUCGGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCACAGCC-3′

mRNA of the *SKIDA1* gene contains a cluster for three miRNAs BSs with a total length of 68 nt, starting at 2917 nt and ending 2985 nt.

Supplementary Table S6 shows the characteristics of mRNA of 11 human genes with bta-miRNAs in the clusters with a length of 27–36 nt. The mRNA of the *DLX6* gene has multiple BSs for bta-miR-11975 and bta-miR-11976. The mRNAs of 10 genes contains clusters for bta-miR-11975 and bta-miR-11976 and bta-miR-2885. In all mRNAs of 10 genes BSs of bta-miR-11976 and bta-miR-2885 are located with overlapping.

The mRNA of the *DMRTA2* gene contains four BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885. The mRNA of the *FOXF2* gene contains three BSs for these miRNAs. The mRNA of the *IRF2BPL* gene contains five BSs for bta-miR-11975 and bta-miR-11976 and three BSs for bta-miR-2885. The mRNA of the *MNX1* gene has a cluster of BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885. The cluster size is 45 nt and ΔG varied from −110 to −127 kJ/mol. The mRNA of the *NKX2-3* gene contains a single BS for bta-miR-11976, six BSs for bta-miR-11975, and three BSs for bta-miR-2885. The mRNA of the *ZNF703* gene contains a cluster for three miRNA BSs, the length of the cluster is 31 nt, extending from 1711 to 1742 nt. The mRNAs of *CASKIN1, FOXE1,* and *ZSWIM6* genes contain a cluster consisting of multiple BSs for bta-miR-11976, bta-miR-11975, and bta-miR-2885. The mRNA of the *ZIC5* gene is characterized by a single BS for bta-miR-2885 and bta-miR-11976 and two BSs for bta-miR-11976 located through six nucleotides. The second cluster has 10 BSs for bta-miR-11976, eight BSs for bta-miR-11976, and five BSs for bta-miR-2885, extending from 1467 nt to 1517 nt, the length is 50 nt, and ΔG value ranges from −110 to −127 kJ/mol.

Characteristics of bta-miR-11975, bta-miR-11976, and bta-miR-2885 interactions with the 5′UTR of mRNAs of human genes.

Clusters of BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885 in 5′UTR mRNA were identified in 52 human genes. **Table 4** shows nucleotide sequences of bta-miRNAs BSs in 5′UTR mRNA of human genes formed from GCC trinucleotides clusters with a length of 18 nt in mRNA of 18 genes, with a length of 21 nt in mRNA of 11 genes, with a length of 24 nt in mRNA of 14 genes, with a length of 27 nt in mRNA of two genes, with a length of 30 nt in mRNA of two genes, with a length of 36 nt in mRNA of two genes, and with a length of 39, 42, and 48 nt in the mRNA of one gene.

Supplementary Tables S7–S10 show characteristics of bta-miR-11975, bta-miR-11976, and bta-miR-2885 BSs in mRNAs of human genes located in a cluster. The *ABCC1* mRNA is predicted to be targeted by three cow miRNAs. Three BSs were identified for bta-miR-11975, two BSs for bta-miR-11976, and one BS for bta-miR-2885. The BSs of bta-miR-11976 and bta-miR-2885 overlap. The mRNA of the *ASH1L* gene has two BSs for bta-miR-11975. The mRNA of the *BTF3L4* gene has two BSs for bta-miR-11975 and bta-miR-11976 located through six nucleotides. The mRNA of the *C2CD4C* gene interacts with a single miRNA.

BSs of bta-miR-11975 and bta-miR-11976 were identified in mRNAs of *CPT1A*, *GTF2E2,* and *MAST1* genes. BSs of bta-miR-11975 and bta-miR-11976 are located through one nucleotide. Bta-miR-11976 has four multiple BSs, bta-miR-11975 has five, four and three BSs in mRNAs of target genes, respectively. The mRNAs of *EGLN1* and *RIMS4* have three BSs for bta-miR-11975 and two BSs for bta-miR-11976. The mRNAs of *GRIN1* and *MEMO1* have single BSs for bta-miR-11975 and bta-miR-11976.

BSs of bta-miR-11975 and bta-miR-11976 are identified in mRNAs of *MPRIP, RNF165,* and *RNF220* genes, located through one nucleotide. In mRNAs of *NOG, SCAP, SEPHS1,* and *SPEN* genes identified clusters for BSs of bta-miR-11975, bta-miR-11976 and bta-miR-2885. In mRNAs of *NOG, SCAP,* and *SPEN* genes BSs of bta-miR-11976 and bta-miR-2885 are overlapping.

Supplementary Table S8 shows interactions of bta-miRNAs BSs in a cluster with a length of 21 nt in mRNAs of 11 human genes. The mRNAs of *ABCD3, C4orf19, DISP2, HS3ST4, RGP1, UBE2R2,* and *USP25* genes have clusters of multiple BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885. BSs of bta-miR-11976 and bta-miR-2885 in clusters are located with overlapping through three nucleotides.

The mRNAs of *ANKH, ANKRD13D,* and *CA10* genes have BSs for bta-miR-11975 and for bta-miR-11976. The mRNA of *ANKRD13D* gene contains five BSs for bta-miR-11975 and four BSs for bta-miR-11976 located through three nucleotides. The mRNAs of *CA10* and *JARID* genes has three BSs for bta-miR-11975 and mRNA of *ANKH* gene has two BSs. The mRNAs of *ANKH*, *CA10,* and *JARID* genes have two BSs for bta-miR-11976. The BSs for bta-miR-11975 and bta-miR-11976 are located through one nucleotide. The free energy of interactions ΔG ranges from −110 to −127 kJ/mol.

The cluster in mRNA of the *UBE2R2* gene starts at 541 nt and ends at 569 nt, with a length of 28 nt. The total BSs length in the cluster is 281 nt, where the degree of compaction is 10.

BSs of mRNAs of 14 human genes contain a cluster with a length of 24 nt for bta-miRNAs (**Supplementary Table S9**). The free energy of interactions between miRNAs and the mRNAs of genes varied from −110 to −127 kJ/mol. The mRNA of the *AFF2* gene contains one cluster with two BSs for bta-miR-11975 and bta-miR-11976. The cluster starts at 14 nt and ends at 73 nt, with a length of 59 nt. The total length of BSs in the cluster is 724 nt. BSs of bta-miR-11975 and bta-miR-11976 are located through three nucleotides, starting at 101 nt and ending at 131 nt.

The mRNAs of *CUL3, FAM50A, MAP2K3, MSI1, MTHFD1L, NCKAP1, RPRD2, TCEA1, THOC7,* and *ZNF219* genes have a cluster of BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885. BSs of bta-miR-11976 and bta-miR-2885 are located with overlapping. The mRNA of *CUL3* has six BSs for bta-miR-11975 and bta-miR-11976. Bta-miR-2885 has a single BS located at 200 nt. The mRNAs of *FAM50A* and *MAP2K3* genes have five BSs for bta-miR-11975, four BSs for bta-miR-11976, and three BSs for bta-miR-2885. The mRNAs of *MSI1*and *RPRD2* genes contain four BSs for bta-miR-11975, three BSs for bta-miR-11976, and bta-miR-2885. The mRNAs of *MTHFD1L* and *NCKAP1* genes have four and three bta-miR-11975 BSs, respectively. The BSs of bta-miR-11976 and bta-miR-11975 are located with overlapping. The *ZNF219* mRNA has a cluster of BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885. The mRNAs of *TCEA1* and *THOC7* have three BSs for bta-miR-11975 and bta-miR-11976, and two BSs for bta-miR-2885.

The mRNAs of *GSK3B* and *UBTF* contain BSs for bta-miR-11975 and bta-miR-11976. The mRNA of the *USP7* gene has a single BS for bta-miR-11975 and bta-miR-11976 and cluster of BSs for three miRNAs. The cluster of BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885 is located from 94 to 121 nt, 27 nt.

Supplementary Table S10 shows characteristics of interactions of bta-miR-11975, bta-miR-11976, and bta-miR-2885 with nine genes mRNAs containing the BSs clusters with a length of 27–45 nt. The mRNAs of *SBF1* and *SMAD9* genes have four BSs for bta-miR-11975 and bta-miR-11976 and three BSs for bta-miR-2885. The mRNAs of

BCL11A, BCL2L11, and *GNB2* genes have multiple BSs for bta-miR-11975 and bta-miR-11976. The BSs clusters for three miRNAs were identified in mRNA of *WBP4* gene. The mRNA of the *KIF3B* gene is a target for three miRNAs and the total length of BSs is 94 nt. The mRNA of the *NDRG3* gene is a target for three miRNAs, two of which have nine BSs. The mRNA of *RHOT1* has a cluster of BSs for bta-miR-11975, bta-miR-11976, and bta-miR-2885. The cluster size is 47 nt and the total BS length in the cluster is 560 nt. The free energy value is higher than −127 kJ/mol for the interactions of bta-miR-11976 with the mRNAs of all nine human genes.

The miRNAs with Multiple Targets Have Been Shown to Have Several Effects on Various Diseases in Humans, Including Cancer

Supplementary Table S11 shows data for genes targeted by bta-miRNAs that may be involved in the development of various diseases: six genes associated with breast cancer; *CTNND2* – liposarcoma; *CEBPA* - myeloid leukemia; *DGKI* - gastric cancer; *FBXL17* - medulloblastoma; *FMNL1*, *FOXD1,* and *SKOR2* – cell carcinoma; *FOXE1* - thyroid cancer; *FOXK1, IRS2* cholangiocarcinoma; *IRX3* - hepatocellular carcinoma; *MNX1, SOX12, TRIM67,* and *ZNF839* - colorectal cancer; *HOXA2* and *HOXA13* - prostate cancer; *LOXL1* - pancreatic cancer; *DLX6, SOBP,* and *UNCX* - lung cancer; *LHFPL3* - melanoma; *LTBP1* - glioblastoma; *SOX21* - cervical cancer; *SP8* - hepatoblastoma; *SLC24A3* – meningiomas; *TGFBR3L* - neuroendocrine tumors; *TPRN*–myeloma; *ZCCHC2* - retinoblastoma; *ZNF703* - thyroid carcinoma and neurodegenerative diseases: *ARX* - interneuron development; *CCDC177, CHD3, TSPYL2, FOXG1,* and *ZSWIM6* - neurodevelopmental syndrome; *DMRTA2* - cortical development; *GABBR2* - autism; *GPR88* and *POU3F3* - Parkinson's disease; *HCN2* - epilepsy; *HTT* - Huntington's disease; *IRF2BPL* - neurological phenotypes; *LCORL* and *MMP24* - Alzheimer's disease*; MECP2* - neurodevelopmental disorder; *CASZ1, GBX2, IRX4, IRX5, NKX2-3, JUND,* and *ZIC3* - cardiac diseases; *GPR150* - liver disease. This list of candidate genes for various human diseases indicates the great potential of bta-miRNAs in the regulation of the pathogenesis of the listed diseases. Increasing evidence sheds light on the potential implications of ex-miRs identified in human biofluids derived from non-human species to cross-kingdom gene regulation and human disease pathogenesis (*Perge et al., 2017*; *Sanchita et al., 2018*; *Zhao et al., 2018*). The biogenesis and function of such exogenous miRNAs are evidently health related (*Arnold et al., 2012*; *Izumi et al., 2012*; *Liu et al., 2012*; *Baier et al., 2014*).

DISCUSSION

In this study, bioinformatics analyses were employed to predict exogenous miRNAs that target human mRNAs. The miRNAs have been predicted as highly transportable candidates; several of them have identical sequences with their homologs in human (**Table 2**). It was found experimentally and *in silico* that hsa-miR-127-5p, hsa-miR-136, hsa-miR-431, hsa-miR-432, and hsa-miR-433 bind to mRNA of the human *RTL1* gene in identical positions with bta-miR-127, bta-miR-136, bta-miR-431, bta-miR-432, and bta-miR-433 (*Davis et al., 2005*; *Yurikova et al., 2019*).

Humans have a number of identical miRNAs for bta-miR-135a, bta-miR-136, bta-miR-432, bta-miR-127, bta-miR-431, bta-miR-433, bta-miR-1282, and bta-miR196a. The identical sequence may indicate a higher probability that the exogenous miRNA will regulate human genes after transportation into circulation. Moreover, miRNAs have been identified to be conserved across mammalian and non-mammalian species and this highlights that homologous miRNAs may share similar functional roles in common pathways in the evolutionary mechanism of distinct species (*Liu et al., 2012; van Herwijnen et al., 2018*). Human breast milk contains hsa-miR-136 and hsa-miR-135a, hsa-miR-432, hsa-miR-433 in colostrum, hsa-miR-196a and hsa-miR-431 in milk and colostrum (*Weber et al., 2010*). Previously 245 miRNAs have been found in cow milk by Chen et al. (*Chen et al., 2010*). The bta-miR-135a, bta-miR-136, bta-miR-127-3p, bta-miR-196a, and bta-miR-432 were found in bovine milk and colostrum (**Table 2**) (*Chen et al., 2010*). These experiments confirm that these five miRNAs are expressed in such mature milk-specific miRNAs and colostrum-specific miRNAs. Overall, 95% of the miRNA expressed in human milk was also expressed in cow milk. Milk-derived miRNAs can be transferred from bovine milk to humans and regulate gene expression in target tissues and cells. Milk derived miRNAs can enter normal and malignant cells and can regulate biological signals (*Golan-Gerstl et al., 2017; Zempleni et al., 2018*). Most milk miRNAs can affect human protein-coding genes. The bta-miR-151-5p, bta-miR-151-3p, bta-miRNA-320 have BSs in 11 genes, while bta-miRNA-345-5p, bta-miRNA-614, bta-miRNA-1296b and bta-miRNA-149 have BSs in 12, 14, 15 and 26 genes, respectively. The bta-miR-574-5p from cow's milk had from one to 25 repeating BSs in mRNAs of 209 human genes, indicating this miRNA's significant biological role. Clearly, experiments are needed next to determine possibly exosomal transport pathways. Therefore, future studies will need to address if milk-derived exosomal miRNAs can be transferred from bovine milk to humans and regulate gene expression in humans. Alternatively, such exosomes may act directly on cells such as MECs (mammary epithelial cells): Kelleher et al. could recently show that miRNAs are involved in MEC signaling and intriguingly, also appear to be required for healthy breast functions (*Kelleher et al., 2019*).

This study identified three miRNAs: bta-miR-11975, bta-miR-11976 and bta-miR-2885, in a cluster that targets multiple human-associated genes. The bta-miR-11976 and bta-miR-11975 recently discovered in a bovine (*Morenikeji et al., 2020*). The data obtained in our work on the possible effect of these miRNAs on human genes that cause many diseases provide a basis for further investigation of the effect of these exogenous miRNAs on human genes. Thus, from the data shown in **Supplementary Table S11**, it can be seen that almost half of the target genes are transcription factors, which makes it difficult to establish the effect of miRNAs on these target genes, since transcription factors can affect the expression of genes that control various functions. The miRNAs can affect oncogenes or onco-suppressors, showing the opposite effect on oncogenesis. For example, an increased level of miR-222 expression suppressed the effect of a breast cancer tumor suppressor, and a decreased miR-222 level increased the expression level of a tumor suppressor (*Said et al., 2021*). Increased expression of miRNA-200a-3p promotes the growth and development of gastric cancer tumors by suppressing the tumor suppressor (*Li et al., 2021*). Conversely, miR-345-5p plays the role of a tumor suppressor in lung adenocarcinoma cells, suppressing the oncogene (*Zhou Y. et al., 2021*).

Conducting *in silico* research leads to the question of how reliable the results are. The MirTarget program has been used in several studies and predicted the results of earlier and later studies. The high predictive advantage of the MirTarget program lies in the fact that it finds BSs for the entire nucleotide sequence of miRNA and determines important quantitative characteristics of the interaction of miRNA with mRNA. This approach in determining the properties of the interaction of exosomal miRNA with mRNA made it possible to reveal fundamentally new properties of miRNA forming in mRNA clusters to reduce the length of the 5′UTR and CDS regions (*Kondybayeva et al., 2018*; *Aisina et al., 2019*; *Mukushkina et al., 2020*; *Akimniyazova et al., 2021*; *Kamenova et al., 2021*). Studying the interaction of exogenous miRNAs with mRNA genes of humans *in silico* makes it possible to significantly reduce the material and time costs for determining effective associations of miRNAs and their target genes. A fundamentally important result of this work is to establish the possibility of the effect of exogenous miRNAs on the expression of human genes, including candidate genes for socially significant human diseases.

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

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

FUNDING

We gratefully acknowledge support by the European Union (RISE project 645648) for support (SL and MM).

ACKNOWLEDGMENTS

We are thankful to Pyrkova A. Yu for making calculations.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.705350/full#supplementary-material

REFERENCES

Aarts, J., Boleij, A., Pieters, B., Feitsma, A. L., van Neerven, R., Ten Klooster, J. P., et al. (2021). Flood Control: How Milk-Derived Extracellular Vesicles Can Help to Improve the Intestinal Barrier Function and Break the Gut-Joint Axis in Rheumatoid Arthritis. *Front. Immunol.* 12, 703277. doi:10.3389/fimmu.2021.703277

Aisina, D., Niyazova, R., Atambayeva, S., and Ivashchenko, A. (2019). Prediction of Clusters of miRNA Binding Sites in mRNA Candidate Genes of Breast Cancer Subtypes. *PeerJ* 7, e8049. doi:10.7717/peerj.8049

Akimniyazova, A., Pyrkova, A., Uversky, V., and Ivashchenko, A. (2021). Predicting Associations of miRNAs and Candidate Gastric Cancer Genes for Nanomedicine. *Nanomaterials* 11 (3), 691. doi:10.3390/nano11030691

Arnold, C. N., Pirie, E., Dosenovic, P., McInerney, G. M., Xia, Y., Wang, N., et al. (2012). A Forward Genetic Screen Reveals Roles for Nfkbid, Zeb1, and Ruvbl2 in Humoral Immunity. *Proc. Natl. Acad. Sci. USA* 109, 12286–12293. doi:10.1073/pnas.1209134109

Atambayeva, S., Niyazova, R., Ivashchenko, A., Pyrkova, A., Pinsky, I., Akimniyazova, A., et al. (2017). The Binding Sites of miR-619-5p in the mRNAs of Human and Orthologous Genes. *BMC Genomics* 18, 428. doi:10.1186/s12864-017-3811-6

Auerbach, A., Vyas, G., Li, A., Halushka, M., and Witwer, K. (2016). Uptake of Dietary Milk miRNAs by Adult Humans: A Validation Study. *F1000Research* 5, 721. doi:10.12688/f1000research.8548.1

Bahrami, A., Aledavood, A., Anvari, K., Hassanian, S., Maftouh, M., Yaghobzade, A., et al. (2018). The Prognostic and Therapeutic Application of microRNAs in Breast Cancer: Tissue and circulatigmicroRNAs. *J. Cel Physiol* 233, 774–786. doi:10.1002/jcp.25813

Baier, S. R., Nguyen, C., Xie, F., Wood, J. R., and Zempleni, J. (2014). MicroRNAs Are Absorbed in Biologically Meaningful Amounts from Nutritionally Relevant Doses of Cow Milk and Affect Gene Expression in Peripheral Blood Mononuclear Cells, HEK-293 Kidney Cell Cultures, and Mouse Livers. *J. Nutr.* 144, 1495–1500. doi:10.3945/jn.114.196436

Bari, A., Orazova, S., and Ivashchenko, A. (2013). miR156- and miR171-Binding Sites in the Protein-Coding Sequences of Several Plant Genes. *Biomed. Res. Int.* 2013, 1–7. doi:10.1155/2013/307145

Bari, A., Sagaidak, I., Pinskii, I., Orazova, S., and Ivashchenko, A. (2014). Binding of miR396 to mRNA of Genes Encoding Growth Regulating Transcription Factors of Plants. *Russ. J. Plant Physiol.* 61, 807–810. doi:10.1134/S1021443714050033

Bayraktar, R., Van Roosbroeck, K., and Calin, G. A. (2017). Cell-to-cell Communication: MicroRNAs as Hormones. *Mol. Oncol.* 11, 1673–1686. doi:10.1002/1878-0261.12144

Benmoussa, A., and Provost, P. (2019). Milk microRNAs in Health and Disease. *Compr. Rev. Food Sci. Food Saf.* 18, 703–722. doi:10.1111/1541-4337.12424

Billa, P., Faulconnier, Y., Ye, T., Bourdon, C., Pires, J., and Leroux, C. (2021). Nutrigenomic Analyses Reveal miRNAs and mRNAs Affected by Feed Restriction in the Mammary Gland of Midlactation Dairy Cows. *PLoS One* 16 (4), e0248680. doi:10.1371/journal.pone.0248680

Bryniarski, K., Ptak, W., Martin, E., Nazimek, K., Szczepanik, M., Sanak, M., et al. (2015). Free Extracellular miRNA Functionally Targets Cells by Transfecting Exosomes from Their Companion Cells. *PLoS One* 10 (4), e0122991. doi:10.1371/journal.pone.0122991

Chen, X., Gao, C., Li, H., Huang, L., Sun, Q., and Dong, Y. (2010). Identification and Characterization of microRNAs in Raw Milk During Different Periods of Lactation, Commercial Fluid, and Powdered Milk Products. *Cell Res* 20, 1128–1137. doi:10.1038/cr.2010.80

Dai, X., Zhuang, Z., and Zhao, P. (2011). Computational Analysis of miRNA Targets in Plants: Current Status and Challenges. *Brief. Bioinformatics* 12, 115–121. doi:10.1093/bib/bbq065

Davis, E., Caiment, F., Tordoir, X., Cavaillé, J., Ferguson-Smith, A., Cockett, N., et al. (2005). RNAi-Mediated Allelic Trans-interaction at the Imprinted Rtl1/Peg11 Locus. *Curr. Biol.* 15, 743–749. doi:10.1016/j.cub.2005.02.060

Dever, J. T., Kemp, M. Q., Thompson, A. L., Keller, H. G. K., Waksmonski, J. C., Scholl, C. D., et al. (2015). Survival and Diversity of Human Homologous Dietary MicroRNAs in Conventionally Cooked Top Sirloin and Dried Bovine Tissue Extracts. *PLOS ONE* 10 (9), e0138275. doi:10.1371/journal.pone.0138275

Dickinson, B., Zhang, Y., Petrick, J. S., Heck, G., Ivashuta, S., and Marshall, W. S. (2013). Lack of Detectable Oral Bioavailability of Plant MicroRNAs After Feeding in Mice. *Nat. Biotechnol.* 31, 965–967. doi:10.1038/nbt.2737

Diomaiuto, E., Principe, V., De Luca, A., Laperuta, F., Alterisio, C., and Di Loria, A. (2021). Exosomes in Dogs and Cats: An Innovative Approach to Neoplastic and Non-neoplastic Diseases. *Pharmaceuticals* 14, 766. doi:10.3390/ph14080766

Fabian, M. R., and Sonenberg, N. (2012). The Mechanics of miRNA-Mediated Gene Silencing: A Look under the Hood of miRISK. *Nat. Struct. Mol. Biol.* 19, 586. doi:10.1038/nsmb.2296

Friedman, R. A., and Honig, B. (1995). A Free Energy Analysis of Nucleic Acid Base Stacking in Aqueous Solution. *Biophys. J.* 69 (4), 1528–1535. doi:10.1016/S0006-3495(95)80023-8

Gao, H. N., Ren, F. Z., Wen, P. C., Xie, L. X., Wang, R., Yang, Z. N., et al. (2021). Yak Milk-Derived Exosomal microRNAs Regulate Intestinal Epithelial Cells on Proliferation in Hypoxic Environment. *J. Dairy Sci.* 104, 1291–1303. doi:10.3168/jds.2020-19063

Garg, A., and Heinemann, U. (2018). A Novel Form of RNA Double Helix Based on G·U and C·A+ Wobble Base Pairing. *RNA* 24, 209–218. doi:10.1261/rna.064048.117

Golan-Gerstl, R., Lavi-Moshayoff, V., Elbaum, Y. S., and Leshkowits, D. (2017). Characterization and Biological Function of Milk Derived miRNAs. *Mol. Nutr. Food Res.* 61, 1. doi:10.1002/mnfr.201700009

Gu, Y., Li, M., Wang, T., Liang, Y., Zhong, Z., and Wang, X. (2012). Lactation-related MicroRNA Expression Profiles of Porcine Breast Milk Exosomes. *PLoS One* 7, e43691. doi:10.1371/journal.pone.0043691

Huntzinger, E., and Izaurralde, E. (2011). Gene Silencing by MicroRNAs: Contributions of Translational Repression and mRNA Decay. *Nat. Rev. Genet.* 12, 99. doi:10.1038/nrg2936

Ipsaro, J. J., and Joshua-Tor, L. (2015). From Guide to Target: Molecular Insights into Eukaryotic RNA-interference Machinery. *Nat. Struct. Mol. Biol.* 22, 20. doi:10.1038/nsmb.2931

Ivashchenko, A., Berillo, O., Pyrkova, A., Niyazova, R., and Atambayeva, S. (2014). MiR-3960 Binding Sites with mRNA of Human Genes. *Bioinformation* 10, 423–427. doi:10.6026/97320630010423

Ivashchenko, A. T., Pyrkova, A. Y., Niyazova, R. Y., Alybayeva, A., and Baskakov, K. (2016). Prediction of miRNA Minding Sites in mRNA. *Bioinformation* 12, 237–240. doi:10.6026/97320630012237

Izumi, H., Kosaka, N., Shimizu, T., Sekine, K., Ochiya, T., and Takase, M. (2012). Bovine Milk Contains MicroRNA and Messenger RNA that are Stable Under Degradative Conditions. *J. Dairy Sci.* 95, 4831–4841. doi:10.3168/jds.2012-5489

Jiang, X., You, L., Zhang, Z., Cui, X., Zhong, H., Sun, X., et al. (2021). Biological Properties of Milk-derived Extracellular Vesicles and their Physiological Functions in Infant. *Front Cell Dev Biol.* 9, 693534. doi:10.3389/fcell.2021.693534

Jonas, S., and Izaurralde, E. (2015). Towards a Molecular Understanding of MicroRNA-mediated Gene Silencing. *Nat. Rev. Genet.* 16, 421. doi:10.1038/nrg3965

Kamenova, S., Aralbayeva, A., Kondybayeva, A., Akimniyazova, A., Pyrkova, A., and Ivashchenko, A. (2021). Evolutionary Changes in the Interaction of MiRNA with mRNA of Candidate Genes for Parkinson's Disease. *Front. Genet.* 12, 647288. doi:10.3389/fgene.2021.647288

Kelleher, S. L., Gagnon, A., Rivera, O. C., Hicks, S. D., Carney, M. C., and Samina, A. (2019). Milk-derived miRNA Profiles Elucidate Molecular Pathways that Underlie Breast Dysfunction in Women with Common Genetic Variants in SLC30A2. *Sientific Rep.* 9, 1. doi:10.1038/s41598-019-48987-4

Kim, K. U., Kim, W. H., Jeong, C. H., Yi, D. Y., and Min, H. (2020). More Than Nutrition: Therapeutic Potential of Breast Milk-Derived Exosomes in Cancer. *Int. J. Mol. Sci.* 21, 7327. doi:10.3390/ijms21197327

Komine-Aizawa, S., Ito, S., Aizawa, S., Namiki, T., and Hayakawa, S. (2020). Cow Milk Exosomes Activate NK Cells and γδT Cells in Human PBMCs In Vitro. *Immunological Med.* 43, 161–170. doi:10.1080/25785826.2020.1791400

Kondybayeva, A. M., Akimniyazova, A. N., Kamenova, S. U., and Ivashchenko, A. T. (2018). The Characteristics of miRNA Binding Sites in mRNA of ZFHX3 Gene and its Orthologs. *Vavilov J. Genet. Breed.* 22, 438–444. doi:10.18699/VJ18.380

Kool, E. T. (2001). Hydrogen Bonding, Base Stacking, and Steric Effects in DNA Replication. *Annu. Rev. Biophys. Biomol.Struct.* 30, 1–22. doi:10.1146/annurev.biophys.30.1.1

Kosaka, N., Izumi, H., Sekine, K., and Ochiya, T. (2010). MicroRNA as a New Immune-regulatory Agent in Breast Milk. *Silence* 1 (1), 7. doi:10.1186/1758-907X-1-7

Kusuma, R. J., Manca, S., Frieme, T., Sukreet, S., Nguyen, C., and Zempleni, J. (2016). Human Vascular Endothelial Cells Transport Foreign Exosomes from Cow's Milk by Endocytosis. *Am. J. Physiol. Cel. Physiol.* 310, 800–807. doi:10.1152/ajpcell.00169.2015

Laubier, J., Castille, J., Le Guillou, S., and Le Provost, F. (2015). No Effect of an Elevated miR-30b Level in Mouse Milk on its Level in Pup Tissues. *RNA Biol.* 12, 26–29. doi:10.1080/15476286.2015.1017212

Leontis, N. B., Stombaugh, J., and Westhof, E. (2002). The Non-watson-crick Base Pairs and Their Associated Isostericity Matrices. *Nucleic Acids Res.* 30, 3497–3531. doi:10.1093/nar/gkf481

Leroux, C., Chervet, M. L., and German, J. B. (2021). Perspective: Milk microRNAs as Important Players in Infant Physiology and Development. *Adv. Nutr.* 12, 1625. doi:10.1093/advances/nmab059

Li, Z., Wang, Y., Liu, S., Li, W., Wang, Z., Jia, Z., et al. (2021). MiR-200a-3p Promotes Gastric Cancer Progression by Targeting DLC-1. *J. Mol. Histol.* 22, 4697. doi:10.1007/s10735-021-10037-7

Lin, D., Chen, T., Xie, M., Li, M., Zeng, B., Sun, R., et al. (2020). Oral Administration of Bovine and Porcine Milk Exosome Alter miRNAs Profiles in Piglet Serum. *Sci. Rep.* 10, 6983. doi:10.1038/s41598-020-63485-8

Liu, R., Ma, X., Xu, L., Wang, D., Jiang, X., Zhu, W., et al. (2012). Differential microRNA Expression in Peripheral Blood Mononuclear Cells from Graves' Disease Patients. *J. Clin. Endocrinol. Metab.* 97 (6), 968–972. doi:10.1210/jc.2011-2982

Lowry, D., Paul, H., and Reimer, R. (2021). Impact of Maternal Obesity and Prebiotic Supplementation on Select Maternal Milk microRNA Levels and Correlation with Offspring Outcomes. *Br. J. Nutr.* 1, 1. doi:10.1017/S0007114521001197

Lukasik, A., and Zielenkiewicz, P. (2014). In Silico identification of Plant miRNAs in Mammalian Breast Milk Exosomes a Small Step Forward? *PLoSOne* 1 (6), e99963. doi:10.1371/journal.pone.0099963

Manca, S., Upadhyaya, B., Mutai, E., Desaulniers, A. T., Cederberg, R. A., White, B. R., et al. (2018). Milk Exosomes Are Bioavailable and Distinct microRNA Cargos Have Unique Tissue Distribution Patterns. *Sci. Rep.* 8, 11321. doi:10.1038/s41598-018-29780-1

Mar-Aguilar, F., Arreola-Triana, A., Mata-Cardona, D., Gonzalez-Villasana, V., Rodríguez-Padilla, C., and Reséndez-Pérez, D. (2020). Evidence of Transfer of miRNAs from the Diet to the Blood Still Inconclusive. *PeerJ* 8, e9567. doi:10.7717/peerj.9567

Marsh, S. R., Pridham, K. J., Jourdan, J., and Gourdie, R. G. (2021). Novel Protocols for Scalable Production of High Quality Purified Small Extracellular Vesicles from Bovine Milk. *Nanotheranostics* 5, 488–498. doi:10.7150/ntno.62213

McNeill, E. M., and Hirschi, K. D. (2020). Roles of Regulatory RNAs in Nutritional Control. *Annu. Rev. Nutr.* 40, 77–104. doi:10.1146/annurev-nutr-122319-035633

Melnik, B. C., Stremmel, W., Weiskirchen, R., John, S. M., and Schmitz, G. (2021). Exosome-Derived MicroRNAs of Human Milk and Their Effects on Infant Health and Development. *Biomolecules* 11, 851. doi:10.3390/biom11060851

Melnik, B. C. (2021b). Synergistic Effects of Milk-Derived Exosomes and Galactose on α-Synuclein Pathology in Parkinson's Disease and Type 2 Diabetes Mellitus. *Int. J. Mol. Sci.* 22, 1059. doi:10.3390/ijms22031059

Melnik, B. (2021a). Lifetime Impact of Cow's Milk on Overactivation of mTORC1: from Fetal to Childhood Overgrowth, Acne, Diabetes, Cancers, and Neurodegeneration. *Biomolecules* 11 (3), 404. doi:10.3390/biom11030404

Miao, C., Wang, X., Zhou, W., and Huang, J. (2021). The Emerging Roles of Exosomes in Autoimmune Diseases, with Special Emphasis on microRNAs in Exosomes. *Pharmacol. Res.* 169, 105680. doi:10.1016/j.phrs.2021.105680

Morenikeji, O. B., Wallace, M., Strutton, E., Bernard, K., Yip, E., and Thomas, B. N. (2020). Integrative Network Analysis of Predicted miRNA-Targets Regulating Expression of Immune Response Genes in Bovine Coronavirus Infection. *Front. Genet.* 11, 584392. doi:10.3389/fgene.2020.584392

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

Munagala, R., Aqil, F., Jeyabalan, J., and Gupta, R. C. (2016). Bovine Milk-Derived Exosomes for Drug Delivery. *Cancer Lett.* 371, 48–61. doi:10.1016/j.canlet.2015.10.020

Perge, P., Nagy, Z., Decmann, A., Igaz, I., and Igaz, P. (2017). Potential Relevance of microRNAs in Inter-species Epigenetic Communication, and Implications for Disease Pathogenesis. *RNA Biol.* 14, 391–401. doi:10.1080/15476286.2016.1251001

Rajagopal, C., and Harikumar, K. B. (2018). The Origin and Functions of Exosomes in Cancer. *Front. Oncol.* 8, 1–13. doi:10.3389/fonc.2018.00066

Rakhmetullina, A., Pyrkova, A., Aisina, D., and Ivashchenko, A. (2020). In Silico Prediction of Human Genes as Potential Targets for Rice miRNAs. *Comput. Biol. Chem.* 87, 107305. doi:10.1016/j.compbiolchem.2020.107305

Reif, S., Elbaum-Shiff, Y., Koroukhov, N., Shilo, I., Musseri, M., and Golan-Gerstl, R. (2020). Cow and Human Milk-Derived Exosomes Ameliorate Colitis in DSS Murine Model. *Nutrients* 12, 2589. doi:10.3390/nu12092589

Said, M. N., Hanafy, S. M., Helal, A., Fawzy, A., Allam, R. M., and Shafik, N. F. (2021). Regulation of CDK Inhibitor P27 by microRNA 222 in Breast Cancer Patients. *Exp. Mol. Pathol.* 6, 104718. doi:10.1016/j.yexmp.2021.104718

Sanchita, R., Trivedi, R., Asif, M. H., and Trivedi, P. K. (2018). Dietary Plant miRNAs as an Augmented Therapy: Cross-Kingdom Gene Regulation. *RNA Biol.* 15, 1433–1439. doi:10.1080/15476286.2018.1551693

Shah, K. B., Chernausek, S. D., Garman, L. D., Pezant, N. P., Plows, J. F., Kharoud, H. K., et al. (2021). Human Milk Exosomal MicroRNA: Associations with Maternal Overweight/Obesity and Infant Body Composition at 1 Month of Life. *Nutrients* 13, 1091. doi:10.3390/nu13041091

Shu, J., Chiang, K., Zempleni, J., and Cui, J. (2015). Computational Characterization of Exogenous MicroRNAs that Can Be Transferred into Human Circulation. *PLoS One* 10, e0140587. doi:10.1371/journal.pone.0140587

Title, A. C., Denzler, R., and Stoffel, M. (2015). Uptake and Function Studies of Maternal Milk-Derived MicroRNAs. *J. Biol. Chem.* 290, 23680–23691. doi:10.1074/jbc.M115.676734

van Herwijnen, M. J. C., Driedonks, T. A. P., Snoek, B. L., Kroon, A. M. T., Kleinjan, M., Jorritsma, R., et al. (2018). Abundantly Present miRNAs in Milk-Derived Extracellular Vesicles Are Conserved Between Mammals. *Front. Nutr.* 5, 81. doi:10.3389/fnut.2018.00081

Wang, J., Li, Z., Liu, B., Chen, G., Shao, N., Ying, X., et al. (2016a). Systematic Study of Cis-Antisense miRNAs in Animal Species Reveals miR-3661 to Target PPP2CA in Human Cells. *RNA* 22, 87–95. doi:10.1261/rna.052894.115

Wang, K., Li, H., Yuan, Y., Etheridge, A., Zhou, Y., Huang, D., et al. (2012). The Complex Exogenous RNA Spectra in Human Plasma: An Interface with Human Gut Biota? *PLoS One* 7, e51009. doi:10.1371/journal.pone.0051009

Wang, L., Wang, X., Shi, Z., Shen, L., Zhang, J., and Zhang, J. (2021a). Bovine Milk Exosomes Attenuate the Alteration of Purine Metabolism and Energy Status in IEC-6 Cells Induced by Hydrogen Peroxide. *Food Chem.* 350, 129142. doi:10.1016/j.foodchem.2021.129142

Wang, X., Fan, Y., He, Y., Han, Z., Gong, Z., Peng, Y., et al. (2021b). Integrative Analysis of miRNA and mRNA Expression Profiles in Mammary Glands of Holstein Cows Artificially Infected with staphylococcus Aureus. *Pathogens* 10, 506. doi:10.3390/pathogens10050506

Wang, Y., Li, L., Tang, Sh., Liu, J., Zhang, H., Zhi, H., et al. (2016b). Combined Small RNA and Degradome Sequencing to Identify miRNAs and Their Targets in Response to Drought in Foxtail Millet. *BMC Genomics* 17, 57. doi:10.1186/s12863-016-0364-7

Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., Huang, K. H., and Lee, M. J. (2010). The MicroRNA Spectrum in 12 Body Fluids. *Clin. Chem.* 56, 1733–1741. doi:10.1373/clinchem.2010.147405

Wehbe, Z., and Kreydiyyeh, S. (2021). Cow's Milk May Be Delivering Potentially Harmful Undetected Cargoes to Humans. Is it Time to Reconsider Dairy Recommendations? *Nutr. Rev.* 1, nuab046. doi:10.1093/nutrit/nuab046

Xia, L., Zhao, Z., Yu, X., Lu, C., Jiang, P., Yu, H., et al. (2021). Integrative Analysis of miRNAs and mRNAs Revealed Regulation of Lipid Metabolism in Dairy Cattle. *Funct. Integr. Genomics* 21, 393. doi:10.1007/s10142-021-00786-9

Xiao, J., Feng, S., Wang, X., Long, K., Luo, Y., Wang, Y., et al. (2018). Identification of Exosome-like Nanoparticle-Derived microRNAs from 11 Edible Fruits and Vegetables. *PeerJ* 6, e5186. doi:10.7717/peerj.5186

Yim, N., Ryu, S. W., Choi, K., Lee, K. R., Lee, S., Choi, H., et al. (2016). Exosome Engineering for Efficient Intracellular Delivery of Soluble Proteins Using Optically Reversible Protein-Protein Interaction Module. *Nat. Commun.* 7, 12277. doi:10.1038/ncomms12277

Yun, B., Kim, Y., Park, D. J., and Oh, S. (2021). Comparative Analysis of Dietary Exosome-Derived microRNAs from Human, Bovine and Caprine Colostrum and Mature Milk. *J. Anim. Sci. Technol.* 63, 593–602. doi:10.5187/jast.2021.e39

Yurikova, O. Y., Aisina, D. E., Niyazova, R. E., Atambayeva, S. A., Labeit, S., and Ivashchenko, A. T. (2019). The Interactions of miRNA-5p and miRNA-3p with the mRNAs of Ortolologous Genes. *Mol. Biol.* 53 (4), 692–704. doi:10.1134/s0026893319040174

Zempleni, J., Sukreet, S., Zhou, F., Wu, D., and Mutai, E. (2018). Milk-derived Exosomes and Metabolic Regulation. *Annu. Rev. Anim. Biosci.* 7, 245. doi:10.1146/annurev-animal-020518-115300

Zeng, B., Chen, T., Luo, J., Xie, M., Wei, L., Xi, Q., et al. (2020). Exploration of Long Non-coding RNAs and Circular RNAs in Porcine Milk Exosomes. *Front. Genet.* 11, 652. doi:10.3389/fgene.2020.00652

Zeng, B., Chen, T., Luo, J. Y., Zhang, L., Xi, Q. Y., Jiang, Q. Y., et al. (2021). Biological Characteristics and Roles of Noncoding RNAs in Milk-Derived Extracellular Vesicles. *Adv. Nutr.* 12, 1006–1019. doi:10.1093/advances/nmaa124

Zhang, L., Hou, D., Chen, X., Li, D., Zhu, L., Zhang, Y., et al. (2012a). Exogenous Plant MIR168a Specifically Targets Mammalian LDLRAP1: Evidence of Cross-Kingdom Regulation by microRNA. *Cel Res* 22, 107. doi:10.1038/cr.2011.158

Zhang, Y., Wiggins, B. E., Lawrence, C., Petrick, J., Ivashuta, S., and Heck, G. (2012b). Analysis of Plant-Derived miRNAs in Animal Small RNA Datasets. *BMC Genomics* 13, 381. doi:10.1186/1471-2164-13-381

Zhang, Y., Xu, Q., Hou, J., Huang, G., Zhao, S., and Zheng, N. (2021). Loss of Bioactive MicroRNAs in Cow's Milk by Ultra-High-Temperature Treatment but not by Pasteurization Treatment. *J. Sci. Food Agric.* [Epub ahead of print]. doi:10.1002/jsfa.11607

Zhao, Q., Liu, Y., Zhang, N., Hu, M., Zhang, H., Joshi, T., et al. (2018). Evidence for Plant-Derived xenomiRs Based on a Large-Scale Analysis of Public Small RNA Sequencing Data from Human Samples. *PLoS One* 13, e0187519. doi:10.1371/journal.pone.0187519

Zhou, Q., Gui, S., Zhang, P., and Wang, M. (2021a). Upregulation of miR-345-5p Suppresses Cell Growth of Lung Adenocarcinoma by Regulating Ras Homolog Family Member A (RhoA) and Rho/Rho Associated Protein Kinase (Rho/ROCK) Pathway. *Chin. Med. J.* 134 (21), 2619–2628. doi:10.1097/CM9.0000000000001804

Zhou, Q., Li, M., Wang, X., Li, Q., Wang, T., Zhu, Q., et al. (2012). Immune-related microRNAs Are Abundant in Breast Milk Exosomes. *Int. J. Biol. Sci.* 8, 118–123. doi:10.7150/ijbs.8.118

Zhou, Y., Yu, Z., Wang, X., Chen, W., Liu, Y., Zhang, Y., et al. (2021b). Exosomal circRNAs Contribute to Intestinal Development via the VEGF Signalling Pathway in Human Term and Preterm Colostrum. *Aging* 13, 11218–11233. doi:10.18632/aging.202806

Zhou, Z., Li, X., Liu, J., Dong, L., Chen, Q., Liu, J., et al. (2015). Honeysuckle-encoded Atypical microRNA2911 Directly Targets Influenza Viruses. *Cel Res* 25, 39–49. doi:10.1038/cr.2014.130

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