



COMPUTATIONAL GENOMICS AND STRUCTURAL BIOINFORMATICS IN PERSONALIZED MEDICINES

EDITED BY: George Priya Doss C., Thirumal Kumar D. and Balu Kamaraj
PUBLISHED IN: Frontiers in Medicine



frontiers

Frontiers eBook Copyright Statement

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

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

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

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

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

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

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

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

ISSN 1664-8714

ISBN 978-2-88976-224-8

DOI 10.3389/978-2-88976-224-8

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

COMPUTATIONAL GENOMICS AND STRUCTURAL BIOINFORMATICS IN PERSONALIZED MEDICINES

Topic Editors:

George Priya Doss C., VIT University, India

Thirumal Kumar D., Meenakshi Academy of Higher Education and Research, India

Balu Kamaraj, Imam Abdulrahman Bin Faisal University, Saudi Arabia

Citation: Doss, C. G. P., Kumar, D. T., Kamaraj, B., eds. (2022). Computational Genomics and Structural Bioinformatics in Personalized Medicines. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-224-8

Table of Contents

- 05** *Phytochemical Moieties From Indian Traditional Medicine for Targeting Dual Hotspots on SARS-CoV-2 Spike Protein: An Integrative in-silico Approach*
V. Umashankar, Sanjay H. Deshpande, Harsha V. Hegde, Ishwar Singh and Debprasad Chattopadhyay
- 25** *Potential Prognostic Immune Biomarkers of Overall Survival in Ovarian Cancer Through Comprehensive Bioinformatics Analysis: A Novel Artificial Intelligence Survival Prediction System*
Tingshan He, Liwen Huang, Jing Li, Peng Wang and Zhiqiao Zhang
- 41** *Bioactive Molecules of Tea as Potential Inhibitors for RNA-Dependent RNA Polymerase of SARS-CoV-2*
Vijay Kumar Bhardwaj, Rahul Singh, Jatin Sharma, Vidya Rajendran, Rituraj Purohit and Sanjay Kumar
- 52** *Saudi Familial Hypercholesterolemia Patients With Rare LDLR Stop Gain Variant Showed Variable Clinical Phenotype and Resistance to Multiple Drug Regimen*
Zuhier Ahmed Awan, Omran M. Rashidi, Bandar Ali Al-Shehri, Kaiser Jamil, Ramu Elango, Jumana Y. Al-Aama, Robert A. Hegele, Babajan Banaganapalli and Noor A. Shaik
- 65** *rs12537 Is a Novel Susceptibility SNP Associated With Estrogen Receptor Positive Breast Cancer in Chinese Han Population*
Jingkai Xu, Guozheng Li, Mengyun Chen, Wenjing Li, Yaxing Wu, Xuejun Zhang, Yong Cui and Bo Zhang
- 71** *Whole-Genome Sequencing Reveals Exonic Variation of ASIC5 Gene Results in Recurrent Pregnancy Loss*
Nourah H. Al Qahtani, Sayed AbdulAzeez, Noor B. Almandil, Norah Fahad Alhur, Hind Saleh Alsuwat, Hatoon Ahmed Al Taifi, Ahlam A. Al-Ghamdi, B. Rabindran Jermy, Mohamed Abouelhoda, Shazia Subhani, Lubna Al Asoom and J. Francis Borgio
- 79** *Identification of Key Genes and Pathways in Persistent Hyperplastic Primary Vitreous of the Eye Using Bioinformatic Analysis*
Derin M. Thomas, Chitra Kannabiran and D. Balasubramanian
- 97** *Multi-Trait Genomic Risk Stratification for Type 2 Diabetes*
Palle Duun Rohde, Mette Nyegaard, Mads Kjolby and Peter Sørensen
- 108** *Novel MYO1D Missense Variant Identified Through Whole Exome Sequencing and Computational Biology Analysis Expands the Spectrum of Causal Genes of Laterality Defects*
Rabab Said Alsafwani, Khalidah K. Nasser, Thoraia Shinawi, Babajan Banaganapalli, Hanan Abdelhalim ElSokary, Zhaheer F. Zaher, Noor Ahmad Shaik, Gaser Abdelmohsen, Jumana Yousuf Al-Aama, Adam J. Shapiro, Osman O. Al-Radi, Ramu Elango and Turki Alahmadi
- 121** *Development and Verification of an Autophagy-Related lncRNA Signature to Predict Clinical Outcomes and Therapeutic Responses in Ovarian Cancer*
Yan Li, Juan Wang, Fang Wang, Chengzhen Gao, Yuanyuan Cao and Jianhua Wang

- 137 Alpha-Synuclein Aggregation in Parkinson's Disease**
E. Srinivasan, G. Chandrasekhar, P. Chandrasekar, K. Anbarasu, A. S. Vickram, Rohini Karunakaran, R. Rajasekaran and P. S. Srikumar
- 151 Identification of Cross-Pathway Connections via Protein-Protein Interactions Linked to Altered States of Metabolic Enzymes in Cervical Cancer**
Krishna Kumar, Sarpita Bose and Saikat Chakrabarti
- 171 Clinical Significance and Prognostic Value of Human Soluble Resistance-Related Calcium-Binding Protein: A Pan-Cancer Analysis**
Jinguo Zhang, Jian Chen, Benjie Shan, Lin Lin, Jie Dong, Qingqing Sun, Qiong Zhou and Xinghua Han
- 190 Identification and Structure Prediction of Human Septin-4 as a Biomarker for Diagnosis of Asthenozoospermic Infertile Patients—Critical Finding Toward Personalized Medicine**
A. S. Vickram, K. Anbarasu, Palanivelu Jeyanthi, G. Gulothungan, R. Nanmaran, S. Thanigaivel, T. B. Sridharan and Karunakaran Rohini
- 199 Development and Validation of a Novel Prognosis Prediction Model for Patients With Stomach Adenocarcinoma**
Tong Wang, Weiwei Wen, Hongfei Liu, Jun Zhang, Xiaofeng Zhang and Yu Wang
- 215 Immune Checkpoint Gene Expression Profiling Identifies Programmed Cell Death Ligand-1 Centered Immunologic Subtypes of Oral and Squamous Cell Carcinoma With Favorable Survival**
Yang Yu, Huiwen Tang, Debora Franceschi, Prabhakar Mujagond, Aneesha Acharya, Yupei Deng, Bernd Lethaus, Vuk Savkovic, Rüdiger Zimmerer, Dirk Ziebolz, Simin Li and Gerhard Schmalz
- 235 Transcriptome-Based Molecular Networks Uncovered Interplay Between Druggable Genes of CD8⁺ T Cells and Changes in Immune Cell Landscape in Patients With Pulmonary Tuberculosis**
Faten Ahmad Alsulaimany, Nidal M. Omer Zabermawi, Haifa Almukadi, Snijesh V. Parambath, Preetha Jayasheela Shetty, Venkatesh Vaidyanathan, Ramu Elango, Babajan Babanaganapalli and Noor Ahmad Shaik
- 247 Increased CDCA2 Level Was Related to Poor Prognosis in Hepatocellular Carcinoma and Associated With Up-Regulation of Immune Checkpoints**
Mengying Tang, Mingchu Liao, Xiaohong Ai and Guicheng He



Phytochemical Moieties From Indian Traditional Medicine for Targeting Dual Hotspots on SARS-CoV-2 Spike Protein: An Integrative *in-silico* Approach

OPEN ACCESS

Edited by:

Balu Kamaraj,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia

Reviewed by:

Rituraj Purohit,
Institute of Himalayan Bioresource
Technology (CSIR), India
Gnanendra Shanmugam,
Yeungnam University, South Korea

*Correspondence:

V. Umashankar
umashankar.v@icmr.gov.in
Debprasad Chattopadhyay
debprasadc@gmail.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 26 February 2021

Accepted: 31 March 2021

Published: 07 May 2021

Citation:

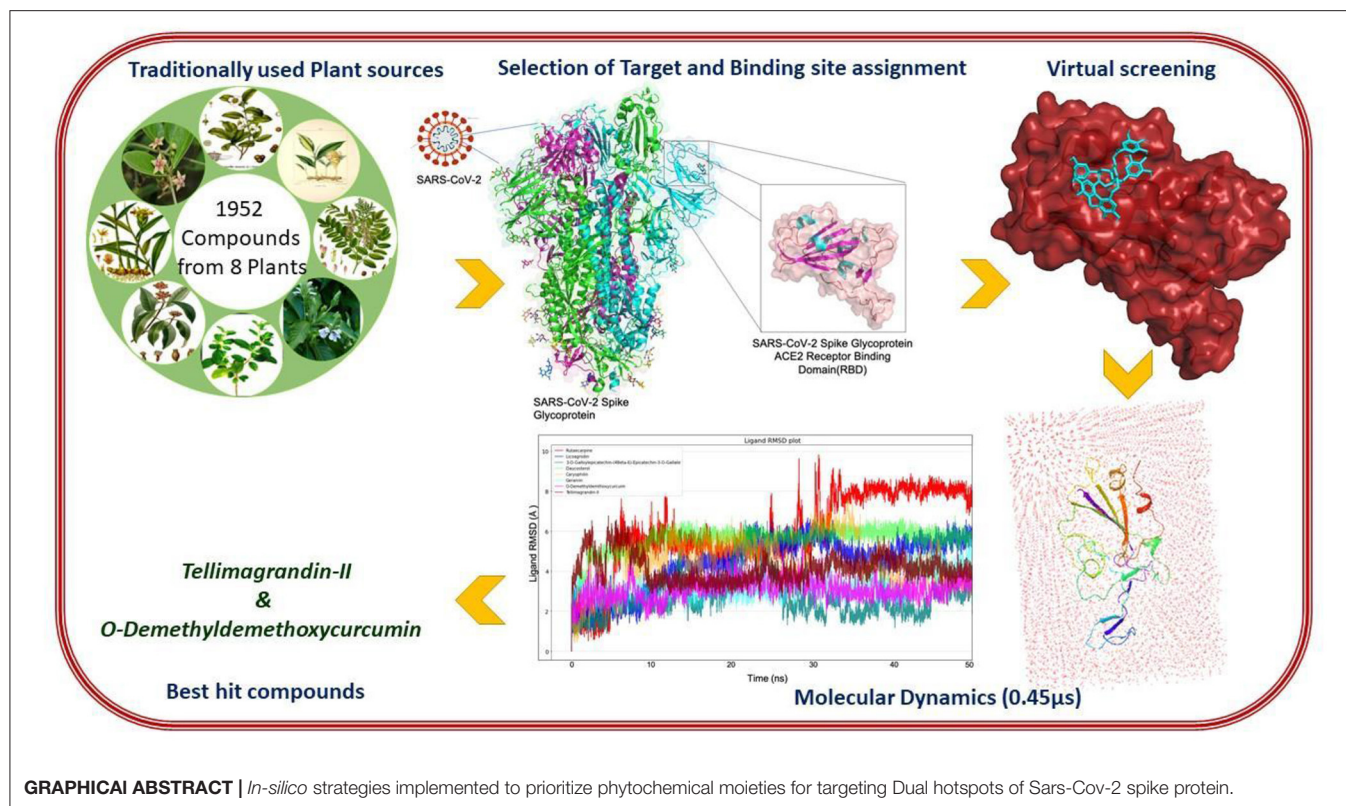
Umashankar V, Deshpande SH,
Hegde HV, Singh I and
Chattopadhyay D (2021)
Phytochemical Moieties From Indian
Traditional Medicine for Targeting Dual
Hotspots on SARS-CoV-2 Spike
Protein: An Integrative *in-silico*
Approach. *Front. Med.* 8:672629.
doi: 10.3389/fmed.2021.672629

V. Umashankar^{*†}, Sanjay H. Deshpande[†], Harsha V. Hegde, Ishwar Singh and
Debprasad Chattopadhyay^{*}

ICMR-National Institute of Traditional Medicine, Indian Council of Medical Research, Department of Health Research
(Government of India), Belagavi, India

SARS-CoV-2 infection across the world has led to immense turbulence in the treatment modality, thus demanding a swift drug discovery process. Spike protein of SARS-CoV-2 binds to ACE2 receptor of human to initiate host invasion. Plethora of studies demonstrate the inhibition of Spike-ACE2 interactions to impair infection. The ancient Indian traditional medicine has been of great interest of Virologists worldwide to decipher potential antivirals. Hence, in this study, phytochemicals (1,952 compounds) from eight potential medicinal plants used in Indian traditional medicine were meticulously collated, based on their usage in respiratory disorders, along with immunomodulatory and anti-viral potential from contemporary literature. Further, these compounds were virtually screened against Receptor Binding Domain (RBD) of Spike protein. The potential compounds from each plant were prioritized based on the binding affinity, key hotspot interactions at ACE2 binding region and glycosylation sites. Finally, the potential hits in complex with spike protein were subjected to Molecular Dynamics simulation (450 ns), to infer the stability of complex formation. Among the compounds screened, Tellimagrandin-II (binding energy of -8.2 kcal/mol and binding free energy of -32.08 kcal/mol) from *Syzygium aromaticum* L. and O-Demethyl-demethoxy-curcumin (binding energy of -8.0 kcal/mol and binding free energy of -12.48 kcal/mol) from *Curcuma longa* L. were found to be highly potential due to their higher binding affinity and significant binding free energy (MM-PBSA), along with favorable ADMET properties and stable intermolecular interactions with hotspots (including the ASN343 glycosylation site). The proposed hits are highly promising, as these are resultant of stringent *in silico* checkpoints, traditionally used, and are documented through contemporary literature. Hence, could serve as promising leads for subsequent experimental validations.

Keywords: COVID-19, traditional medicine, docking, molecular dynamics, drug design



INTRODUCTION

A new respiratory infectious disease was reported in Wuhan, Hubei Province of China, around December 2019 (1, 2). The outbreak at the initial stage was linked to a seafood market with a possibility of animal transmission. In due course of time, human to human infection began that spread across the globe, and the disease was called as COVID-19 (Coronavirus disease 19). The newly emerged virus was named as SARS Corona virus-2 (SARS-CoV-2; **Figure 1**), based on the etiology and symptoms, which is closely associated with SARS-CoV identified in the year 2002 in China (3). The epidemic of COVID-19 has been declared as a pandemic by the WHO on 30th January 2020, which affected the population across the globe to the worst possible extent (4). The SARS-CoV-2 infection has spread across the continents, as of March 25, 2021 a total of 125,429,834 cases with a mortality of 2,756,742 and recoveries of 101,293,629 are reported, based on the registered cases (5). Currently, quarantine, isolation, use of masks, physical distancing, washing of hands with soap water and symptomatic treatment protocol is being strictly followed to manage the disease, as there is no drug available till date to selectively target this virus. These data mainly highlight the extent of spread across the globe. Hence, finding prophylactic or therapeutic agents becomes important and essential.

The virus enters the human cells by the regulation of spike (S) glycoprotein (1,273 amino acids long; **Figure 2**) which is cleaved into 2 main units, namely, S1 (13–685 aa) and S2 (686–1,273 aa).

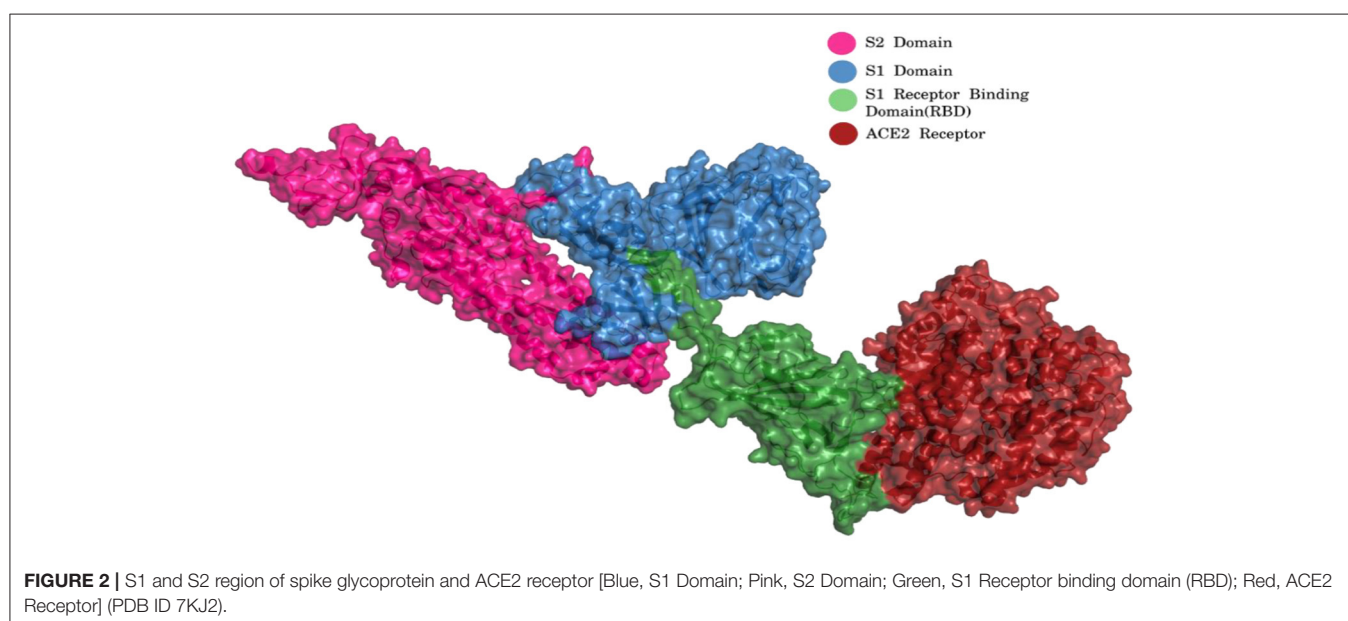
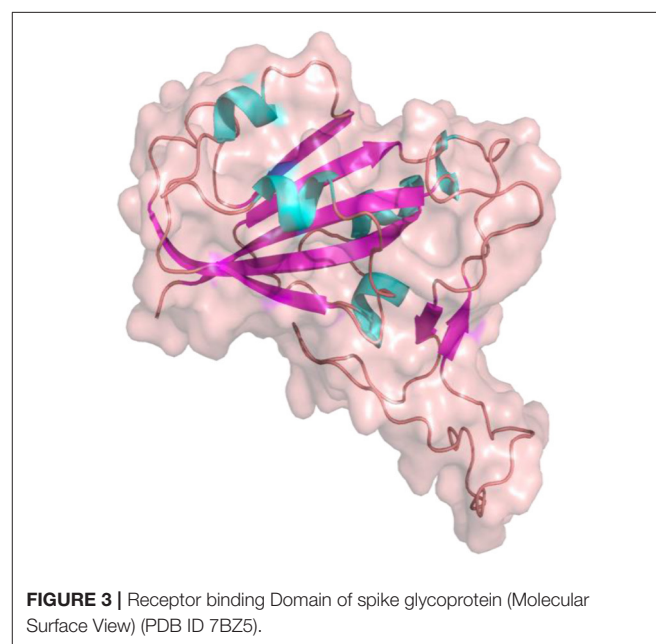
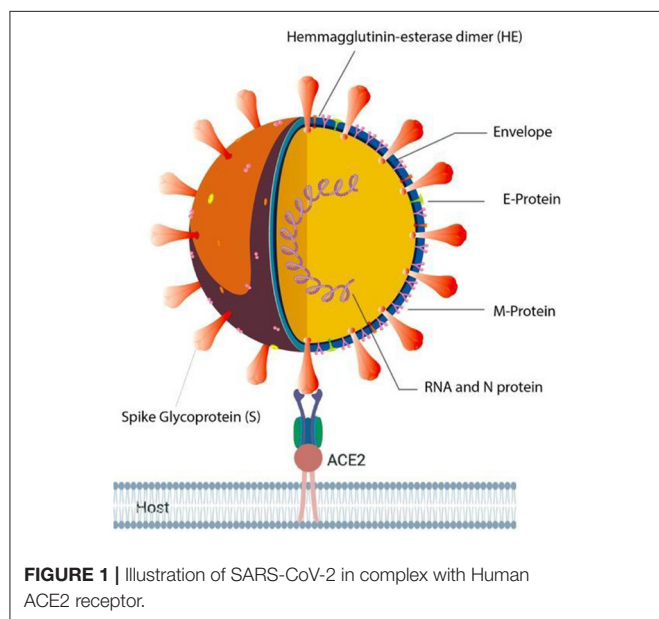
The S1 and S2 domains are present in individual monomers of the spike protein trimer (**Supplementary Figure 1**). This 3D structure of protein complex has been recently elucidated using Cryo-Electron microscopy (PDB id: 6VSB) (rendered using PyMol) (6, 7). The surface unit 1 (S1) helps in the strong attachment of the spike protein to human cell receptors. The cleavage of S1/S2 helps in the entry of viral particles and the fusion of the viral capsid with the host cell membrane is guided by the S2 subunit (8). Several studies established that angiotensin-converting enzyme 2 (ACE2) receptor of the host cell is the mediator that facilitates viral entry (9, 10). Spike protein S1 domain is further divided into multiple regions that are involved in binding to host receptors. In Spike protein, 319–541 aa region (S1) is known as receptor binding domain (RBD) [PDB id: 7BZ5; **Figure 3**; (11)]

and 437–508 as receptor binding motif (RBM), which binds to the ACE2 receptor. Other earlier studies have also compared the receptor-binding domain (RBD) of spike protein of both SARS-CoV and SARS-CoV-2 having high residue conservation, indicating that only a small change makes the SARS-CoV-2 binding to the ACE2 receptor different from the other coronaviruses (12, 13). Recent genome sequencing study also showed the spread of mutant form (D614G) of spike protein to have the potential for enhanced ACE2 binding (14). Glycosylation sites of SARS-CoV-2 are reported to modulate the host immune response, and also proposed to be the potential target for future mutations (15). The glycosylation process in coronaviruses mainly occurs to camouflage the immunogenic

process in the host. Targeting glycosylation sites can also help in the early and rapid immune response to neutralize the virion (16). The potential hotspot residues of spike protein namely, THR323, SER325, ASN331, and ASN343 are reported to be involved in glycosylation. Among these residues, ASN343 spans the RBD region of the spike protein (17). Thus, targeting these sites could be an efficient mode for combating SARS-CoV2 infections.

Drug development is a tedious and highly time-consuming process usually takes years to get the newly developed drug for the treatment (18–20). Thus, one of the most preferred method is to find a suitable drug through drug repurposing, as it saves both time and financial resources (21). The most common source

of drug repurposing is existing drugs or molecules of natural origin, and the scientists across the globe feel that the use of compounds from natural sources is one of the best and sought about way to move forward (22), even to find a drug for COVID-19 (23, 24). Computational approaches in identifying potential hits have gained momentum due to their cost effective and time saving efficiencies in drug discovery process (25). Moreover, these approaches have been well-utilized for mining potential chemical moieties from diverse phytochemical libraries (24, 26, 27). In recent times, studies on traditional medicine have climbed to newer heights across the globe due to its immense potential, easy availability, time-tested safety profile, and wide range of



pharmacological actions. The increase in studies is also due to the implementation of technologies to understand the structure and function of phytochemicals from nature (28). With the continuous advancement in the field of computer science, many drugs have been approved from natural sources through computer-aided drug design, like Ponatinib (FDA approval: 2012), Dasatinib (FDA approval: 2006), and Imatinib (FDA approval: 2001). Applications of *in silico* approach helps to calculate and analyze the combinations of compounds and targets as highly accurate, hence gaining more and more importance in the field of drug discovery by saving time and money (29). Several studies revealed that ethnomedicinal “phytophores” belonging to different classes, based on their structure activity relationship (SAR), showed effectiveness in curtailing viral replication in diverse viral infections. The antiviral compounds usually interfere in host-virus interaction points like viral entry process from adhesion/attachment to fusion and penetration, inhibit enzymatic activity, and or block one or more steps of the viral life cycle including replication to release. The scientific evidence and the traditional usage of antiviral plant extracts clearly portray the potential of natural compounds in modulating viral infection (30).

India is a country with rich biodiversity and long history of use of traditional medicine (TM) with a vast knowledge base of useful medicinal plants through the ages. Indian Ayurveda is one of the oldest systems of medicine of the world existing since the world's first civilizations and Vedic era. The main resource of TM is the generation-old time-tested knowledge base of plant-based formulations and wisdom of different communities, known as “ethnomedicine” (31), using different parts of plants from roots to leaves, bark fruits, and seeds (*New Look to Phytomedicine*, 2019). A wide range of plants used in ethnomedicinal practices were shown to be highly effective in the management of diverse viral infections by inhibiting either the viral life cycle or the host-virus interactions (32).

Viral infections are increasing across the globe mainly due to increased anthropogenic activities like land-use change, increased human-animal interaction and lack of proper healthcare infrastructure. Hence, the discovery of antivirals from natural sources, mainly traditionally used medicinal plants have gained importance. Since ages, plants have been used as a source of therapeutics in diverse ethnomedicinal practices. Many ethnomedicinal plant extracts and phytochemicals are known to modulate host immune responses (33) and may exert antimicrobial and antiviral effects (34). A variety of plant compounds including alkaloids, coumarins, essential oils, flavonoids, polyphenols, phytosterols, proteins, peptides, saponins, and tannins play diverse roles in the human system. Consistent progress has been made in the development of nature-based antiviral drugs in recent years, as natural products like plant extracts and phytochemicals used in TM are novel and broad-based chemical entity that may serve as a potential sources of antiviral drugs (35). The ever-increasing drug resistance, frequent microbial mutations with increased emerging and re-emerging outbreaks of viruses necessitate the development of easily available cost-effective antimicrobials and antivirals for better treatments. Hence, traditional medicines are the hope and

source for novel agents to manage viral diseases (30). A whole range of viral diseases caused by the Dengue, Human herpes viruses, HIV, Rabies, and Severe acute respiratory syndrome (SARS) needs potential therapeutics; while using modern tools the vast knowledge of ethnomedicinal practices can be identified and validated for antiviral applications (36). Thus, a surge of research is being observed in research institutes and universities, particularly the countries rich in TM.

In recent times, several phytochemicals having antiviral potential have been identified with their molecular mechanism of action. Spiroketalenol, isolated from the rhizome extract of *Tanacetum vulgare* L., was found to inhibit HSV-1 and HSV-2 by blocking the virus entry, and inhibit the activity of viral glycoproteins (37). Another compound Samarangenin B from the roots of *Limonium sinense* found to suppress the replication of HSV-1 by inhibiting the expression of HSV-1 immediate early (IE) or α - gene (38). Harmaline (HM), a dihydro-pyrido-indole, from the ethnomedicinal herb *Ophiorrhiza nicobarica* is reported to exhibit anti-HSV activity by suppressing the viral IE gene synthesis through epigenetic blocking of LSD-1 with a different mode of action than the gold standard antiviral Acyclovir (39). Further, it was reported that the ursolic acid isolated from *Mallotus peltatus* (Geist) Muell. Arg. dose-dependently inhibits the plaque formation of both HSV-1 and HSV-2 at 10 μ g/ml within 2–5 h post-infection (40). Moreover, *Odina woder* Roxb, a herb used in folklore medicine confer therapeutic effects on the skin infections caused by HSV (41). *Pterocarya stenoptera* traditionally used in the treatment of viral diseases is another potential plant with antiviral activity and its isolated compound Pterocarnin A was shown to inhibit HSV-2, by blocking the penetration of the virion into the host cells (42). Complementarily, many bio-active compounds from plants were shown to have immunomodulatory activities by triggering anti-inflammatory responses, which in turn helps in the control of viral infection (34). Earlier studies have revealed that modulation of NF- κ B signaling mediated anti-inflammatory response triggered by *Pedilanthus tithymaloides* L. confer a higher level of anti-HSV activity (43, 44). Further, ultrasound-induced Gallic acid based gold nanoparticles can inhibit HSV infection with EC₅₀ of 32.3 and 38.6 μ M against HSV-1 and HSV-2, respectively (45). While oleo-gum resin-extract and β -Boswellic acid of *Boswellia serrata* inhibit HSV-1 infection through modulation of NF- κ B and p38 MAP kinase signaling (46).

Ethnomedicinal literature claims the broad-spectrum antiviral activity of diverse medicinal plant extracts and phytochemicals, as the majority of those antiviral herbs contain flavones, polyphenols, and alkaloids. Due to the rapid emergence of new highly infectious viruses as well as re-emergence of drug-resistance, and difficult-to-treat infections along with the concurrent availability of advanced technological tools, the exploration of antiviral activity of medicinal plants has acquired momentum. In the current scenario of COVID-19, traditional Chinese medicine (TCM) was included in the guideline for the treatment, which claimed to be efficacious in several cases (47). Similarly, many of the Indian ethnomedicinal plants are reported to ameliorate the symptoms related to COVID-19, with antiviral

activities (24). The hits based on such observations can provide the edge for development of drugs to manage/treat COVID-19. Thus, in this study we have performed a meticulous analysis of documented antiviral properties of selected traditionally used Indian medicinal plants. This resulted in eight potential plants to be probed for phytochemical moieties that could target COVID-19 effectively. The rationale on selection of plants is discussed in detail as follows.

MATERIALS AND METHODS

Selection of Plants and the Rationale

Tylophora indica Burm F. Merrill (Asclepiadaceae)

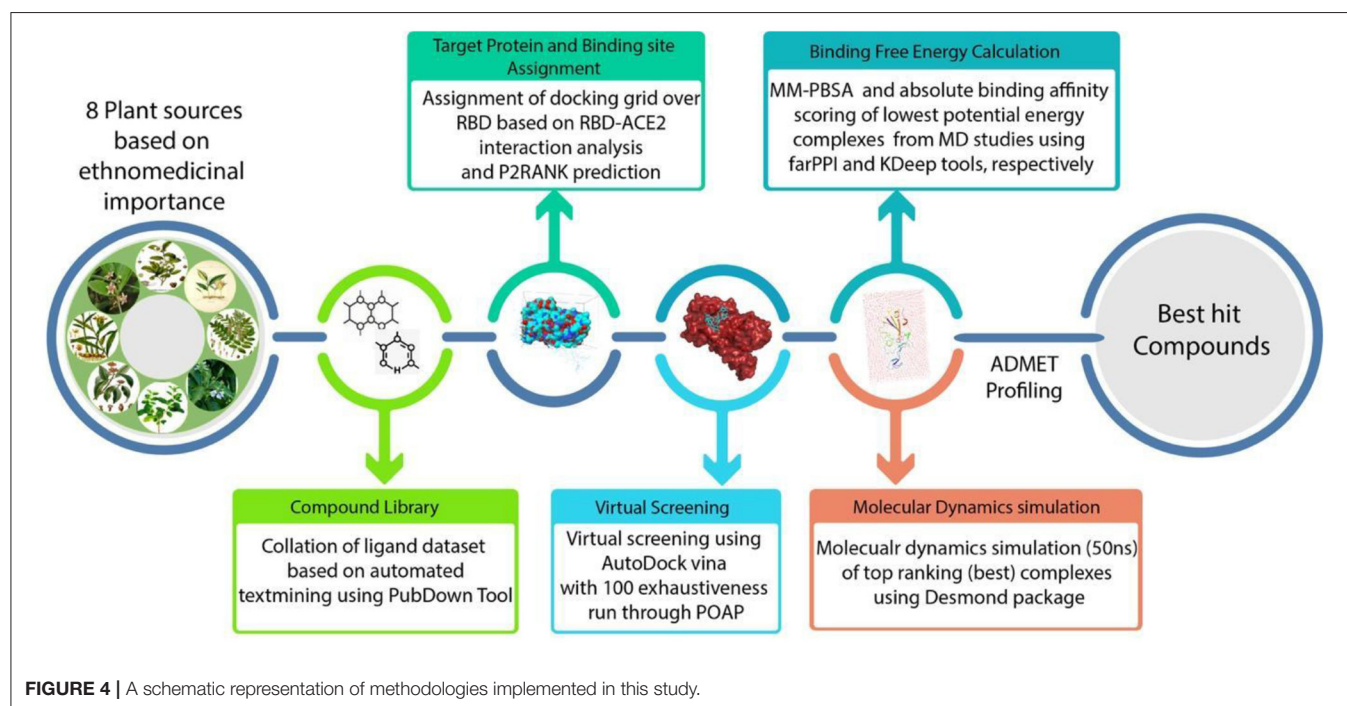
Syn. *T. asthmatica* (Roxb) Wt & Arn.

The detailed flowchart on the insilico methodologies implemented in this study towards prioritization of phytochemical moieties are shown in **Figure 4**. *Tylophora indica*, a perennial climber indigenous to India, commonly called “Antamool” is an important medicinal plant used in Indian medicine, mainly found in the plains, hills, and the forest borders in eastern and southern India. This plant is ethnically used for treating various types of ailments including cancer, respiratory infections, bronchial asthma, whooping cough, and anaphylaxis. The active ingredients of *T. indica* are mainly available in leaves and roots that exhibit most therapeutic effects mostly due to the pharmacologically active alkaloids tylophorine, tylophorinine, and tylophorinidine (48). Some previous studies have shown that Tylophora alkaloids can inhibit viral protease and suppresses viral RNA replication by blocking the JAK2 mediated NF- κ B activation (49). While tylophorine derivatives have inhibitory effects on mouse hepatitis virus (MHV), transmissible gastroenteritis virus (TGEV), and SARS-CoV

(50–52). A recent report revealed that Tylophora alkaloids could inhibit the CoV-infected cells of swine (53). However, an earlier pharmacokinetic study demonstrated moderate to good oral bioavailability of tylophorine (65.7%) in rats (50). Recent studies also showed that alkaloids from *T. indica* possess anti-replication activity and inhibit the cytopathic effect induced by apoptosis, and apoptosis induced by viral infection (54). While kaempferol derived from *T. indica* could effectively block the 3a channel protein in coronavirus (55).

Glycyrrhiza glabra L. (Fabaceae)

Glycyrrhiza glabra rhizome (Yashtimadhu) is used worldwide in various traditional systems of medicines. In Ayurveda it is an important drug component of Dasamoolarishtam, Aswagandharishtam, Madhu-yastyaditaila etc., as mentioned in Charaka Samhita. In folk medicine it is used as a laxative, emmenagogue, contraceptive, galactagogue, anti-asthmatic, antitussive, and antiviral agent. Being a member of the pea and bean family, the plant is best known for its use in making liquorice-flavored confectionery while roots and rhizomes are used for medicinal purposes. A number of pharmacological effects including expectorant and antitussive, antiviral against SARS-CoV, HIV, and in the treatment of diabetes, cancer, and hepatitis (56) have been studied for this plant. The main chemical constituent of liquorice is glycyrrhizin, a triterpene saponin with a low haemolytic index; while the root contains glycyrrhetic (Glycyrrhetic) acid, the aglycone of glycyrrhizin. Other active constituents of liquorice include isoflavonoids, chalcones, coumarins, triterpenoids, sterols, lignans, amino acids, amines, gums, and volatile oils (57). Chemically *G. glabra* comprises of 20 triterpenoids and around 300 flavonoid compounds. Among these 18 β -glycyrrhetic acid, glycyrrhizin,



glabridin, licochalcone A, licochalcone E, and liquiritigenin have antimicrobial activity (58). While glycyrrhizin A and 18 β -glycyrrhetic acid can elicit anti-HCV activity through inhibition of core protein expression and by blocking the degradation of NF κ B inhibitor I κ B, followed by activation of T lymphocyte proliferation (58). The glycyrrhizin and its analogs have significant inhibitory effect against hepatitis, herpes, influenza, and SARS viruses (59). Oral administration of *G. glabra* extract has an antitussive effect by promoting pharyngeal and bronchial secretions leading to good expectorant action. Liquiritigenin, a flavonoid from the root extracts demonstrated anti-asthmatic activity (60). The antiviral activity of glycyrrhizin have been assessed against two clinical isolates of coronavirus (FFM-1 and FFM-2) from SARS patients and found that it could inhibit viral adsorption, penetration, and replication (59). The crude Glycyrrhizin was also demonstrated to have low antiviral activity against varicella zoster virus (VZV) better than acyclovir and interferon (61). The roots of *G. glabra* had an accumulation of molecules having 3D similarities to influenza neuraminidase (NA) inhibitors. Further, it was elaborated in chemiluminescence (CL)-based NA inhibition assays on different influenza virus strains including an oseltamivir-resistant isolate A/342/09 (H1N1) that 11 out of 12 compounds had IC₅₀ in nanomolar to micromolar range (62). A study with *G. glabra* leaf extract also revealed antiviral activity against Newcastle disease virus (NDV) with an highest embryo survival rate at 300 μ g/ml (63).

***Camellia sinensis* L. (Theaceae)**

The use of *Camellia sinensis* or tea as beverage and medicine has a long history of almost 5000 years. Chemically tea contains polyphenols, flavonoids, tannins, and caffeine derivatives with amino acids, having antioxidant and diverse therapeutic effects. Black tea is prepared from the green tea leaves by a series of fermentation when catechin (30%) of green leaves oxidized into theaflavins (theaflavin, theaflavin-3-gallate, theaflavin-3'-gallate, and theaflavin-3,3'-digallate) by dimerization and into thearubigins (17%) through polymerization. Tea flavonoids help in the reduction of inflammation, possess antimicrobial effects, and are used in the treatment of respiratory diseases such as asthma. A number of compounds like theaflavins and tannins from black and green tea have antiviral activities, mainly against bovine rotavirus and bovine coronavirus (32, 64). *In vitro* studies have shown that theaflavin di-gallate inhibited the infectivity of influenza A and B viruses (65). Green tea is widely used as a beverage across the world, mainly for its antioxidant nature. It is rich in polyphenolic compounds (flavonoids) and bonded benzene rings combined with multiple hydroxyl functional groups. A study on water-soluble phenols like tannic acid and theaflavin-3,3'-digallate have shown to inhibit 3-chymotrypsin like protease (3CLpro) of SARS Coronavirus. Hence, it can be considered as a starting point for molecules against the SARS-CoV-2 (32, 66). A recent docking study revealed that the bioactive molecules of *C. sinensis*: Barrigenol, Kaempferol, and Myricetin have significant binding affinity with the active site of SARS-CoV2

Nsp15 protein (67). In a similar study, Oolonghomobisflavan-A, Theasinensin-D, and Theaflavin-3-O-gallate from tea were compared with repurposed antivirals (Atazanavir, Darunavir, and Lopinavir) for their binding affinity with Mpro of SARS-CoV-2. The results revealed that Oolonghomobisflavan-A to be highly significant in terms of binding affinity and intermolecular interactions when compared to all the other repurposed antiviral inhibitors (68).

***Justicia adhatoda* L. (Acanthaceae) Syn. *Adhatoda vasica* Nees**

Justicia adhatoda (synonym *Adhatoda vasica*), known as Vasaka in Ayurveda, is a well-known medicinal plant in indigenous system of medicine, mostly effective in treating respiratory ailments, as the leaf extract has a stimulant effect. Vasica leaf is an antispasmodic cum expectorant and has been used for centuries to treat asthma, chronic bronchitis, and other problems including fever, swelling, asthma, pneumonia, malaria, tuberculosis, cough, and cold (69). The infusion of *A. Vasica* leaf is known to relieve headaches. The root is used as an expectorant and antispasmodic; while the root infusion has an anthelmintic property. The phytochemical profiling of this plant showed the presence of alkaloids, anthraquinones, flavonoids, phenols, saponins, and tannins (70). Several studies showed that the aqueous and methanolic extracts of leaves can directly interfere with the envelop proteins of many viruses. In particular, methanolic extract had a higher level of inhibition of influenza virus, by blocking viral attachment and inhibition of viral hemagglutinin (HA) protein. Detailed study revealed that the methanolic extract mainly comprised of Vasicine alkaloids have antiviral activity (71). Aqueous extract of leaves is reported to inhibit the arachidonic acid metabolites through COX (TXB2) and LOX (LP1 and 12-HETE) pathways; while platelet aggregation studies showed butanol extract to exert strong inhibition against arachidonic acid, platelet activating factor, and collagen-induced aggregation (72). Methanolic extract also possess antiviral activity against HSV-2, while aqueous extract against HSV-1. Moreover, the methanolic extract showed 100% reduction in HA at 10 mg/ml; while the aqueous extracts at 5–10 mg/ml dose reduced the HA levels to 33 and 16.67%, respectively, suggesting strong anti-influenza activity by inhibiting viral attachment and/or replication (73).

***Ocimum Tenuiflorum* L. (Lamiaceae) Syn. *O. sanctum* L. (Tulsi)**

Ocimum sanctum or Tulsi has been used for thousands of years for its diverse therapeutic activities, and is known as the “Queen of herbs” or the legendary “Incomparable one” of India with strong aroma and astringent taste. It is the holiest and most cherished plant for its healing and health-promoting properties and in TM, Tulsi is known as an adaptogen that balances different processes in the body and helps in adapting stress. Ayurveda treats it as a kind of “elixir of life” and is believed to promote longevity and a healthy body. Thus, extracts from Tulsi are used in many Ayurvedic remedies including the common cold, headache, stomach ailments, inflammation and heart disease (74). Several studies with *O. sanctum* leaf extracts showed

therapeutic, prophylactic, and virucidal activities. A study *in ovo* model indicated its therapeutic activity against H9N2 virus by reducing the infection level (75); while crude extracts or individual compounds isolated from Tulsi have a wide spectrum of antiviral activity against HSV, Adenovirus, Cocksackievirus B1, and Enteroviruses (76). Tulsi is used in diverse formulations including mouthwash, sanitizer and water purifiers (77). The purified components apigenin, linalool, and ursolic acid of *O. basilicum* showed a broad spectrum of antiviral activities against DNA viruses (HSV, adenoviruses, hepatitis B virus) and RNA viruses (Cocksackievirus B1, Enterovirus 71), among which ursolic acid showed the strongest activity against HSV (40), ADV-8, CVB1, and EV71 (76). Crude, terpenoid, and polyphenol-rich extract of *O. sanctum* showed significant virucidal activity ($p < 0.001$ – 0.01) and was found to decrease the virus genome copy numbers at the lowest dose up to 72 h post-infection (77). Recently, molecular docking studies also suggest that tulsinol A-G and dihydro-dieuginol B as potential inhibitors of SARS Coronavirus Main Protease (Mpro) and Papain-like Protease (PLpro), indicating that *O. sanctum* can be used as preventive against CoV due to its potential immunomodulatory, ACE2 blocking and viral replication inhibition properties (78).

***Zingiber Officinale* Roscoe (Zingiberaceae)**

Zingiber officinale (Ginger), native to South-East Asia, is used as a common spice across the world. It encompasses several diverse chemical moieties with antiarthritic, anti-inflammatory, antidiabetic, antibacterial, antifungal, and anticancer activities and is one of the major medicinal sources of Ayurveda, Unani, Siddha, and various traditional medicine systems of India (79). Fresh ginger is used to treat cold, nausea, colic, heart palpitations, respiratory illnesses, dyspepsia and dry cough. During the nineteenth century a popular formulation from Ginger was used in the treatment of asthma and cough, consisted of the mixture of fresh ginger, and fresh garlic juice with honey (80). The ginger rhizome contains highly pungent vanillyl ketones like Gingerol and paradol derivatives having therapeutic effect on a wide range of diseases (81). Fresh rhizomes have inhibitory activity against the human respiratory syncytial virus (RSV) that infects the respiratory tract of humans (82). It was shown that the water-grown ginger has greater inhibitory activity against Chikungunya (CHIK) virus (83). The ginger oil was also reported to inhibit HSV-2 plaque formation (84) while the dried rhizomes containing sesquiterpenes have anti-rhinoviral activity in plaque reduction assay, but the best activity was found with the beta-sesquiphellandrene at an IC_{50} of $0.44 \mu\text{M}$ *in vitro* (85).

***Curcuma longa* L. (Zingiberaceae)**

Curcuma longa (Turmeric) belongs to the ginger family Zingiberaceae; and turmeric rhizome has been traditionally used in India for various ailments and diseases. Indian traditional and folklore medicine used turmeric to treat inflammation, infections, respiratory illness, gastric, hepatic, and blood disorders. Curcumin, the marker compound of turmeric is a well-studied therapeutic phyto-molecule, while curcumin and its derivatives are the major polyphenols of the rhizome (86). The antiviral activity of curcumin and its derivatives have been

established against a wide variety of pathogenic viruses including hepatitis, herpes simplex, human immune deficiency, human papilloma, influenza, and zika. The mechanism is mainly by inhibition of viral entry, replication, particle production, viral protease, and gene expression (87). Curcumin and its analogs also modulate the regulation of renin-angiotensin-aldosterone system (RAAS) which is involved in anti-inflammatory, anti-oxidant and anti-hypertensive activity, that are highly elevated in viral infection (88). Crude aqueous and ethanolic extracts of *C. longa* confer significant antiviral activity against H5N1 virus *in vitro* by inhibiting viral replication with significant upregulation of TNF- α and IFN- β mRNA expressions (89). Anti-influenza activity of curcumin was earlier assessed by computational methods, wherein curcumin derivatives were docked against the HA protein of influenza (H1N1) virus. The results inferred that specific curcumin derivatives can be successfully used against influenza virus infection. Moreover, curcuminoids from the methanol extract of *C. longa* also provide strong inhibitory effects on the neuraminidases of H1N1 and H9N2, as non-competitive inhibitors (90).

***Syzygium Aromaticum* (L.) Merr. & L. M. Perry (Myrtaceae)**

Clove, the aromatic flower bud of *Syzygium aromaticum* is one of the ancient traditionally used spices in almost every household in India, with several therapeutic properties for dental, digestive, and respiratory disorders, including asthma (91). The other application of clove includes food preservation. Cloves contain various classes of phytochemicals including sesquiterpenes, monoterpenes, hydrocarbons, and phenolics along with Eugenyl acetate, eugenol, and β -caryophyllene as principal components of clove oil. Various pharmacological studies with clove have shown its inhibitory effects on pathogenic bacteria, *Plasmodium*, and Herpes simplex and Hepatitis C viruses (92). The essential oil of clove contains 85–95% eugenol and is shown to be highly effective in the treatment of HSV and HCV by blocking viral replication. The synergistic action of acyclovir and *S. aromaticum* extract have a significant impact on the inhibition of viral replication (92). Aqueous extract of clove showed antiviral activity against Feline Calicivirus (FCV) as a surrogate for human norovirus. Pre-treatment of FCV with clove oil reduced viral titer to 6.0 logs. The antiviral activity of the pure eugenol was similar to the clove extract, albeit at a lower level (93). The silver nanoparticles prepared from the aqueous extract of the flower buds of *S. aromaticum* were found to be novel and effective against the Newcastle Viral Disease (NDV) *in vitro* and in embryonated eggs (94).

Data Sources

In this study, the dataset of phytochemicals of eight plants were acquired from different sources like CMAUP (95), NPASS (96), Dr. Dukes Database and KnapSack (97) database (Supplementary Table 1). Initially, the list of all the phytochemicals were collated out from individual sources by manual curation. In the next step, all the duplicate entries and the ubiquitous chemicals were removed from the list. The final list was taken as input for downloading structural files from the

PubChem database using an automation script created using Python programming (<https://github.com/sandes89/PubDown>). The structures which were accessible online and documented were only considered for screening.

Docking Studies

Preparation of Protein

The crystal structure of RBD (PDB id: 7BZ5) of spike protein presented in **Figure 3** was downloaded from RCSB PDB (Protein Data Bank) (98, 99). GUI based “Auto-Dock Tools” was used to prepare and execute the docking studies. Kollman atom charges, solvation parameters, and polar hydrogens were added to the protein and proceeded for docking studies. As the ligands used are not peptides, Gasteiger charges were assigned only to the protein and the non-polar hydrogens were merged. Based on the literature and predicted active regions, a grid box was assigned around the active sites using AutoGrid application (100).

Preparation of Ligands

The 2D/3D structures were retrieved from PubChem Database using a custom written python script which is hosted on GitHub portal (<https://github.com/sandes89/PubDown>). List of compounds with their chemical names were prepared as an input to Python script and searched iteratively on PubChem ftp database and the compounds were downloaded in sdf (Structure Data File) format. A total of 1,952 compounds were downloaded from PubChem database. In this study, POAP (101) was used for the preparation of ligands and for virtual screening. POAP tool is a bash shell script-based pipeline which can be used to optimize ligands for docking using Open Babel (102) and to perform virtual screening using Autodock Suite. POAP implements dynamic file handling methods for efficient memory usage and data organization, ligand minimization (5,000 steps), MMFF94 force-field was employed with the addition of hydrogens. A total of 50 conformations for each compound were generated using the weighted rotor search method, with minimization using the steepest descent method. Finally, the best conformation was retained in.pdbqt format for utilization in further docking studies.

Active Residues Definition and Cavity Prioritization

The most important aspect in docking studies is the identification of important residues and favorable cavity for ligand binding. In this study, the cavity definition was mainly performed based on the literature insights on important residues (ACE2 binding site) coupled with the cavity prediction using P2RANK (103). The P2RANK predicted cavity spanned the ACE2 binding residues, as well as on few glycosylation sites. It should be noted that viral glycosylation has many roles in viral pathogenesis and biology, as it affects protein folding and stable interaction with host cells (15). As discussed earlier, glycosylation process in coronaviruses mainly occur to camouflage the immunogenic process in the host. Targeting glycosylation sites can aid in the primary and rapid immune response to neutralize the virus (16). Hence, along with ACE2 binding residues, glycosylation sites spanning the active cavity predicted by P2RANK were also considered for grid box generation. Considering the importance of the binding site with

ACE2, glycosylation sites as well as the active cavity predicted by the P2RANK tool, the docking grid for molecular docking was fixed.

Virtual Screening Using POAP and ADMET Prediction

The geometry optimized compounds were subjected to Molecular docking with SARS-CoV-2 Spike glycoprotein. In the docking process, the ligands were considered as flexible and protein was considered as rigid body. The Grid box was prepared based on the active site residues as inferred from earlier ligand co-crystallized complex of spike protein and P2RANK based binding pocket prediction. For the docking process, an exhaustiveness value of 100 was fixed in Vina. The resulting Protein-ligand complexes were analyzed for intermolecular interactions using PLIP tool (104). The top-ranking ligands were subjected to ADMET profiling using pKCSM server (105).

Molecular Dynamics (MD) Simulation

The MD simulation of Apo protein and docked complexes were carried out using Desmond version 2020. Here, OPLS_2005 force field was used to initiate the MD simulation, and the system was solvated using SPC (Simple point charge) water model (106). The neutralization of the system was performed by adding counter ions and the details of ions and concentration added to complexes are given in **Supplementary Table 11**. Energy minimization of the entire system was performed using OPLS_2005, as it is an all-atom type force field (107). In the studies on natural compounds, the application of OPLS_2005 force field was found to be highly optimal. Hence, it was adopted for this study (27). The geometry of water molecules, the bond lengths and the bond angles of heavy atoms was restrained using the SHAKE algorithm (108). Simulation of the continuous system was executed by applying periodic boundary conditions (109) and long-range electrostatics was maintained by the particle mesh Ewald method (110, 111). The equilibration of the system was done using NPT ensemble with temperature at 300 k and pressure at 1.0 bar. The coupling of temperature-pressure parameters was done using the Berendsen coupling algorithm (112). On post-minimization and equilibration of the system, the Apo protein system consisted of 28,645 atoms in total and number of atoms for all the complexes are given in **Supplementary Table 11**. On post-preparation of the system, the production run was performed for 50 ns with a time step of 1.2 fs and trajectory recording was done for every 5.0 ps summing up to the recording of 10,000 frames. The calculation of the RMSD (Root mean square deviation) was done for the backbone atoms and was analyzed graphically to understand the nature of protein-ligand interactions (113, 114). RMSF (Root Mean Square Fluctuation) for every residue was calculated to understand the major conformational changes in the residues in comparison between the initial state and dynamics state (115). The compactness of the protein-ligand complexes in comparison to Apo form was calculated using radius of gyration rGyr (116). The 2D interactions of Protein-ligand complex showing the stability of the complexes and interaction sites were generated for the complete run time.

KDeep Based Absolute Binding Affinity (ΔG) Calculation

On post-molecular dynamics simulation, the top ranking complexes from each plant were energy minimized and was also analyzed for absolute binding affinity (ΔG) using KDeep (117). KDeep employs machine learning approach with implementation of 3D convolutional neural networks. KDeep analyses the input and voxelizes into pharmacophore features like (aromatic, hydrophobic, hydrogen-bond acceptor/donor, positive, and negative ionizable). The prepared input is passed to the DCNN (Deep Convolutional Neural Network) model which is pre-trained by PDB bind v.2016 database, wherein, based on adaptability to the model the absolute free energy of the protein-ligand complex is calculated.

MM-PBSA Calculation of Topmost Stable Complexes

Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) calculation is one of the most commonly used method for enumerating binding free energy of protein-ligand complexes. The MM-PBSA combines energy calculations based on molecular mechanics and implicit solvent model. This method precisely estimates the binding free energy of the protein-ligand complex, which is estimated by the differences between the free energy of the complex and free energies of unbound individual components of the complex (118). In this study, the MM-PBSA (PB3) based binding free energy of top most stable complexes were calculated using farPPI server. While PB3 option was considered as it was benchmarked to be highly accurate when compared to all the other methods in FarPPI (119). The force fields, GAFF2 and ff14SB as provided in the server were applied for ligand and protein, respectively. This calculation was performed for the lowest potential energy conformation of the top most stable complexes.

RESULTS

Binding Site Assignment

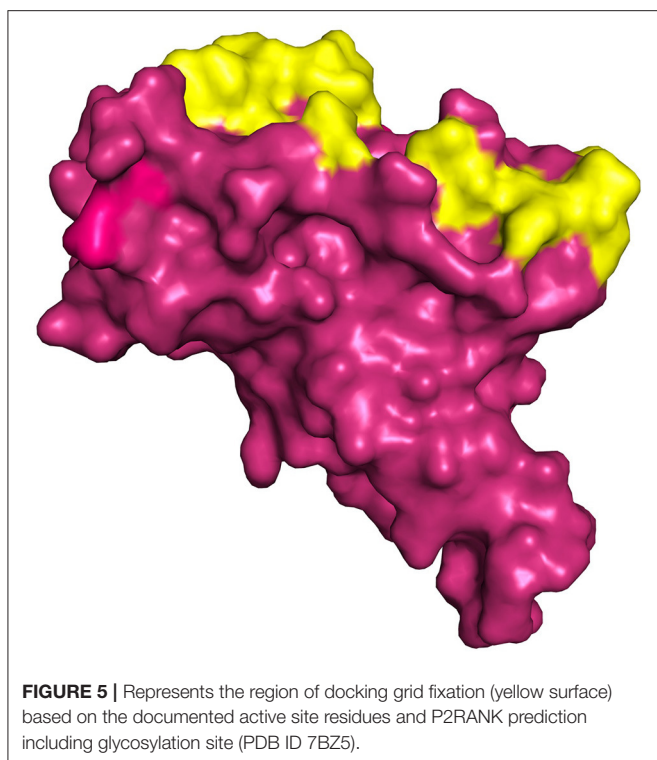
The structure of spike protein co-crystallized with the ligand is yet to be available in PDB, hence P2RANK was used to identify the potential drug binding pockets in concurrence with key hotspot residues reported in the literature. Based on the interaction plot of Spike protein (RBD)-ACE2 receptor complex (PDB ID: 6M0J), the residues of Spike protein that are involved in the interactions were identified as GLY502, THR500, LYS417, TYR449, GLY446, ASN487, and GLY496. In addition, the P2RANK prediction also covered the ASN343 glycosylation site spanning the RBD of spike glycoprotein. Considering all these important residues based on the literature, cavities and solvent accessible areas as predicted by P2RANK, an optimal docking grid box was created over the spike protein (PDB ID: 7BZ5). It should be noted that the 6M0J structure was only utilized for mapping the ACE2-Binding region on the spike protein, and was not used for docking studies, as some atoms were found missing in the structure. Hence, the 7BZ5 structure (ACE2 unbound form) with no such issues were utilized for docking and simulation studies. The Grid file was generated with the following coordinates ($x = -85.56$, $y = -23.44$, $z = -16.90$)

using Autodock tools program and was proceeded for molecular docking using Vina as shown in **Figure 5**.

Molecular Docking Studies

In pursuit of finding an important candidate for managing COVID-19 from selected plants, molecular docking studies were carried out with phytochemicals listed from eight plants on the binding pocket of COVID-19 spike glycoprotein (PDB ID: 7BZ5). Based on literature review, it was clearly found that the virus enters the human cell via the ACE2 receptor. Hence, we prioritized the receptor binding domain of spike glycoprotein PDB ID: 7BZ5, wherein the spike protein interacts with ACE2 for docking studies. The geometry optimized compounds from all the plants were docked against active-cavity as discussed above, and were ranked based on their corresponding docking score. Compounds having the docking score of < -7.0 kcal/mol were considered for further evaluation. This cut-off was adopted, based on earlier studies, wherein it was found to be optimal (120, 121). A comprehensive evaluation of all the compounds was performed based on the binding affinity score and the involvement of key residues in the binding cavity (**Table 1**). Also the ADMET profiling of the compounds with topmost binding affinity was carried out using pKCSM server (105) and the data is provided in **Supplementary Table 10**.

The compound Rutaecarpine from *T. indica* was found to be interacting with spike protein at SER371 and SER373 with a binding affinity of -7.9 kcal/mol (**Figure 6A**). Tylophorinidine showed a hydrogen bond with ASN343 (glycosylation site) and binding energy of -6.9 kcal/mol. Licoagrodin from *G. glabra* was found to interact with GLY339, ASP364, VAL367, and SER371 of RBD region with a binding affinity of -8.7 kcal/mol (**Figure 6B**). Further analysis also showed that Hispaglabridin-B, Licoagron, and Licocoumarin-A from *G. glabra* to form hydrogen-bonded interactions with glycosylation site ASN343 with a binding energy of -8.2 kcal/mol, respectively. Cryptoxanthin showed hydrogen bond with ASN440, stabilized by 7 hydrophobic interactions having a binding energy of -8.4 kcal/mol (**Figure 6C**); while 3-O-Galloylepicatchin-(4Beta-6)-Epigallo-catechin-3-O-Gallate (-8.3 kcal/mol) showed hydrogen bonded interactions at positions: PHE338, ASN370, SER371, SER373, ASN437, and ASN440. Based on this significance it was considered for further studies. Furthermore, compounds namely, Camelliquercetinside-B (-7.8 kcal/mol), Procyanidin C1 (-7.8 kcal/mol), 3-O-Galloylepicatchin-(4Beta-6)-Epigallo-catechin-3-O-Gallate (-7.7 kcal/mol), Theasinensin B (-7.6 kcal/mol), Epigallocatechin-(2 Beta-7,4 Beta-8)-Epigallocatechin-3-O-Gallate (-7.5 kcal/mol), 3-O-Galloylepicatchin-(4 Beta-8)-Epicatechin-3-O-Gallate (-7.3 kcal/mol), and Theasinensin-C (-7.2 kcal/mol) from *C. sinensis* also showed hydrogen bond with residue ASN343. Compounds from *J. adhatoda* did not show any interactions with active residues, however interactions were observed with residues proximal to the active region: PHE342, SER371, and SER373. Among all the compounds studied from *J. adhatoda*, Daucosterol showed the highest binding affinity with a score of -7.6 kcal/mol (**Figure 6D**). In case of *O. tenuiflorum*, Stigmastanol (-7.7 kcal/mol) showed interactions



with surrounding residues of active cavity and (+)-Taxifolin (-7.0 kcal/mol) formed hydrogen bonded interaction with ASN343, while other compounds showed interactions only with other residues proximal to active residues and showed higher binding affinity. Among these compounds, Caryophyllene featured the highest binding affinity with a score of -8.1 kcal/mol (**Figure 6E**). Compounds Isoginkgetin (-7.6 kcal/mol), Rutin (-7.3 kcal/mol), 4'-Methoxyglabridin (-7.2 kcal/mol), Curcumin (-7.2 kcal/mol), Cubebin (-7.2 kcal/mol), Cyanin (-7.0 kcal/mol), and (Z)-1,7-bis(4-hydroxy-3-methoxy-phenyl) hept-4-en-3-one (-7.0 kcal/mol) from *Z. officinale* showed hydrogen bonded interaction with active cavity residues. The marker compounds of *Z. officinale* like Gingerenone-A and B, and Isogingerenone-B showed interactions around the active residues PHE342, SER371, and SER373, wherein Geraniin showed highest affinity with a score of -8.2 kcal/mol (**Figure 6F**). In case of *C. longa*, Letestuianin A (-7.2 kcal/mol) showed hydrogen bonded interaction with residue ASN343. Moreover, the marker compounds like Curcumin (-7.2 kcal/mol) and its derivatives showed interactions around the active site residues CYS336, PHE342, SER371, and SER373; wherein O-Demethyldemethoxycurcumin showed significant binding affinity with a score of -8 kcal/mol (**Figure 6G**). In case of *S. aromaticum* Tellimagrandin-II (-8.2 kcal/mol), Rugosin-D (-7.9 kcal/mol), Syzyginin-A (-7.8 kcal/mol), Campesterol glucoside (-7.6 kcal/mol), Sitoglucide (-7.7 kcal/mol), Cirrhoptalanthrins (-7.4 kcal/mol), Tellimagrandin-I (-7.4 kcal/mol), Rugosin-E (-7.3 kcal/mol), Strictinin (-7.3 kcal/mol), Tannin (-7.1 kcal/mol), Myricetin (-7.0 kcal/mol),

and Quercetin (-7.0 kcal/mol) showed hydrogen bonded interactions at active residue ASN343 and other residues. Among these compounds Tellimagrandin-II showed the highest affinity with a score of -8.2 kcal/mol (**Figure 6H**). The 2D interaction diagrams of all the compounds are given in **Supplementary Figures 2–9**.

Molecular Dynamics Simulation of Top Docked Complexes

Molecular dynamics simulation for all the top-ranking complexes per plant was carried out with Desmond for a duration of 50 ns. The data from the trajectory was analyzed and tabulated in **Supplementary Table 11**. The RMSD and RMSF values of the protein backbone for all the 8 complexes were plotted and are shown in **Figures 7, 8**.

Based on the stability, compactness, and ligand contacts during the simulation process, Spike protein O-Demethyldemethoxycurcumin and Spike protein Tellimagrandin-II complexes were found to be more stable and were analyzed further in detail.

Spike Protein O-Demethyldemethoxycurcumin Complex

The simulation system of Spike protein O-Demethyldemethoxycurcumin consisted of 25,902 atoms with 7,659 water molecules. To further neutralize the system, 3 Cl^- (7.122 mM) were added and the system was subjected to 50 ns run of production run. The RMSD plot showed a convergence at 10 ns with ~ 1.5 Å difference in the ligand bound state (**Figure 9**). The Ligand RMSD values remained within the range of 1.0–2.5 Å with average RMSD value being 1.75 Å (**Figure 9**). The lowest potential energy conformation was found at 21.5 ns with energy value of $-84,977$ kcal/mol with a binding free energy (MM-PBSA) of -12.48 kcal/mol.

The mobility of the compound in the complex during simulation with residue-wise calculations was plotted as RMSF trajectory. The analysis of the RMSF plot inferred that there was a minimum fluctuation around ~ 1 Å, and the trajectory to remain stable throughout the simulation with maximum deviation of ~ 2.4 Å (**Supplementary Figure 10**). Further the radius of gyration (rGyr) trajectory was plotted for the entire production run, wherein the deviation was ~ 4.70 – 5.5 Å, thereby implying the higher compactness during the simulation process (**Supplementary Figure 11**).

Protein-Ligand contact analysis inferred that CYS336 to form hydrogen-bonded interactions for around 80% of the duration, followed by PHE338, which showed 60% of time to interaction by means of hydrogen bond and hydrophobic interactions. PHE342, ASP364, VAL367, and TRP436 showed around 50% of the time with interaction fraction which includes hydrophobic and water-bridge interactions (**Figures 10A,B**).

Spike Protein—Tellimagrandin-II Complex

The simulation system of Spike protein Tellimagrandin-II complex comprised of 25,917 atoms with 7,645 water molecules. To further neutralize the system, 3 Cl^- (7.135 mM) were added.

TABLE 1 | Compounds from each plant with best binding energy/score.

Plant id	Plant name	Compound name	Binding energy (kcal/mol)	KDeep ΔG (Kcal/mol)	Interactions		
					Hydrogen bond (Distance in Å)	Hydrophobic interaction (Distance in Å)	Pi-stacking
NITM1	<i>Tylophora indica</i> (Burm. F.) Merrill Syn. <i>Tylophora asthmatica</i> (Roxb.) Wt & Arn.	Rutaecarpine	−7.9	−6.15	SER371 (3.23), SER373 (3.14)	PHE342 (3.16), VAL367 (3.39), LEU368 (3.67), PHE374 (3.64)	TRP436
NITM2	<i>Glycyrrhiza glabra</i> L.	Licoagrodin	−8.7	−9.45	GLY339 (3.13), ASP364 (3.49), VAL367 (4.06), SER371 (3.13)	PHE338 (3.89), GLU340 (3.71), ASP364 (2.93), VAL367 (3.69), LEU368 (3.46), PHE374 (3.93), TRP436 (3.83)	
NITM3	<i>Camellia sinensis</i> L.	3-O-Galloylepicatechin-(4Beta-6)-Epicatechin-3-O-Gallate	−8.3	−10.95	PHE338 (2.46), ASN370 (2.17), SER71 (2.57), SER373 (2.21), ASN437 (3.44), ASN440 (2.67)	PHE338 (3.77), PHE342 (3.63), VAL367 (3.58)	
NITM4	<i>Justicia adhatoda</i> L., Syn. <i>Adhatoda vasica</i> Nees	Daucosterol	−7.6	−12.77	PHE342 (3.23), SER373 (3.72), TRP436 (3.13), ARG509 (2.94)	LEU335 (3.41), PHE338 (3.55), PHE342 (3.51), ASP364 (3.94), VAL367 (3.72), PHE374 (3.80)	
NITM5	<i>Ocimum tenuiflorum</i> L., Syn. <i>Ocimum sanctum</i> L.	Caryophyllene	−8.1	−13.77	PHE338 (2.87), GLY339 (2.88), SER373 (2.70)	LEU335 (3.00), PHE338 (3.75), PHE342 (3.89), ASP364 (3.30), VAL367 (3.62), LEU368 (3.19), PHE374 (3.72)	
NITM6	<i>Zingiber officinale</i> Roscoe	Geraniin	−8.2	−9.40	VAL362 (3.79), ASP364 (3.04), VAL367 (3.29), SER371 (3.81)	LEU368 (3.68)	
NITM7	<i>Curcuma longa</i> L.	O-Demethyl demethoxycurcumin	−8.0	−7.84	CYS336 (3.04), ASP364 (3.84)	LEU335 (3.70), PHE338 (3.42), ASP364 (3.67), VAL367 (3.55), LEU368 (3.95), PHE374 (3.85)	PHE374, TRP436
NITM8	<i>Syzygium aromaticum</i> L.	Tellimagrandin-II	−8.2	−15.68	CYS336 (2.90), PRO337 (3.52), GLY339 (3.72), GLU340 (3.37), ASN343 (3.15), ASP364 (3.73), VAL367 (3.72), SER371 (2.91), SER373 (2.93)	PHE338 (3.28), VAL367 (3.47), LEU368 (3.39)	

Binding energy—Autodock binding score; KDeep ΔG —absolute protein ligand binding affinity calculated using KDeep tool.

The system was subjected to a 50 ns run of the production run. The RMSD plot for this complex showed convergence at 10 ns with ~ 1 Å (admissible range) of deviation in intermolecular interactions during the entire production run (**Figure 11**). The Ligand RMSD values remained within the range of ~ 2.0 – 3.0 Å with an average mean value of 2.5 Å (**Figure 11**). The lowest potential energy conformation was found at 40.8 ns with energy value of $-83,747$ kcal/mol with a binding free energy (MM-PBSA) of -32.08 kcal/mol.

The mobility of the compound in the complex during simulation with residue-wise calculation was visualized as (root mean square fluctuation) RMSF plot. On further analysis, the plot inferred a minimum fluctuation in the ligand bound position ~ 1 Å. Moreover, the trajectory remained stable throughout the simulation with a maximal deviation of ~ 2.5 Å (**Supplementary Figure 12**). Furthermore, radius of gyration (rGyr) trajectory was also plotted, which inferred rGyr to have maintained ~ 5.70 – 6 Å, thereby implying the higher compactness of ligand (**Supplementary Figure 13**). Protein-Ligand contact

analysis for the simulation period of 50 ns inferred that GLU340 and ASP364 to show hydrogen bonded and water-bridge interaction fraction for around 100% of the duration, with interactions at more than one position. ASN343 showed interactions around 60% of the simulation run, which mainly includes hydrogen bonds and water bridges. VAL367 showed interactions around 100% of the time includes Hydrogen bonds, Hydrophobic, and water-bridge interactions (**Figures 12A,B**).

DISCUSSION

On cumulative analysis of all the results, it could be inferred that *Tellimagrandin-II* and *O-Demethyldemethoxycurcumin*, were highly potential hits, as these compounds feature significant interactions with ACE2 binding region coupled with key glycosylation site (ASN343) of spike protein. Recent mutagenesis studies strongly suggest that the targeting ASN343 glycosylation to be the most potential inhibitory mode. Moreover, the

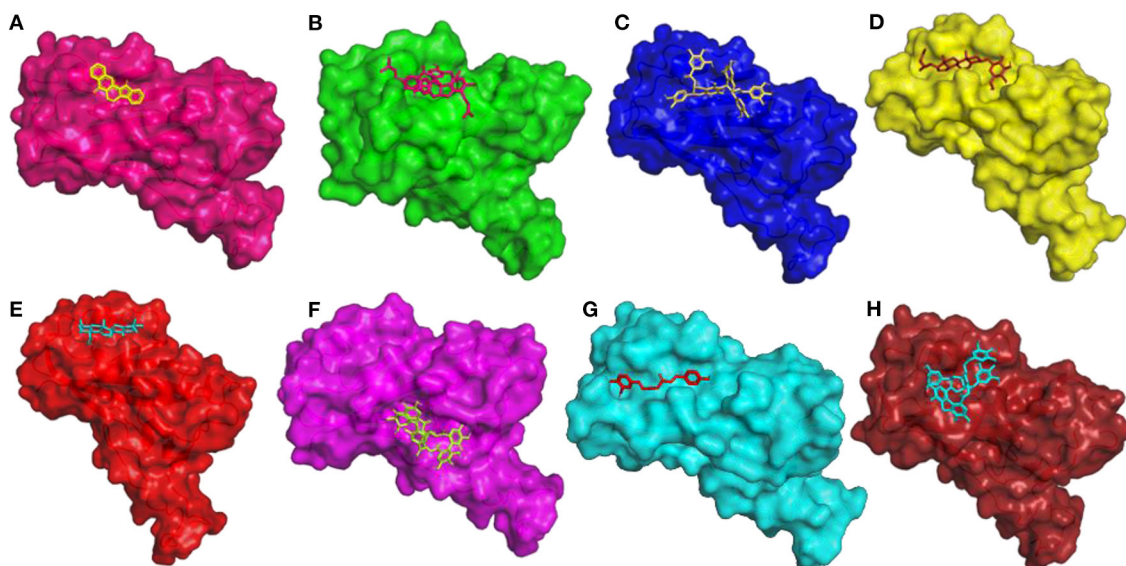


FIGURE 6 | 3D diagram of RBD of Spike Glycoprotein in complex with (A) Rutacarpine, (B) Licoagrodin, (C) 3-O-Galloylepicatechin-(4Beta-6)-Epicatechin-3-O-Gallate, (D) Daucosterol, (E) Caryophyllene, (F) Geraniin, (G) O-Demethyldemethoxycurcumin, and (H) Tellimagrandin-II.

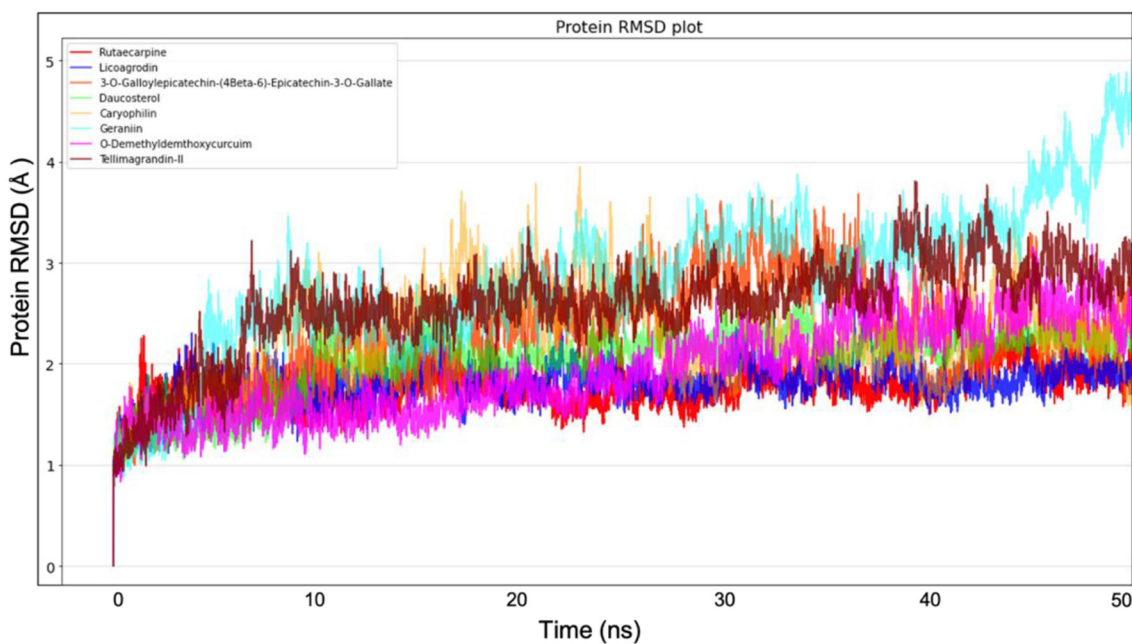


FIGURE 7 | Protein Backbone RMSD plots from Molecular dynamics for all the 8 top ranking protein-ligand complexes.

infectivity of SARS-CoV-2 showed reduction to almost 1,200-folds when both ASN331 and ASN343 were mutated in spike protein. This shows the significance of blocking these glycosylation sites on the receptor binding domain (122). Viral glycosylation holds a major role in pathogenesis, as it mediates protein folding, shaping viral tropism, and host

invasion (123). Blocking of glycosylation not only aids in preventing viral pathogenesis, but also facilitates immune recognition of the virus (124, 125). Based on the number of intermolecular interactions with active residues, glycosylation sites, and proximal residues to active site, Tellimagrandin-II with a binding energy of -8.2 kcal/mol from *S. aromaticum* may

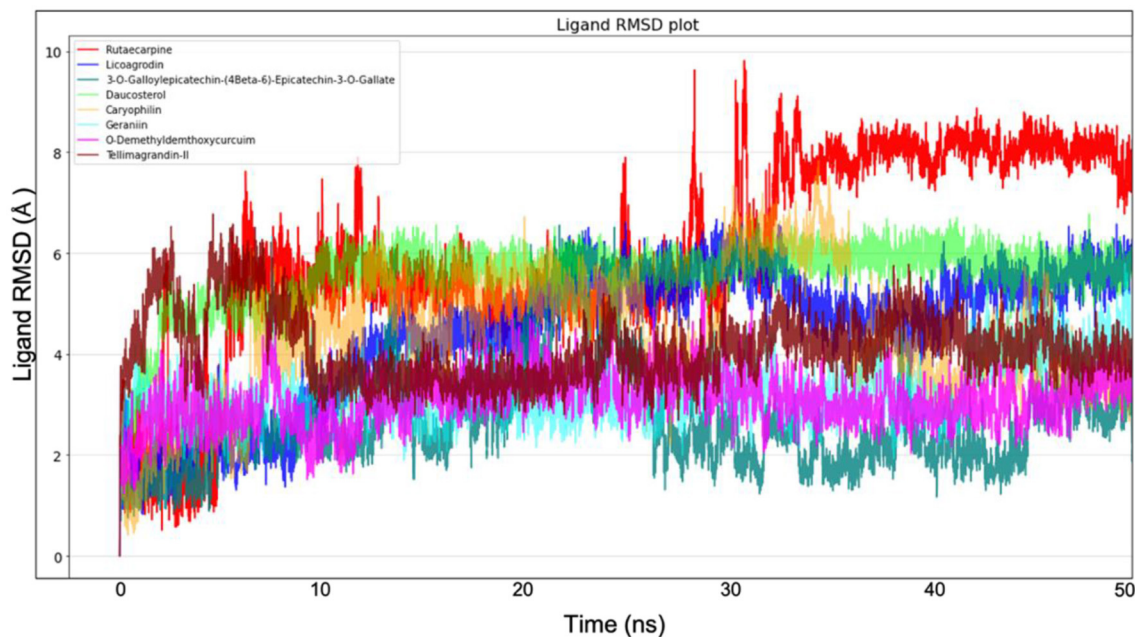


FIGURE 8 | Ligand RMSD plot from Molecular dynamics for all the 8 molecules.

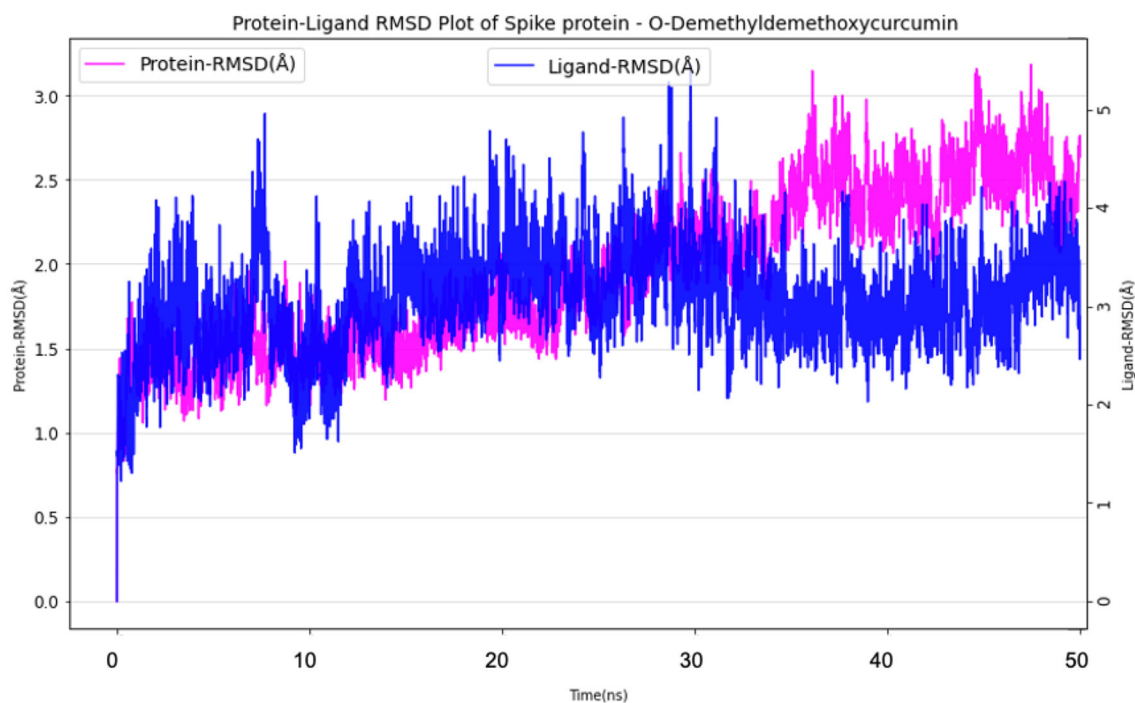
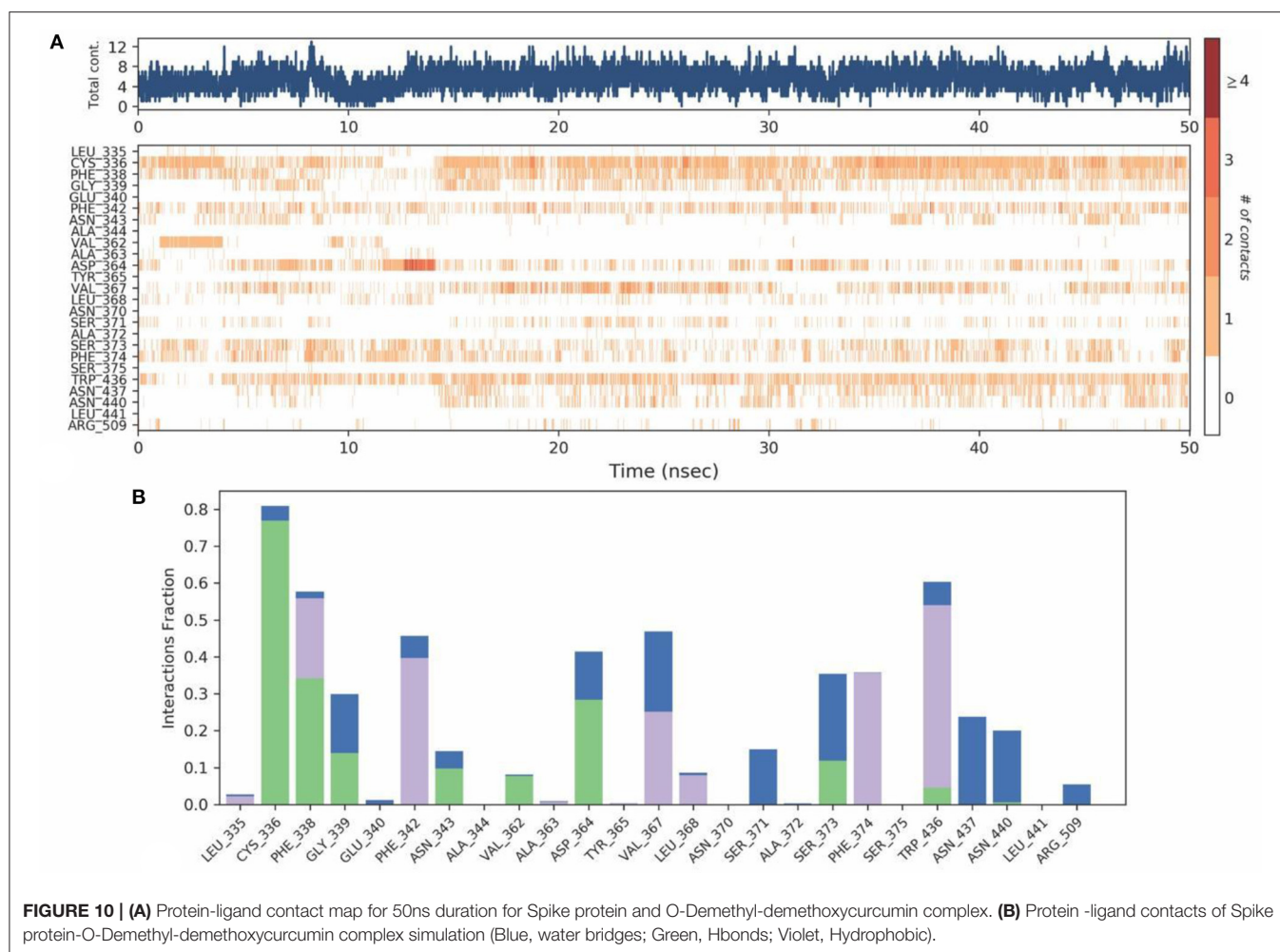


FIGURE 9 | Protein-Ligand RMSD plot of Spike protein O-Demethyldemethoxycurcumin complex.

have a higher affinity toward the spike protein in comparison with all other compounds. Moreover, it also formed a stable complex, as inferred by molecular dynamics simulation. The

hydrolysable tannin Tellimagrandin-II is traditionally known to possess antiviral activity; while hydrolysable tannins as a whole class are well-known antiviral agents (126). Tannins are



known to inhibit various viral activities like attachment and penetration of virus and inhibition of reverse transcriptase (127). Tellimagrandin-II is the first polyphenolic ellagitannin formed from 1,2,3,4,6-pentagalloyl-glucose, and is an isomer of punicafolin or nupharin A, also known as Cornustannin 2 or Eugeniiin ($C_{41}H_{30}O_{26}$; Molecular Mass 93,866 g/mol). The compound is isolated from the dried flower bud of *S. aromaticum* (Clove). Earlier studies showed that ethanol extract of *S. aromaticum* to possess strong inhibition of recombinant NS2BNS3 proteases of DENV-2 and 3; while its bioactivity guided fractionation yielded eugeniiin (128, 129), isobiflorin (5,7-dihydroxy-2-methylchromone-8C- β -d-glucopyranoside), and biflorin (5,7-dihydroxy-2-methylchromone-6C- β -d-glucopyranoside). Interestingly the eugeniiin from *S. aromaticum* and *Geum japonicum* is found to inhibit α -glucosidase and possess significant antiviral activity against wild-type HSV-1 and HSV-2. Moreover, eugeniiin also targets thymidine-kinase deficient or acyclovir as well as phosphonoacetic acid (PAA)-resistant HSV-1 at EC_{50} of 5.0 μ g/ml, with CC_{50} of 69.5 μ g/ml (130). Unlike nucleoside analogs, Eugeniiin is reported to inhibit viral DNA polymerase and late protein syntheses in HSV-infected Vero cells, in a non-competitive manner with respect to dTTP

(130). Animal studies revealed that Eugeniiin at 0.3 mg/kg at oral and intraperitoneal dose retard the development of skin lesions of HSV-1-infected mice; while at 6 or 50 mg/kg it significantly prolonged the mean survival times and or reduced mortality without toxicity. However, at an oral dose of 50 mg/kg it reduced virus yields in the skin and brain of infected mice with higher bioavailability. Moreover, Eugeniiin enhance the anti-HSV-1 activity of acyclovir, and interact with the polymerase near PAA-binding site (131). Eugeniiin in pure form demonstrated potent inhibition of NS2BNS3 proteases of DENV-2 and 3 at IC_{50} of 94.7 nM and 7.53 μ M; while moderate inhibition was found with isobiflorin and biflorin at 58.9 and 89.6 μ M (132). Furthermore, the kinetic studies revealed a competitive inhibition at same binding site of both proteases; while the K_i value of eugeniiin is reported as 125.2 nM for DENV2 protease, and 7.1 μ M for DENV3 protease (132).

Secondly, O-Demethyl-demethoxycurcumin from *C. longa* was predicted to be a promising molecule for inhibition of SARS-CoV-2 pathogenesis. It should be noted that O-Demethyl-demethoxycurcumin not only confers inhibitory effects on the SARS-CoV-2 spike protein as per our prediction, but is also well-proven to be involved in Endoplasmic reticulum

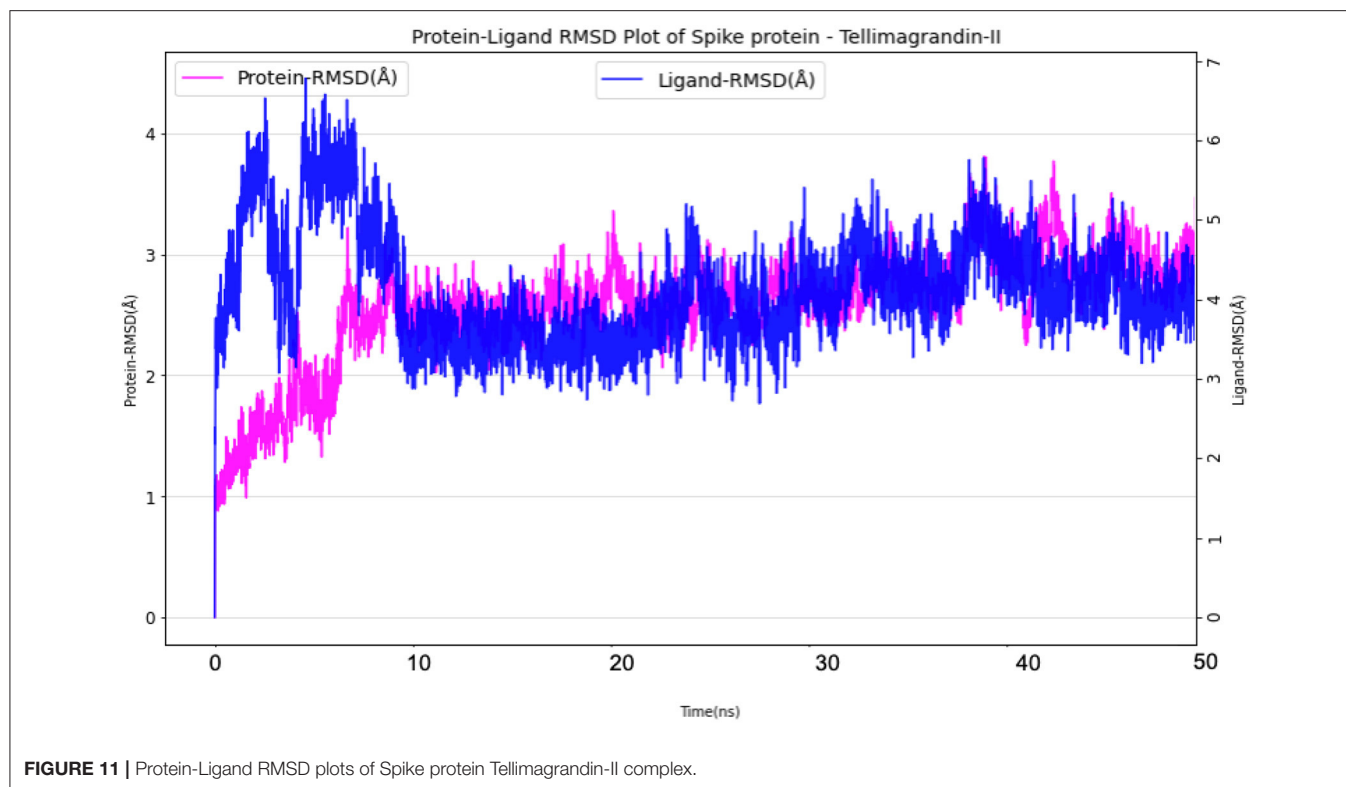


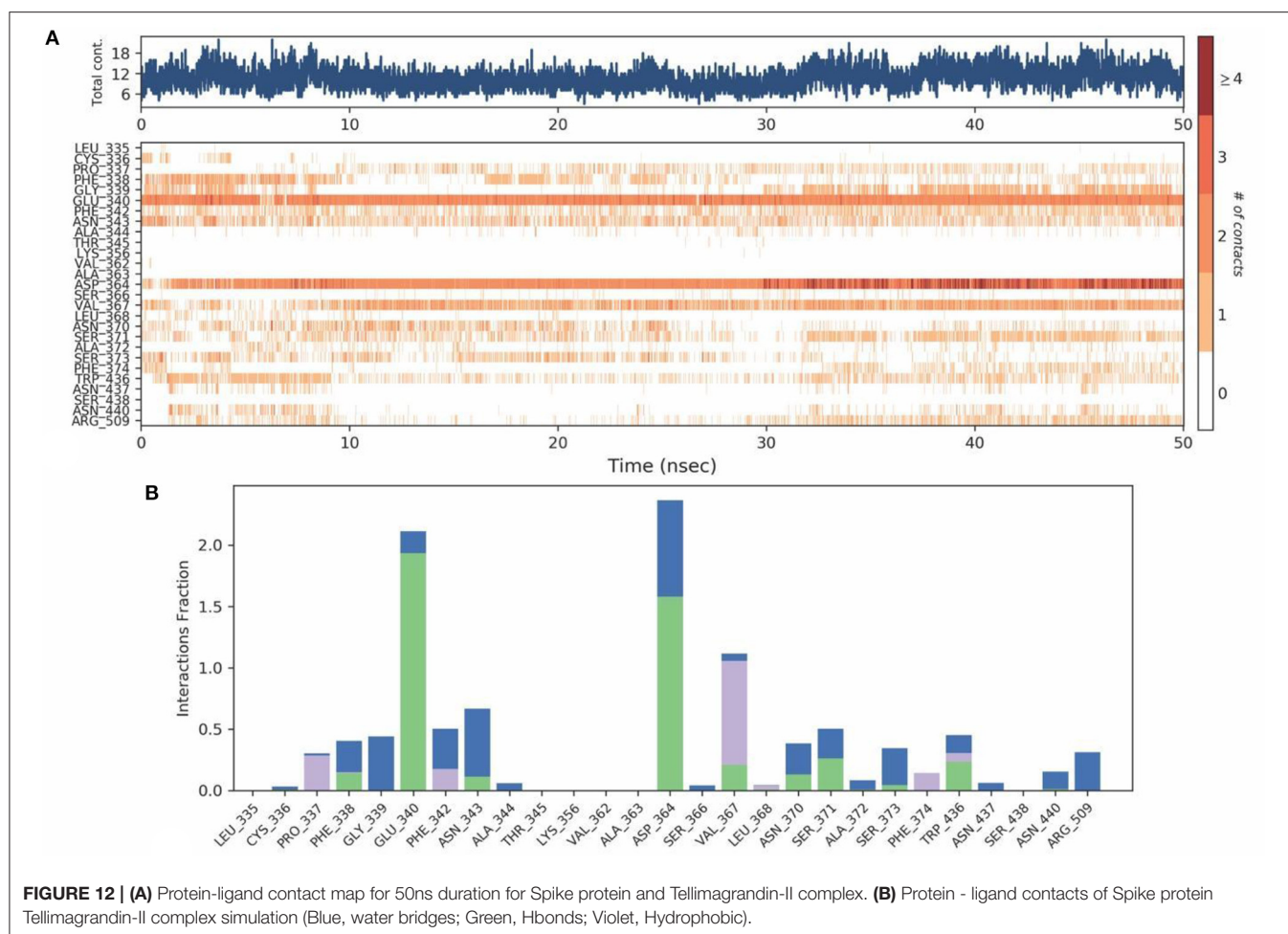
FIGURE 11 | Protein-Ligand RMSD plots of Spike protein Tellimagrandin-II complex.

(ER) stress reduction (133). It is well-known that ER stress reduction is crucial in viral replication and infection, and is an essential aspect in reducing the infection level, as the complete secretory mechanism of the virus occurs in ER (134, 135). Moreover, ER stress is one of the major problems in SARS-CoV-2 infection, as the synthesis and folding of transmembrane protein loses balance and the amount of proteins entering the ER increases drastically. This loss of balance culminates in the aggregation of unfolded proteins in the ER, which in turn triggers the ER stress response that initiates to assist the organelle for homeostasis. SARS-CoV-2 activates the Unfolded protein response (UPR) and hijack the signaling pathways for its benefit to infect rapidly, hence the reduction of ER stress in the body can be a potential way of blocking SARS-CoV-2 infection (136). Curcumin and its derivatives are treated as miraculous molecules in many infectious diseases as well as in immunomodulation. The derivatives of curcumin in combination with advanced drug delivery systems may work in a multi-faceted way for the treatment and prevention of SARS-CoV-2 (88). A recent computational study showed that curcumin exhibit strong binding affinity to Spike protein of SARS-CoV-2, ACE2 receptor of host, and their complex (RBD of viral S protein and ACE2; RBD/ACE2-complex) with the binding affinity values of -7.9 kcal/mol; -7.8 kcal/mol; and -7.6 kcal/mol; -9.1 and -7.6 kcal/mol, respectively. Moreover, molecular dynamics simulation also substantiated the curcumin's interaction within RBD site, thereby predicts the possibility of therapeutic strategy against SARS-CoV2 (120).

The ADME and toxicity profile of O-Demethyldemethoxycurcumin and Tellimagrandin II were predicted and summarized using pKCSM (105). The Intestinal absorption of O-Demethyldemethoxycurcumin was found to be high compared to Tellimagrandin-II i.e., 76.46 and 41.54%, respectively. O-Demethyl-demethoxycurcumin was found to be CYP3A4 substrate and inhibitor of CYP1A2, CYP2C19, and CYP2C9; whereas Tellimagrandin-II was predicted as a non-inhibitor. However, both the compounds were found to be non-substrate to CYP2D6 and non-inhibitor of CYP2D6 and CYP3A4, thus shall be non-toxic. Further, both these compounds showed a negative effect in the AMES toxicity, which indicates its negligible effect on the different bacterial strains; and a negative effect on the hERG (Ether-à-go-go-Related Gene), thereby unlikely to cause arrhythmia; and do not have hepatotoxic property. The oral acute toxicity of O-Demethyl-demethoxycurcumin was predicted to be 2.23 mol/kg, whereas, for Tellimagrandin II it was 2.48 mol/kg. Similarly, the oral rat chronic toxicity of O-Demethyldemethoxycurcumin was predicted to be 2.715 log mol/kg body weights per day, whereas, for Tellimagrandin II it was 10.618 log mol/kg body weight per day. Both these compounds also scored significant KDeep ΔG (absolute binding affinity) and MM-PBSA values.

CONCLUSION

The RBD of Spike protein is one of the major targets in the inhibition of SARS-CoV-2 and is the most sought-after target



being worked out across the globe. Some of the key residues which are involved in the entry and infection of the SARS-CoV-2 harbors on the RBD of the spike protein. Recently, glycosylation sites are also suggested to hold a key role in viral proliferation, as inferred by mutagenesis studies. Hence, in this study, virtual screening of phytochemical inhibitors targeting RBD domain was carried out, with a key emphasis on ACE2 binding residues along with glycosylation sites. Among the compounds studied, *Tellimagrandin-II* from *S. aromaticum* and *O-Demethyl-demethoxycurcumin* from *C. longa* were found to show stable interactions with key hotspot residues (ACE2 binding) including the glycosylation site. Molecular dynamics simulation of these compounds in complex with RBD also showed higher stability due to intermolecular interactions with active residues, significant binding free energy and optimal shape complementary during the entire production run. The results from this study clearly indicates that the proposed compounds may be considered as potential candidates for the inhibition of SARS-CoV-2 infection, as these are dual-acting in terms of inhibiting ACE2 interactions, as well as targeting the glycosylation of spike protein. However,

further experimental validations are warranted to infer the therapeutic efficacy.

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

VU and SD performed all the computational studies which included data collection, Molecular docking and Molecular dynamics simulation, and also prepared the manuscript. The experimental design and supervision was made by VU. DC conceptualized the whole study, provided inputs and compiled the write up of traditionally used medicinal plants along with their antiviral applications available in contemporary literature, revised the manuscript, and supervised the study. HH provided

initial inputs on the plants, while IS helped in the data collection. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

Authors would thank Professor Dr. Balram Bhargava, Secretary, Department of Health Research, Ministry of Health & Family

Welfare, Government of India and the Director General, Indian Council of Medical Research (ICMR) for the support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.672629/full#supplementary-material>

REFERENCES

- Wang W, Xu Y, Gao R, Lu R, Han K, Wu G, et al. Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA*. (2020) 323:1843–1844. doi: 10.1001/jama.2020.3786
- Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. (2020) 382:727–33. doi: 10.1056/NEJMoa2001017
- Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg Microb Infect*. (2020) 9:221–36. doi: 10.1080/22221751.2020.1719902
- WHO. WHO Media Briefing on COVID-19- 11 March 2020. (2020). Available online at: <https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19--11-march-2020> (accessed September 14, 2020).
- Worldometer (2021). Available online at: <https://www.worldometers.info/coronavirus/> (accessed April 02, 2021).
- Huang Y, Yang C, Xu X, Xu W, Liu S. Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacologica Sinica*. (2020) 41:1141–9. doi: 10.1038/s41401-020-0485-4
- Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*. (2020) 367:1260–3. doi: 10.1126/science.abb2507
- Hoffmann M, Müller MA, Drexler JF, Glende J, Erdt M, Gürtkow T, et al. Differential sensitivity of bat cells to infection by enveloped RNA viruses: coronaviruses, paramyxoviruses, filoviruses, and influenza viruses. *PLoS ONE*. (2013) 8:e72942. doi: 10.1371/journal.pone.0072942
- Hasan A, Paray BA, Hussain A, Qadir FA, Attar F, Aziz FM, et al. A review on the cleavage priming of the spike protein on coronavirus by angiotensin-converting enzyme-2 and furin. *J Biomol Struct Dyn*. (2020). 1–9. doi: 10.1080/07391102.2020.1754293
- Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veasler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell*. (2020) 181:281–92.e6. doi: 10.1016/j.cell.2020.02.058
- Wu Y, Wang F, Shen C, Peng W, Li D, Zhao C, et al. A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science*. (2020) 368:1274–8. doi: 10.1126/science.abc2241
- Ge X-Y, Li J-L, Yang X-L, Chmura AA, Zhu G, Epstein JH, et al. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. (2013) 503:535–8. doi: 10.1038/nature12711
- Gralinski LE, Menachery VD. Return of the coronavirus: 2019-nCoV. *Viruses*. (2020) 12:135. doi: 10.3390/v12020135
- Singh I, Vetrivel U, Harish DR, Chattopadhyay D. Coding-complete genome sequences of NITMA1086 and NITMA1139, two SARS-CoV-2 isolates from Belagavi District, Karnataka State, India, Harboring the D614G mutation. *Microbiol Resour Annu*. (2021) 10:e00016-21. doi: 10.1128/MRA.00016-21
- Shajahan A, Supekar NT, Gleinich AS, Azadi P. Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology*. (2020) 30:981–8. doi: 10.1093/glycob/cwaa042
- Watanabe Y, Allen JD, Wrapp D, McLellan JS, Crispin M. Site-specific glycan analysis of the SARS-CoV-2 spike. *Science*. (2020) 369:eabb9983. doi: 10.1126/science.abb9983
- Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature*. (2020) 581:221–4. doi: 10.1038/s41586-020-2179-y
- Enayatkhani M, Hasaniazad M, Faezi S, Gouklani H, Davoodian P, Ahmadi N, et al. Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against COVID-19: an *in silico* study. *J Biomol Struct Dyn*. (2020). doi: 10.1080/07391102.2020.1756411. [Epub ahead of print].
- Muralidharan N, Sakthivel R, Velmurugan D, Gromiha MM. Computational studies of drug repurposing and synergism of lopinavir, oseltamivir and ritonavir binding with SARS-CoV-2 protease against COVID-19. *J Biomol Struct Dyn*. (2020). doi: 10.1080/07391102.2020.1752802. [Epub ahead of print].
- Wahedi HM, Ahmad S, Abbasi SW. Stilbene-based natural compounds as promising drug candidates against COVID-19. *J Biomol Struct Dyn*. (2020). doi: 10.1080/07391102.2020.1762743. [Epub ahead of print].
- Parvathaneni V, Gupta V. Utilizing drug repurposing against COVID-19- efficacy, limitations, and challenges. *Life Sci*. (2020) 259:118275. doi: 10.1016/j.lfs.2020.118275
- Enmozhi SK, Raja K, Sebastine I, Joseph J. Andrographolide as a potential inhibitor of SARS-CoV-2 main protease: an *in silico* approach. *J Biomol Struct Dyn*. (2020). doi: 10.1080/07391102.2020.1760136. [Epub ahead of print].
- Elmezayen AD, Al-Obaidi A, Sahin AT, Yeleki K. Drug repurposing for coronavirus (COVID-19): *in silico* screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. *J Biomol Struct Dyn*. (2020). doi: 10.1080/07391102.2020.1758791. [Epub ahead of print].
- Naik SR, Bharadwaj P, Dingelstad N, Kalyanamoorthy S, Mandal SC, Ganesan A, et al. Structure-based virtual screening, molecular dynamics and binding affinity calculations of some potential phytochemicals against SARS-CoV-2. *J Biomol Struct Dyn*. (2021). doi: 10.1080/07391102.2021.1891969. [Epub ahead of print].
- Ansar S, Vetrivel U. Structure-based design of small molecule and peptide inhibitors for selective targeting of ROCK1: an integrative computational approach. *J Biomol Struct Dyn*. (2021). doi: 10.1080/07391102.2021.1898470. [Epub ahead of print].
- Nagarajan H, Vetrivel U. Membrane dynamics simulation and virtual screening reveals potential dual natural inhibitors of endothelin receptors for targeting glaucomatous condition. *Life Sci*. (2021) 269:119082. doi: 10.1016/j.lfs.2021.119082
- Sivashanmugam M, Sulochana KN, Umashankar V. Virtual screening of natural inhibitors targeting ornithine decarboxylase with pharmacophore scaffolding of DFMO and validation by molecular dynamics simulation studies. *J Biomol Struct Dyn*. (2019) 37:766–80. doi: 10.1080/07391102.2018.1439772
- Yi Y, Hsieh I-Y, Huang X, Li J, Zhao W. Glioblastoma stem-like cells: characteristics, microenvironment, and therapy. *Front Pharmacol*. (2016) 7:477. doi: 10.3389/fphar.2016.00477
- Yi F, Li L, Xu L, Meng H, Dong Y, Liu H, et al. In silico approach in reveal traditional medicine plants pharmacological material basis. *Chin Med*. (2018) 13:33. doi: 10.1186/s13020-018-0190-0
- Chattopadhyay D, Khan MTH. Ethnomedicines and ethnomedicinal phytochemicals against herpesviruses. *Biotechnol Annu Rev*. (2008) 14:297–348. doi: 10.1016/S1387-2656(08)00012-4
- Debprasad C. *Ethnomedicine: A Source of Complementary Therapeutics*. Trivandrum: Research Signpost (2010).

32. Chattopadhyay D, Naik T. Antivirals of ethnomedicinal origin: structure-activity relationship and scope. *Mini Rev Med Chem.* (2007) 7:275–301. doi: 10.2174/138955707780059844
33. Nair A, Chattopadhyay D, Saha B. Chapter 17 - plant-derived immunomodulators. In: Ahmad Khan MS, Ahmad I, Chattopadhyay D, editors. *New Look to Phytomedicine*. London: Academic Press (2019). p. 435–99. doi: 10.1016/B978-0-12-814619-4.00018-5
34. Das MA, Bhowmik P, Banerjee A, Das A, Ojha D, Chattopadhyay D. Chapter 3 - ethnomedicinal wisdom: an approach for antiviral drug development. In: Ahmad Khan MS, Ahmad I, Chattopadhyay D, editors. *New Look to Phytomedicine*. London: Academic Press (2019). p. 35–61. doi: 10.1016/B978-0-12-814619-4.00003-3
35. Chattopadhyay D, Sarkar MC, Chatterjee T, Sharma Dey R, Bag P, Chakraborti S, et al. Recent advancements for the evaluation of antiviral activities of natural products. *N Biotechnol.* (2009) 25:347–68. doi: 10.1016/j.nbt.2009.03.007
36. Chattopadhyay D, Ojha D, Mondal S, Goswami D. “Validation of antiviral potential of herbal ethnomedicine. In: *Evidence-Based Validation of Herbal Medicine* Elsevier (2015). p. 175–200. doi: 10.1016/B978-0-12-800874-4.00008-8
37. Fernandes MJB. Screening of Brazilian plants for antiviral activity against animal herpesviruses. *J Med Plants Res.* (2012) 6:2261–5. doi: 10.5897/JMPR10.040
38. Kuo Y-C, Lin L-C, Tsai W-J, Chou C-J, Kung S-H, Ho Y-H. Samarangenin B from limonium sinense suppresses herpes simplex virus type 1 replication in vero cells by regulation of viral macromolecular synthesis. *Antimicrob Agents Chemother.* (2002) 46:2854–64. doi: 10.1128/AAC.46.9.2854-2864.2002
39. Bag P, Ojha D, Mukherjee H, Halder UC, Mondal S, Biswas A, et al. A dihydro-pyrido-indole potentially inhibits HSV-1 infection by interfering the viral immediate early transcriptional events. *Antiv Res.* (2014) 105:126–34. doi: 10.1016/j.antiviral.2014.02.007
40. Bag P, Chattopadhyay D, Mukherjee H, Ojha D, Mandal N, Sarkar M, et al. Anti-herpes virus activities of bioactive fraction and isolated pure constituent of *Mallotus peltatus*: an ethnomedicine from Andaman Islands. *Virol J.* (2012) 9:98. doi: 10.1186/1743-422X-9-98
41. Ojha D, Mukherjee H, Ghosh S, Bag P, Mondal S, Chandra NS, et al. Evaluation of anti-infective potential of a tribal folklore *Odina woderi* Roxb against some selected microbes and herpes simplex virus associated with skin infection. *J Appl Microbiol.* (2013) 115:1317–28. doi: 10.1111/jam.12330
42. Cos P, Berghe D, Bruyne T, Vlietinck A. Plant substances as antiviral agents: an update (1997–2001). *Curr Organ Chem.* (2003) 7:1163–80. doi: 10.2174/1385272033486558
43. Ghosh S, Chattopadhyay D, Mandal A, Kaity S, Samanta A. Bioactivity guided isolation of antiinflammatory, analgesic, and antipyretic constituents from the leaves of *Pedilanthus tithymaloides* (L.). *Med Chem Res.* (2013) 22:4347–59. doi: 10.1007/s00044-012-0449-4
44. Ojha D, Das R, Sobia P, Dwivedi V, Ghosh S, Samanta A, et al. *Pedilanthus tithymaloides* inhibits HSV infection by modulating NF- κ B signaling. *PLoS ONE.* (2015) 10:e0139338. doi: 10.1371/journal.pone.0139338
45. Halder A, Das S, Ojha D, Chattopadhyay D, Mukherjee A. Highly monodispersed gold nanoparticles synthesis and inhibition of herpes simplex virus infections. *Mater Sci Eng C.* (2018) 89:413–21. doi: 10.1016/j.msec.2018.04.005
46. Goswami D, Das Mahapatra A, Banerjee S, Kar A, Ojha D, Mukherjee PK, et al. Boswellia serrata oleo-gum-resin and β -boswellic acid inhibits HSV-1 infection *in vitro* through modulation of NF- κ B and p38 MAP kinase signaling. *Phytomedicine.* (2018) 51:94–103. doi: 10.1016/j.phymed.2018.10.016
47. Ren J, Zhang A-H, Wang X-J. Traditional Chinese medicine for COVID-19 treatment. *Pharmacol Res.* (2020) 155:104743. doi: 10.1016/j.phrs.2020.104743
48. Shahzad A, Sharma S, Siddiqui SA editors. *Biotechnological Strategies for the Conservation of Medicinal and Ornamental Climbers*. Cham: Springer International Publishing (2016). doi: 10.1007/978-3-319-19288-8
49. Yang C-W, Lee Y-Z, Hsu H-Y, Shih C, Chao Y-S, Chang H-Y, et al. Targeting coronavirus replication and cellular JAK2 mediated dominant NF- κ B activation for comprehensive and ultimate inhibition of coronavirus activity. *Sci Rep.* (2017) 7:4105. doi: 10.1038/s41598-017-04203-9
50. Yang C-W, Lee Y-Z, Kang I-J, Barnard DL, Jan J-T, Lin D, et al. Identification of phenanthroindolizines and phenanthroquinolizidines as novel potent anti-coronaviral agents for porcine enteropathogenic coronavirus transmissible gastroenteritis virus and human severe acute respiratory syndrome coronavirus. *Antiv Res.* (2010) 88:160–8. doi: 10.1016/j.antiviral.2010.08.009
51. Yang CW, Chang HY, Hsu HY, Lee YZ, Chang HS, Chen IS, et al. Identification of anti-viral activity of the cardenolides, Na(+)/K(+)-ATPase inhibitors, against porcine transmissible gastroenteritis virus. *Toxicol Appl Pharmacol.* (2017) 332:129–37. doi: 10.1016/j.taap.2017.04.017
52. Yang C-W, Lee Y-Z, Hsu H-Y, Jan J-T, Lin Y-L, Chang S-Y, et al. Inhibition of SARS-CoV-2 by highly potent broad-spectrum anti-coronaviral tylophorine-based derivatives. *Front Pharmacol.* (2020) 11:606097. doi: 10.3389/fphar.2020.606097
53. Islam MT, Sarkar C, El-Kersh DM, Jamaddar S, Uddin SJ, Shilpi JA, et al. Natural products and their derivatives against coronavirus: a review of the non-clinical and pre-clinical data. *Phytother Res.* (2020) 34:2471–92. doi: 10.1002/ptr.6700
54. Fielding BC, da Silva Maia Bezerra Filho C, Ismail NSM, de Sousa DP. Alkaloids: therapeutic potential against human coronaviruses. *Molecules.* (2020) 25:5496. doi: 10.3390/molecules25235496
55. Schwarz S, Sauter D, Wang K, Zhang R, Sun B, Karioti A, et al. Kaempferol derivatives as antiviral drugs against the 3a channel protein of coronavirus. *Planta Med.* (2014) 80:177–82. doi: 10.1055/s-0033-1360277
56. Sharma V, Katiyar A, Agrawal RC. Glycyrrhiza glabra: chemistry and pharmacological activity. In: Mérillon JM, Ramawat KG, editors. *Sweeteners: Pharmacology, Biotechnology, and Applications*. Cham: Springer International Publishing (2017). p. 87–100. doi: 10.1007/978-3-319-27027-2_21
57. Khare, CP editor. *Indian Medicinal Plants*. New York, NY: Springer New York (2007). doi: 10.1007/978-0-387-70638-2
58. Wang L, Yang R, Yuan B, Liu Y, Liu C. The antiviral and antimicrobial activities of licorice, a widely-used Chinese herb. *Acta Pharmaceut Sin B.* (2015) 5:310–5. doi: 10.1016/j.apsb.2015.05.005
59. Cinatl J, Morgenstern B, Bauer G, Chandra P, Rabenau H, Doerr H. Glycyrrhizin, an active component of liquorice roots, and replication of SARS-associated coronavirus. *The Lancet.* (2003) 361:2045–6. doi: 10.1016/S0140-6736(03)13615-X
60. Chang H-M, But PP-H, Yao S-C, Wang L-L, Yeung SC-S. *Pharmacology and Applications of Chinese Materia Medica*. Singapore: World Scientific (1987). doi: 10.1142/0377
61. Ibrahim E, Mohamed A, Amin M, Emad-Eldin A, Dajem S, Bin SR, et al. Antiviral activity of liquorice powder extract against varicella zoster virus isolated from Egyptian patients. *Biomed J.* (2012) 35:231. doi: 10.4103/2319-4170.106149
62. Grienke U, Richter M, Braun H, Kirchmair J, Grafenstein S, von Liedl K, et al. New insights into the anti-influenza activity of licorice constituents. *Planta Med.* (2013) 79:38. doi: 10.1055/s-0033-1352073
63. Ashraf A, Ashraf MM, Rafiqe A, Aslam B, Galani S, Zafar S, et al. *In vivo* antiviral potential of *Glycyrrhiza glabra* extract against Newcastle disease virus. *Pakistan J Pharmac Sci.* (2017) 30:567–72.
64. Clark K, Grant P, Sarr A, Belakere J, Swaggerty C, Phillips T, et al. An *in vitro* study of theaflavins extracted from black tea to neutralize bovine rotavirus and bovine coronavirus infections. *Vet Microbiol.* (1998) 63:147–57. doi: 10.1016/S0378-1135(98)00242-9
65. Sharangi AB. Medicinal and therapeutic potentialities of tea (*Camellia sinensis* L.) – a review. *Food Res Int.* (2009) 42:529–35. doi: 10.1016/j.foodres.2009.01.007
66. Palit P, Chattopadhyay D, Thomas S, Kundu A, Kim HS, Rezaei N. Phytopharmaceuticals mediated Furin and TMPRSS2 receptor blocking: can it be a potential therapeutic option for Covid-19? *Phytomedicine.* (2021) 85:153396. doi: 10.1016/j.phymed.2020.153396
67. Sharma J, Kumar Bhardwaj V, Singh R, Rajendran V, Purohit R, Kumar S. An *in-silico* evaluation of different bioactive molecules of tea for their inhibition potency against non structural protein-15 of SARS-CoV-2. *Food Chem.* (2021) 346:128933. doi: 10.1016/j.foodchem.2020.128933

68. Bhardwaj VK, Singh R, Sharma J, Rajendran V, Purohit R, Kumar S. Identification of bioactive molecules from tea plant as SARS-CoV-2 main protease inhibitors. *J Biomol Struct Dyn.* (2020). doi: 10.1080/07391102.2020.1766572. [Epub ahead of print].
69. Sharma MP, Ahmad J, Hussain A, Khan S. Folklore medicinal plants of Mewat (Gurgaon District), Haryana, India. *Int J Pharmac.* (1992) 30:129–34. doi: 10.3109/13880209209053975
70. Singh B, Sharma RA. *Secondary Metabolites of Medicinal Plants*. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA (2020). doi: 10.1002/9783527825578
71. Amber R, Adnan M, Tariq A, Mussarat S. A review on antiviral activity of the Himalayan medicinal plants traditionally used to treat bronchitis and related symptoms. *J Pharm Pharmacol.* (2017) 69:109–22. doi: 10.1111/jphp.12669
72. Ahmed S, Gul S, Gul H, Bangash MH. Dual inhibitory activities of *Adhatoda vasica* against cyclooxygenase and lipoxygenase. *Int J Endors Health Sci Res.* (2013) 1:14. doi: 10.29052/IJEHSR.v1.i1.2013.14-17
73. Chavan R, Gohil D, Shah V, Kothari S, Chowdhary A. Antiviral activity of Indian medicinal plant *Justicia adhatoda* against herpes simplex virus: an *in-vitro* study. *Int J Pharm Bio Sci.* (2013) 4:769–78. Available online at: <https://ijpbs.net/abstract.php?article=Mjc2NQ>
74. Pattanayak P, Behera P, Das D, Panda S. *Ocimum sanctum* Linn. A reservoir plant for therapeutic applications: An overview. *Pharmacogn Rev.* (2010) 4:95. doi: 10.4103/0973-7847.65323
75. Ghoke SS, Sood R, Kumar N, Pateriya AK, Bhatia S, Mishra A, et al. Evaluation of antiviral activity of *Ocimum sanctum* and *Acacia arabica* leaves extracts against H9N2 virus using embryonated chicken egg model. *BMC Comple Alternat Med.* (2018) 18:174. doi: 10.1186/s12906-018-2238-1
76. Chiang L-C, Ng L-T, Cheng P-W, Chiang W, Lin C-C. Antiviral activities of extracts and selected pure constituents of *Ocimum basilicum*. *Clin Exp Pharmacol Physiol.* (2005) 32:811–6. doi: 10.1111/j.1440-1681.2005.04270.x
77. Cohen M. *Tulsi-Ocimum sanctum*: a herb for all reasons. *J Ayurveda Integr Med.* (2014) 5:251. doi: 10.4103/0975-9476.146554
78. Shree P, Mishra P, Selvaraj C, Singh SK, Chaube R, Garg N, et al. Targeting COVID-19 (SARS-CoV-2) main protease through active phytochemicals of ayurvedic medicinal plants *Withania somnifera* (Ashwagandha), *Tinospora cordifolia* (Giloy) and *Ocimum sanctum* (Tulsi) – a molecular docking study. *J Biomol Struct Dyn.* (2020) 1–14. doi: 10.1080/07391102.2020.1810778
79. Penna SC, Medeiros MV, Aimbire FSC, Faria-Neto HCC, Sertié JAA, Lopes-Martins RAB. Anti-inflammatory effect of the hydraalcoholic extract of *Zingiber officinale* rhizomes on rat paw and skin edema. *Phytomedicine.* (2003) 10:381–5. doi: 10.1078/0944-7113-00271
80. Karunakaran R, Sadanandan SP. Chapter 13 - *zingiber officinale*: antiinflammatory actions and potential usage for arthritic conditions. In: Watson RR, Preedy VR, editors. *Bioactive Food as Dietary Interventions for Arthritis and Related Inflammatory Diseases*. London: Academic Press (2019). p. 233–44. doi: 10.1016/B978-0-12-813820-5.00013-1
81. Abd El-Wahab A, El-Adawi H, El-Demellawy M. *In vitro* study of the antiviral activity of *Zingiber officinale*. *Planta Med.* (2009) 75:F7. doi: 10.1055/s-0029-1234649
82. Chang JS, Wang KC, Yeh CF, Shieh DE, Chiang LC. Fresh ginger (*Zingiber officinale*) has anti-viral activity against human respiratory syncytial virus in human respiratory tract cell lines. *J Ethnopharmacol.* (2013) 145:146–51. doi: 10.1016/j.jep.2012.10.043
83. Kaushik S, Jangra G, Kundu V, Yadav JP, Kaushik S. Anti-viral activity of *Zingiber officinale* (Ginger) ingredients against the Chikungunya virus. *Virus Dis.* (2020) 31:270–6. doi: 10.1007/s13337-020-00584-0
84. Koch C, Reichling J, Schneele J, Schnitzler P. Inhibitory effect of essential oils against herpes simplex virus type 2. *Phytomedicine.* (2008) 15:71–8. doi: 10.1016/j.phymed.2007.09.003
85. Denyer C V, Jackson P, Loakes DM, Ellis MR, Young DAB. Isolation of antirhinoviral sesquiterpenes from Ginger (*Zingiber officinale*). *J Natural Products.* (1994) 57:658–62. doi: 10.1021/np50107a017
86. Tung BT, Nham DT, Hai NT, Thu DK. Chapter 10 - curcuma longa, the polyphenolic curcumin compound and pharmacological effects on liver. In: Watson RR, Preedy VR, editors. *Bioactive Food as Dietary Interventions for Arthritis and Related Inflammatory Diseases*. London: Academic Press (2019). p. 125–34. doi: 10.1016/B978-0-12-814466-4.00010-0
87. Praditya D, Kirchhoff L, Brüning J, Rachmawati H, Steinmann J, Steinmann E. Anti-infective properties of the golden spice curcumin. *Front Microbiol.* (2019) 10:912. doi: 10.3389/fmicb.2019.00912
88. Manoharan Y, Haridas V, Vasanthakumar KC, Muthu S, Thavoorullah FF, Shetty P. Curcumin: a wonder drug as a preventive measure for COVID19 management. *Indian J Clin Biochem.* (2020) 35:373–5. doi: 10.1007/s12291-020-00902-9
89. Sornpet B, Potha T, Tragoolpua Y, Pringproa K. Antiviral activity of five Asian medicinal plant crude extracts against highly pathogenic H5N1 avian influenza virus. *Asian Pacific J Trop Med.* (2017) 10:871–6. doi: 10.1016/j.apjtm.2017.08.010
90. Dao TT, Nguyen PH, Won HK, Kim EH, Park J, Won BY, et al. Curcuminoids from *Curcuma longa* and their inhibitory activities on influenza A neuraminidases. *Food Chem. istry.* (2012) 134:21–8. doi: 10.1016/j.foodchem.2012.02.015
91. Shah B, Sheth F, Parabia, M. Folk herbal knowledge on the management of respiratory disorders prevailing in ethnic society of Valsad district, Gujarat. *Indian J Natural Products Resour.* (2012) 3:438–47. Available online at: <http://nopr.niscair.res.in/handle/123456789/14827>
92. El-Saber Batiha G, Alkazmi LM, Wasef LG, Beshbishy AM, Nadwa EH, Rashwan EK. *Syzygium aromaticum* L. (Myrtaceae): traditional uses, bioactive chemical constituents, pharmacological and toxicological activities. *Biomolecules.* (2020) 10:202. doi: 10.3390/biom10020202
93. Aboubakr HA, Nauertz A, Luong NT, Agrawal S, El-Sohaimy SAA, Youssef MM, et al. *In vitro* antiviral activity of clove and ginger aqueous extracts against feline calicivirus, a surrogate for human norovirus. *J Food Protect.* (2016) 79:1001–12. doi: 10.4315/0362-028X.JFP-15-593
94. Mehmood Y, Farooq U, Yousaf H, Riaz H, Mahmood RK, Nawaz A, et al. Antiviral activity of green silver nanoparticles produced using aqueous buds extract of *Syzygium aromaticum*. *Pakistan J Pharmac Sci.* (2020) 33:839–45. doi: 10.36721/PJPS.2020.33.2.SUP.839-845.1
95. Zeng X, Zhang P, Wang Y, Qin C, Chen S, He W, et al. CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Res.* (2019) 47:D1118–27. doi: 10.1093/nar/gky965
96. Zeng X, Zhang P, He W, Qin C, Chen S, Tao L, et al. NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Res.* (2018) 46:D1217–22. doi: 10.1093/nar/gkx1026
97. Shinbo Y, Nakamura Y, Altaf-UI-Amin M, Asahi H, Kurokawa K, Arita M, et al. KNApSACk: a comprehensive species-metabolite relationship database. In: *Plant Metabolomics*. Berlin; Heidelberg: Springer-Verlag (2006). p. 165–81. doi: 10.1007/3-540-29782-0_13
98. Berman HM. The protein data bank. *Nucleic Acids Res.* (2000) 28:235–42. doi: 10.1093/nar/28.1.235
99. Sarma P, Shekhar N, Prajapat M, Avti P, Kaur H, Kumar S, et al. *In-silico* homology assisted identification of inhibitor of RNA binding against 2019-nCoV N-protein (N terminal domain). *J Biomol Struct Dyn.* (2020). doi: 10.1080/07391102.2020.1753580. [Epub ahead of print].
100. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem.* (2009) 30:2785–91. doi: 10.1002/jcc.21256
101. Samdani A, Vetrivel U. POAP: a GNU parallel based multithreaded pipeline of open babel and AutoDock suite for boosted high throughput virtual screening. *Comput Biol Chem.* (2018) 74:39–48. doi: 10.1016/j.compbiolchem.2018.02.012
102. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. *J Cheminform.* (2011) 3:33. doi: 10.1186/1758-2946-3-33
103. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform.* (2018) 10:39. doi: 10.1186/s13321-018-0285-8
104. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. *Nucleic Acids Res.* (2015) 43:W443–7. doi: 10.1093/nar/gkv315
105. Pires DEV, Blundell TL, Ascher DB. pkCSM: predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures. *J Med Chem.* (2015) 58:4066–72. doi: 10.1021/acs.jmedchem.5b00104

106. DuBay KH, Hall ML, Hughes TF, Wu C, Reichman DR, Friesner RA. Accurate force field development for modeling conjugated polymers. *J Chem Theor Comput.* (2012) 8:4556–69. doi: 10.1021/ct300175w
107. Bernardes CES, Canongia Lopes JN, da Piedade MEM. All-atom force field for molecular dynamics simulations on organotransition metal solids and liquids. Application to $M(CO)_n$ ($M = Cr, Fe, Ni, Mo, Ru, or W$) compounds. *J Phys Chem A.* (2013) 117:11107–13. doi: 10.1021/jp407739h
108. Barth E, Kuczera K, Leimkuhler B, Skeel RD. Algorithms for constrained molecular dynamics. *J Comput Chem.* (1995) 16:1192–209. doi: 10.1002/jcc.540161003
109. Bulatov VV, Justo JF, Cai W, Yip S, Argon AS, Lenosky T, et al. Parameter-free modelling of dislocation motion: the case of silicon. *Philos Magaz A.* (2001) 81:1257–81. doi: 10.1080/01418610108214440
110. Petersen HG. Accuracy and efficiency of the particle mesh Ewald method. *J Chem Phys.* (1995) 103:3668–79. doi: 10.1063/1.470043
111. Harvey MJ, De Fabritiis G. An implementation of the smooth particle Mesh Ewald method on GPU hardware. *J Chem Theor Comput.* (2009) 5:2371–7. doi: 10.1021/ct900275y
112. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. *J Chem Phys.* (1984) 81:3684–90. doi: 10.1063/1.448118
113. Damm KL, Carlson HA. Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys J.* (2006) 90:4558–73. doi: 10.1529/biophysj.105.066654
114. Kufareva I, Abagyan R. Methods of protein structure comparison. *Methods Mol Biol.* (2012) 857:231–57. doi: 10.1007/978-1-61779-588-6_10
115. Fuglebakk E, Echave J, Reuter N. Measuring and comparing structural fluctuation patterns in large protein datasets. *Bioinformatics.* (2012) 28:2431–40. doi: 10.1093/bioinformatics/bts445
116. Lobanov MY, Bogatyreva NS, Galzitskaya OV. Radius of gyration as an indicator of protein structure compactness. *Mol Biol.* (2008) 42:623–8. doi: 10.1134/S0026893308040195
117. Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. K DEEP: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inform Model.* (2018) 58:287–96. doi: 10.1021/acs.jcim.7b00650
118. Poli G, Granchi C, Rizzolio F, Tuccinardi T. Application of MM-PBSA methods in virtual screening. *Molecules.* (2020) 25:1971. doi: 10.3390/molecules25081971
119. Wang Z, Wang X, Li Y, Lei T, Wang E, Li D, et al. farPPI: a webserver for accurate prediction of protein–ligand binding structures for small-molecule PPI inhibitors by MM/PB(GB)SA methods. *Bioinformatics.* (2019) 35:1777–9. doi: 10.1093/bioinformatics/bty879
120. Jena AB, Kanungo N, Nayak V, Chainy GBN, Dandapat J. Catechin and curcumin interact with S protein of SARS-CoV2 and ACE2 of human cell membrane: insights from computational studies. *Sci Rep.* (2021) 11:2043. doi: 10.1038/s41598-021-81462-7
121. Mohan RR, Wilson M, Gorham RD, Harrison RES, Morikis VA, Kieslich CA, et al. Virtual screening of chemical compounds for discovery of complement C3 ligands. *ACS Omega.* (2018) 3:6427–38. doi: 10.1021/acsomega.8b00606
122. Li Q, Wu J, Nie J, Zhang L, Hao H, Liu S, et al. The impact of mutations in SARS-CoV-2 spike on viral infectivity and antigenicity. *Cell.* (2020) 182:1284–94.e9. doi: 10.1016/j.cell.2020.07.012
123. Watanabe Y, Bowden TA, Wilson IA, Crispin M. Exploitation of glycosylation in enveloped virus pathobiology. *Biochim Biophys Acta.* (2019) 1863:1480–97. doi: 10.1016/j.bbagen.2019.05.012
124. Dalziel M, Crispin M, Scanlan CN, Zitzmann N, Dwek RA. Emerging principles for the therapeutic exploitation of glycosylation. *Science.* (2014) 343:1235681. doi: 10.1126/science.1235681
125. Watanabe Y, Berendsen ZT, Raghvani J, Seabright GE, Allen JD, Pybus OG, et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat Commun.* (2020) 11:2688. doi: 10.1038/s41467-020-16567-0
126. Yoshida T, Hatano T, Ito H, Okuda T. Chemical and biological perspectives of ellagitannin oligomers from medicinal plants. *Stud Natural Products Chem.* (2000) 23:395–453. doi: 10.1016/S1572-5995(00)80134-9
127. Chattopadhyay D. Ethnomedicinal antivirals: scope and opportunity. In: Ahmad I, Aqil F, Owais M, editors. *Modern Phytomedicine.* Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA (2006). p. 313–39. doi: 10.1002/9783527609987.ch15
128. Nonaka G-I, Harada M, Nishioka I. Eugeninin, a new ellagitannin from cloves. *Chem Pharmac Bull.* (1980) 28:685–7. doi: 10.1248/cpb.28.685
129. Takechi M, Tanaka Y. Purification and characterization of antiviral substance from the bud of *Syzygium aromaticum*. *Planta Med.* (1981) 42:69–74. doi: 10.1055/s-2007-971548
130. Kurokawa M, Hozumi T, Basnet P, Nakano M, Kadota S, Namba T, et al. Purification and characterization of eugeninin as an anti-herpesvirus compound from *Geum japonicum* and *Syzygium aromaticum*. *J Pharmacol Exp Therapeut.* (1998) 284:728–35.
131. Kurokawa M, Hozumi T, Tsurita M, Kadota S, Namba T, Shiraki K. Biological characterization of eugeninin as an anti-herpes simplex virus type 1 compound *in vitro* and *in vivo*. *J Pharmacol Exp Therapeut.* (2001) 297:372–9. Available online at: <https://europepmc.org/article/med/11259565>
132. Saleem HN, Batool F, Mansoor HJ, Shahzad-ul-Hussan S, Saeed M. Inhibition of dengue virus protease by eugeninin, isobiflorin, and biflorin isolated from the flower buds of *Syzygium aromaticum* (Cloves). *ACS Omega.* (2019) 4:1525–33. doi: 10.1021/acsomega.8b02861
133. Janyou A, Changtam C, Suksamrarn A, Tocharus C, Tocharus J. Suppression effects of O-demethylmethoxycurcumin on thapsigargin triggered on endoplasmic reticulum stress in SK-N-SH cells. *Neuro Toxicol.* (2015) 50:92–100. doi: 10.1016/j.neuro.2015.08.005
134. Almanza A, Carlesso A, Chintha C, Creedican S, Doultisinos D, Leuzzi B, et al. Endoplasmic reticulum stress signalling - from basic mechanisms to clinical applications. *FEBS J.* (2019) 286:241–78. doi: 10.1111/febs.14608
135. Aoe T. Pathological aspects of COVID-19 as a conformational disease and the use of pharmacological chaperones as a potential therapeutic strategy. *Front Pharmacol.* (2020) 11:1095. doi: 10.3389/fphar.2020.01095
136. Sureda A, Alizadeh J, Nabavi SF, Berindan-Neagoe I, Cismaru CA, Jeandet P, et al. Endoplasmic reticulum as a potential therapeutic target for covid-19 infection management? *Eur J Pharmacol.* (2020) 882:173288. doi: 10.1016/j.ejphar.2020.173288

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

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



Potential Prognostic Immune Biomarkers of Overall Survival in Ovarian Cancer Through Comprehensive Bioinformatics Analysis: A Novel Artificial Intelligence Survival Prediction System

Tingshan He[†], Liwen Huang[†], Jing Li[†], Peng Wang and Zhiqiao Zhang^{*}

Department of Infectious Diseases, Shunde Hospital, Southern Medical University, Guangzhou, China

OPEN ACCESS

Edited by:

George Priya Doss C,
VIT University, India

Reviewed by:

Pavan Gollapalli,
Nitte University, India
Umashankar Vetrivel,
Indian Council of Medical Research
(ICMR), India
Jiansong Fang,
Guangzhou University of Chinese
Medicine, China

*Correspondence:

Zhiqiao Zhang
sdgrxbk@163.com

[†]These authors share first authorship

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 27 August 2020

Accepted: 19 April 2021

Published: 24 May 2021

Citation:

He T, Huang L, Li J, Wang P and
Zhang Z (2021) Potential Prognostic
Immune Biomarkers of Overall Survival
in Ovarian Cancer Through
Comprehensive Bioinformatics
Analysis: A Novel Artificial Intelligence
Survival Prediction System.
Front. Med. 8:587496.
doi: 10.3389/fmed.2021.587496

Background: The tumour immune microenvironment plays an important role in the biological mechanisms of tumorigenesis and progression. Artificial intelligence medicine studies based on big data and advanced algorithms are helpful for improving the accuracy of prediction models of tumour prognosis. The current research aims to explore potential prognostic immune biomarkers and develop a predictive model for the overall survival of ovarian cancer (OC) based on artificial intelligence algorithms.

Methods: Differential expression analyses were performed between normal tissues and tumour tissues. Potential prognostic biomarkers were identified using univariate Cox regression. An immune regulatory network was constructed of prognostic immune genes and their highly related transcription factors. Multivariate Cox regression was used to identify potential independent prognostic immune factors and develop a prognostic model for ovarian cancer patients. Three artificial intelligence algorithms, random survival forest, multitask logistic regression, and Cox survival regression, were used to develop a novel artificial intelligence survival prediction system.

Results: The current study identified 1,307 differentially expressed genes and 337 differentially expressed immune genes between tumour samples and normal samples. Further univariate Cox regression identified 84 prognostic immune gene biomarkers for ovarian cancer patients in the model dataset (GSE32062 dataset and GSE53963 dataset). An immune regulatory network was constructed involving 63 immune genes and 5 transcription factors. Fourteen immune genes (PSMB9, FOXJ1, IFT57, MAL, ANXA4, CTSH, SCRN1, MIF, LTBR, CTSD, KIFAP3, PSMB8, HSPA5, and LTN1) were recognised as independent risk factors by multivariate Cox analyses. Kaplan-Meier survival curves showed that these 14 prognostic immune genes were closely related to the prognosis of ovarian cancer patients. A prognostic nomogram was developed by using these 14 prognostic immune genes. The concordance indexes were 0.760, 0.733, and 0.765 for 1-, 3-, and 5-year overall survival, respectively. This prognostic model

could differentiate high-risk patients with poor overall survival from low-risk patients. According to three artificial intelligence algorithms, the current study developed an artificial intelligence survival predictive system that could provide three individual mortality risk curves for ovarian cancer.

Conclusion: In conclusion, the current study identified 1,307 differentially expressed genes and 337 differentially expressed immune genes in ovarian cancer patients. Multivariate Cox analyses identified fourteen prognostic immune biomarkers for ovarian cancer. The current study constructed an immune regulatory network involving 63 immune genes and 5 transcription factors, revealing potential regulatory associations among immune genes and transcription factors. The current study developed a prognostic model to predict the prognosis of ovarian cancer patients. The current study further developed two artificial intelligence predictive tools for ovarian cancer, which are available at https://zhangzhigao8.shinyapps.io/Smart_Cancer_Survival_Predictive_System_17_OC_F1001/ and https://zhangzhigao8.shinyapps.io/Gene_Survival_Subgroup_Analysis_17_OC_F1001/. An artificial intelligence survival predictive system could help improve individualised treatment decision-making.

Keywords: ovarian cancer, overall survival, immune gene, transcription factor, prognostic signature

INTRODUCTION

Ovarian cancer (OC) is one of the most lethal malignant tumours in women, with 295,414 new cases and 184,799 deaths in 2018 (1). Although considerable progress has been made in diagnostic and therapeutic techniques, the 5-year survival rate of advanced OC patients remains poor (2). Early identification of patients with high mortality risk and more precise, individualised treatments will help improve the prognosis of OC patients. Regarding precision medicine, developing predictive models to provide early individualised mortality risk prediction and predicting the effectiveness of specific therapeutic schedules would be significant.

Considerable progress in bioinformatics helps scientists explore the intrinsic regulatory mechanisms of tumorigenesis and progression (3–6). The immune microenvironment plays an important role in the initiation and development of tumours (7, 8). Various studies have reported the clinical value of immunotherapy for ovarian cancer (5, 6). Several studies established prognostic models to predict the prognosis of OC patients (7, 8). However, regarding precision medicine, mortality risk prediction for high-risk and low-risk subgroups could not meet the needs of individualised treatment. Individualised treatment needs precise prognostic models to provide individual mortality risk prediction for a specific agent but not for a special subgroup.

Our team constructed two precision medicine predictive tools that predict individualised mortality risk for hepatocellular carcinoma (9, 10). These two precision medicine predictive tools

provide online mortality risk prediction that is convenient and easy to understand. More importantly, these precision medicine predictive tools provide individual and specific mortality risk prediction, which is important for individualised treatment decision-making. Recently, artificial intelligence based on big data and advanced algorithms has been used to improve the accuracy of predictive models for the diagnosis and prognosis in various tumours (11–13). Therefore, the current study aimed to build artificial intelligence predictive tools to predict individualised mortality risk for OC patients based on immune genes.

MATERIALS AND METHODS

Study Datasets

We retrieved the Gene Expression Omnibus (GEO) database according to the following conditions to obtain valuable research datasets: (1) The dataset should have available gene expression profile data; (2) The dataset should have complete clinicopathological data; (3) The dataset should have follow-up survival information. The GSE32062 dataset contained expression profiling data from 260 advanced-stage high-grade OC patients (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32062>). The GSE53963 dataset contained expression profiling data from 174 high-grade OC patients (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53963>). To eliminate the effect of death caused by non-tumour factors on the results of survival analysis, surviving patients with a survival time of <3 months were removed from the current study. Therefore, the GSE32062 dataset and GSE53963 dataset involved 420 patients, and 19,569 mRNAs were downloaded as model datasets for further survival. Probe IDs generated on the GPL6480 platform were converted to gene symbols based on Gencode

Abbreviations: OC, ovarian cancer; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; ROC, receiver operating characteristic; DFS, disease-free survival; HR, hazard ratio; CI, confidence interval; AJCC, American Joint Committee on Cancer; SD, standard deviation.

v29. The TCGA cohort contained 21,586 mRNAs and 370 OC patients as a validation dataset for survival. The gene count values were log2-transformed for the TCGA cohort. The flow chart of patient selection is shown in **Supplementary Figure 1**.

Differential Expression Analyses

We searched the GEO database to explore a dataset containing gene expression information of ovarian cancer samples and normal samples. The GSE26712 dataset was generated on the Affymetrix Human Genome U133A Array platform. The GSE26712 dataset has gene expression profiling information from 185 primary ovarian tumours and 10 normal ovarian surface epithelium (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26712>). Therefore, differential expression analyses were performed between 185 tumour samples and 10 normal samples (GSE26712). Cut-off values for differential expression analyses were \log_2 |fold change| > 1 and $P < 0.05$. The data were normalised using the trimmed mean of M values method with “edgeR” (14).

Immune Genes

The Immunology Database and Analysis Portal database were used to identify the immune gene list (15). Transcription factors were identified via the Cistrome Cancer database (16). Cytoscape v3.6.1 was used to develop an immune regulatory network of prognostic immune genes and their highly related transcription factors (11). Thresholds of |correlation coefficient| > 0.5 and $P < 0.01$ were used to identify transcription factors highly correlated with prognostic immune genes. The biological processes of immune genes were identified using the TISIDB database (<http://cis.hku.hk/TISIDB/index.php>).

Tumour Immune Infiltration

Associations among tumour infiltrating immune cells and immune genes were evaluated by the Tumour Immune Estimation Resource database (16). Twenty-eight tumour immune infiltration scores were generated by single sample gene set enrichment analysis (17, 18).

Statistical Analyses

Statistical analyses were conducted by SPSS Statistics 19.0 (SPSS Inc., USA). Artificial intelligence and bioinformatics analyses were performed using Python language 3.7.2 and R software 3.5.2 with the following artificial intelligence algorithms: random survival forest (RFS) algorithm (19, 20), multitask logistic regression (MTLR) algorithm (21, 22), and Cox survival regression algorithm (23, 24). The important packages included pec, rms, survival, rmda, ggplot2, GOpot, timereg, randomForestSRC, and riskRegression. The threshold for a statistically significant difference was a $P < 0.05$.

RESULTS

Study Datasets

The clinical information of the OC patients is shown in **Table 1**. There were 229 (61.9%) of 370 patients who died in the TCGA cohort (validation dataset), and 260 (61.9%) of 420 patients died

TABLE 1 | Clinical features of included patients.

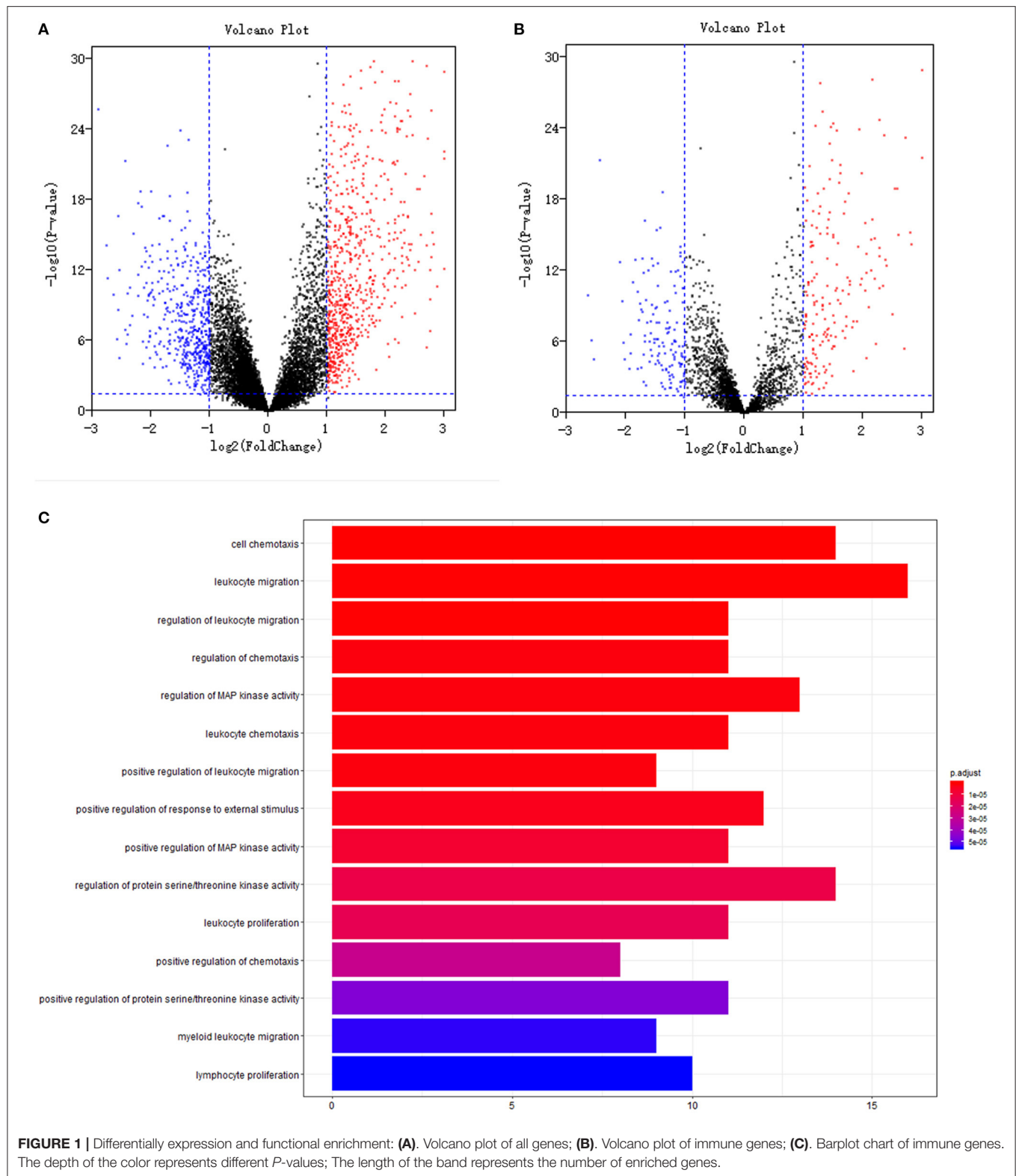
	TCGA dataset	GEO dataset	P-value
Number [n]	370	420	
Deaths [n (%)]	229 (61.9)	260 (61.9)	0.997
Total survival time (mean \pm SD, month)	39.2 \pm 31.0	47.9 \pm 37.3	<0.001
Survival time for dead patients (month)	38.7 \pm 25.7	36.8 \pm 28.3	0.430
Survival time for living patients (month)	40.0 \pm 38.1	66.0 \pm 42.7	<0.001
Age (mean \pm SD, year)	60.0 \pm 11.0	NA	
Grade 4 [n (%)]	1 (0.3)	74 (17.6)	0.615
Grade 3 [n (%)]	316 (85.4)	212 (50.5)	
Grade 2 [n (%)]	42 (11.4)	134 (31.9)	
Grade 1 [n (%)]	1 (0.3)	0	
Grade (NA) [n (%)]	10 (2.7)	0	
Stage 4 [n (%)]	57 (15.4)	93 (22.1)	<0.001
Stage 3 [n (%)]	289 (78.1)	320 (76.2)	
Stage 2 [n (%)]	20 (5.4)	7 (1.7)	
Stage 1 [n (%)]	1 (0.3)	0	
Stage (NA) [n (%)]	3 (0.8)	0	
Vascular invasion (positive) [n (%)]	63 (17.0)	NA	
Vascular invasion (negative) [n (%)]	39 (10.5)	NA	
Vascular invasion (NA) [n (%)]	268 (72.4)	NA	
Lymphovascular invasion (positive) [n (%)]	99 (26.8)	NA	
Lymphovascular invasion (negative) [n (%)]	46 (12.4)	NA	
Lymphovascular invasion (NA) [n (%)]	225 (60.8)	NA	
ECOG score (2–4) [n (%)]	6 (1.6)	NA	
ECOG score (1) [n (%)]	26 (7.0)	NA	
ECOG score (0) [n (%)]	53 (14.3)	NA	
ECOG score (NA) [n (%)]	285 (77.0)	NA	

Continuous variables were presented as mean \pm standard deviation; NA, missing data; AJCC, American Joint Committee on Cancer.

in the GEO cohort (model dataset). As shown in **Table 1**, there was no significant difference regarding mortality, survival time of deceased patients, or grade between the modelling cohort and the validation cohort during the follow-up period ($P > 0.05$). The overall survival time of all patients and the survival time of the patients in the survival subgroup for the model dataset were significantly longer than those of the validation dataset. However, the survival time of patients in the death subgroup of the model dataset was shorter than that of the patients in the death subgroup of the validated dataset, indicating that the difference in survival time between the two datasets might be related to the longer follow-up time of patients in the GEO cohort (model dataset).

Differential Expression Analyses

Volcano plots of 13,216 mRNAs and 3,075 immune genes are shown in **Figures 1A,B**. With a threshold of \log_2 |fold change| > 1 and $P < 0.05$, differential expression analysis identified 779 upregulated and 528 downregulated mRNAs from 13,216 mRNAs (**Figure 1A**) between 185 tumour samples and 10 normal samples (GSE26712 dataset). Differential expression analysis further identified 194 upregulated and 143 downregulated immune mRNAs from 3,075 immune mRNAs (**Figure 1B**)



between 185 tumour samples and 10 normal samples in the GSE26712 dataset.

To explore the gene expression difference of the identified immune biomarkers between the patients who died among the remaining patients with respect to the year of death, we further performed differential expression analysis between 130 tumour samples of patients who died and 10 normal samples of living patients (GSE26712 dataset). Differential expression analysis identified 753 upregulated and 526 downregulated mRNAs from 13,216 mRNAs. Differential expression analysis further identified 190 upregulated and 137 downregulated immune mRNAs from 3,075 immune mRNAs in the GSE26712 dataset.

Functional Enrichment Analyses

Further univariate Cox regression identified 84 prognostic immune gene biomarkers for OC patients in the model dataset (GSE32062 dataset and GSE53963 dataset). The bar plot (Figure 1C) and Gene Ontology chord chart (Figure 2) showed that the biological processes of the previous 84 prognostic immune genes were mainly enriched in leukocyte migration, cell chemotaxis, regulation of protein serine/threonine kinase activity, regulation of MAP kinase activity, positive regulation of response to external stimulus, regulation of leukocyte migration, regulation of chemotaxis, leukocyte chemotaxis, positive regulation of MAP kinase activity, and leukocyte proliferation. The results of the bar plot and Gene Ontology chord chart suggested that the above biological processes might play a role in the occurrence, growth, invasion, and prognosis of ovarian cancer, and the underlying mechanism is worthy of further study.

Immune Regulatory Network

Univariate Cox regression identified 84 prognostic immune biomarkers for the OS of OC patients. Transcription factors that were highly correlated with prognostic immune mRNAs were identified with previous correlation analysis thresholds. To explore the potential regulatory relationships among these immune genes, these previous prognostic immune mRNAs and their highly correlated transcription factors were placed in the STRING database with confidence values of 0.90. Thus, a regulatory network involving 63 immune genes and 5 transcription factors was constructed by using Cytoscape v3.6.1 (Figure 3). As shown in Figure 3, IRF4, GATA4, GATA3, CIITA, and MYH11 were involved in the immune regulatory network, indicating that these five transcription factors might play a role in the immune microenvironment of ovarian cancer.

Construction of a Prognostic Model

Multivariate Cox regression identified fourteen independent prognostic mRNAs for OS (Table 2 and Figure 4), indicating that these 14 prognostic immune genes might be more closely related to the prognosis of ovarian cancer than the prognostic immune genes that were not included in multivariate Cox regression. The formula of the prognostic model based on multivariate Cox regression was as follows: prognostic score = $(-0.472 \times \text{PSMB9}) + (-0.268 \times \text{FOXJ1}) + (0.303 \times \text{IFT57}) + (0.095 \times \text{MAL}) + (0.357 \times \text{ANXA4}) + (-0.339 \times \text{CTSH})$

$+ (0.422 \times \text{SCRN1}) + (-0.301 \times \text{MIF}) + (0.515 \times \text{LTBR}) + (-0.371 \times \text{CTSD}) + (0.503 \times \text{KIFAP3}) + (0.574 \times \text{PSMB8}) + (0.485 \times \text{HSPA5}) + (0.463 \times \text{LTN1})$. A prognostic nomogram is shown in Figure 5. For each prognostic gene, different gene expression values were assigned different risk scores. The total points (overall risk score) of one patient were obtained by adding up the risk scores of 14 prognostic genes. Through the vertical line corresponding to the total points, we can obtain the corresponding mortality rate of individual patients at different times.

Supplementary Figure 2 shows significant differences in survival curves between the high-risk group and the low-risk group. Eight immune factors (IFT57, MAL, ANXA4, SCR1, LTBR, KIFAP3, HSPA5, and LTN1) were positively correlated with poor prognosis of ovarian cancer, whereas six immune factors (PSMB9, FOXJ1, CTSH, MIF, CTSD, and PSMB8) were negatively correlated with poor prognosis of ovarian cancer. Supplementary Figures 3, 4 show the predictive value distribution chart and the survival status scatter plot.

Performance of Model Cohort

Survival curves of the two groups are illustrated in Figure 6A, showing that the mortality rate in the high-risk group was significantly higher than that in the low-risk group. Concordance indexes were 0.760, 0.733, and 0.765 for 1-, 3-, and 5-year survival, respectively (Figure 6B), indicating that the prognostic model has good predictive value for the prognosis of OC patients. Supplementary Figure 5 shows the calibration curves of the model cohort, showing that there was good consistency between the predicted mortality rate and the actual mortality rate.

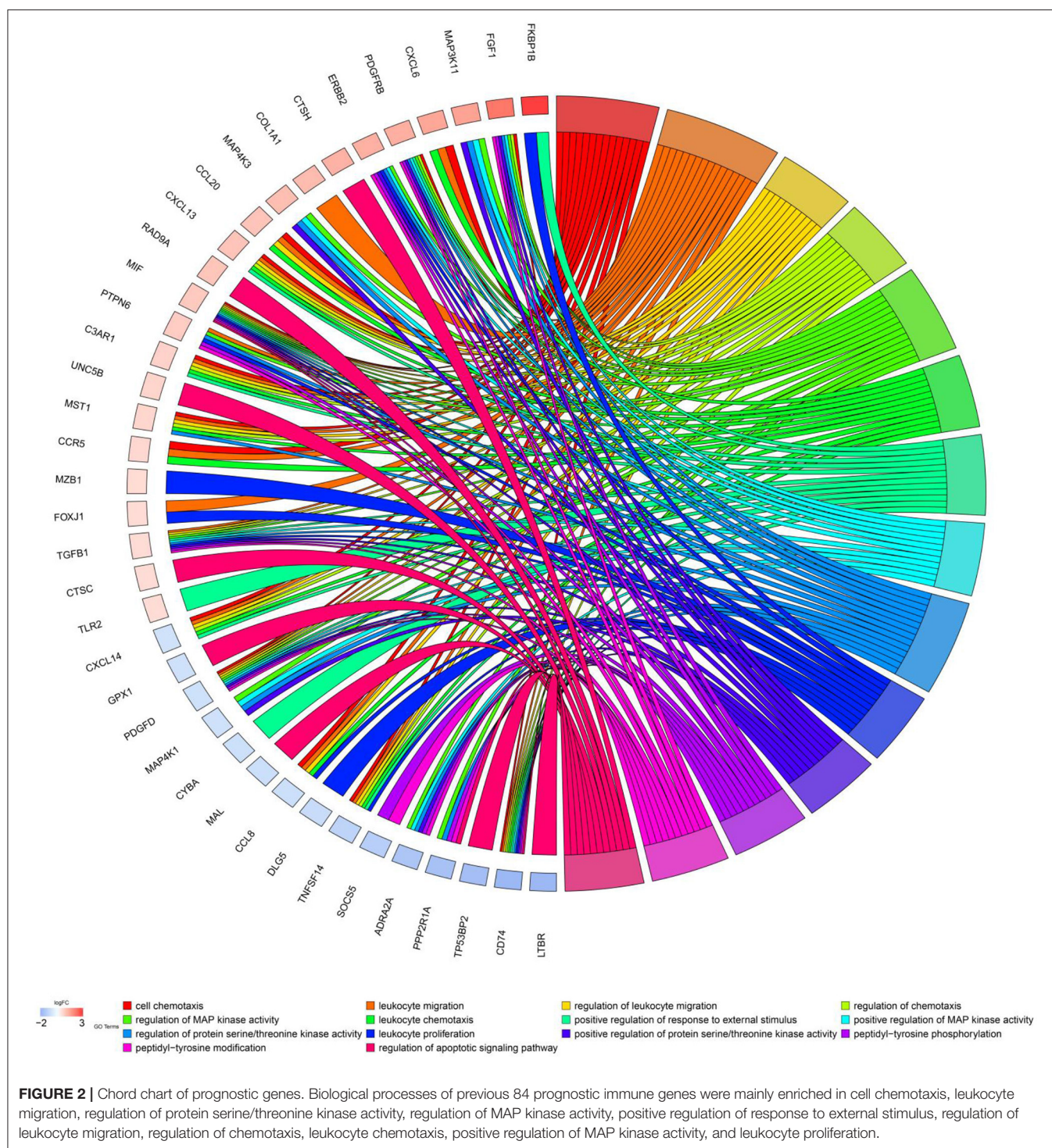
Performance of Validation Cohort

Survival curves of the two groups are illustrated in Figure 7A, showing that the mortality rate in the high-risk group was significantly higher than that in the low-risk group. Concordance indexes were 0.860, 0.715, and 0.679 for 1-, 3-, and 5-year survival, respectively (Figure 7B), indicating that the prognostic model has good predictive value for the prognosis of OC patients. Supplementary Figure 6 shows calibration curves of the validation cohort. Supplementary Figure 7 shows decision curves for 1-, 3-, and 5-year survival, showing that there was consistency between the predicted mortality rate and the actual mortality rate.

Artificial Intelligence Survival Predictive System

An artificial intelligence survival prediction system was constructed for individual mortality risk prediction for OC patients (Figure 8) and is available at https://zhangzhiqiao8.shinyapps.io/Smart_Cancer_Survival_Predictive_System_17_OC_F1001/. After the user inputs the expression values of the prognostic genes and clicks the “predict” button, the survival curve of one individual patient during the follow-up period will be presented.

The artificial intelligence survival prediction system provides three individual mortality risk predictive curves based on

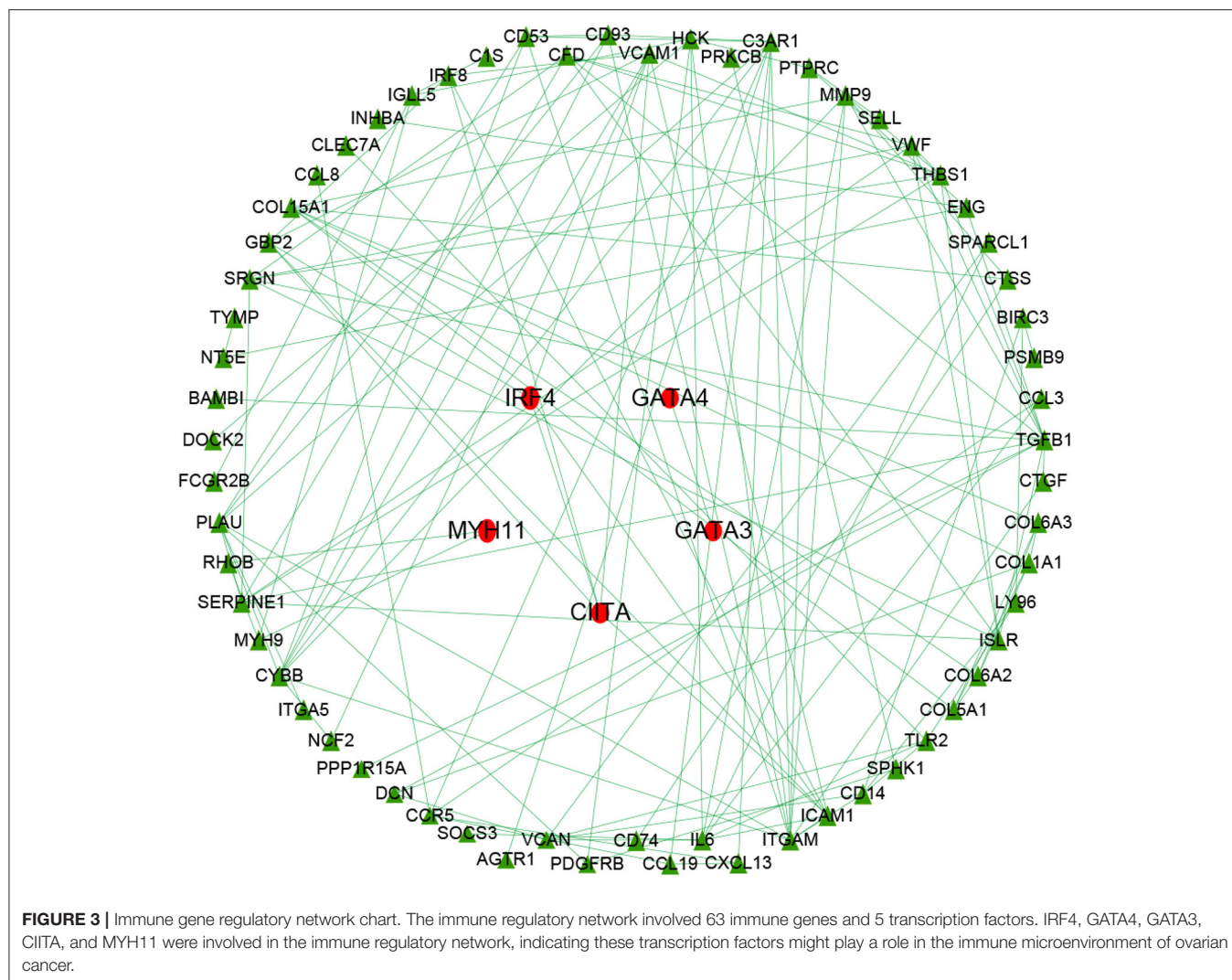


artificial intelligence algorithms: the RFS model (Figure 8A), MTLR model (Figure 8B), and Cox model (Figure 8C).

Gene Survival Analysis Screen System

A Gene Survival Analysis Screen System was constructed for exploratory research of immune genes (Supplementary Figure 8) and is available at <https://zhangz>

hiqiao8.shinyapps.io/Gene_Survival_Subgroup_Analysis_17_OC_F1001/. After the user inputs the parameters and clicks the “survival curve analysis” button, the survival curves of the high-risk group and low-risk group are presented. Users can obtain hazard ratio values of different clinical parameters after clicking the “Univariate Cox survival analysis table” button in the Gene Survival Analysis Screen System.



Independence Assessment

We used multivariate Cox regression to explore the independent effect of the prognostic model on the prognosis of OC patients. The prognostic signature was an independent influencing factor for OS in the model cohort (Table 3). In the validation cohort, the prognostic signature was an independent risk factor for OS. The results of multivariate Cox regression showed that the prognostic model had an independent effect on the prognosis of ovarian cancer, which further supported the value of the prognostic model in predicting ovarian cancer prognosis.

DISCUSSION

The current study identified 1,307 differentially expressed genes and 337 differentially expressed immune genes between tumour samples and normal samples. Further univariate Cox regression identified 84 prognostic immune gene biomarkers for OC patients in the model dataset (GSE32062 dataset

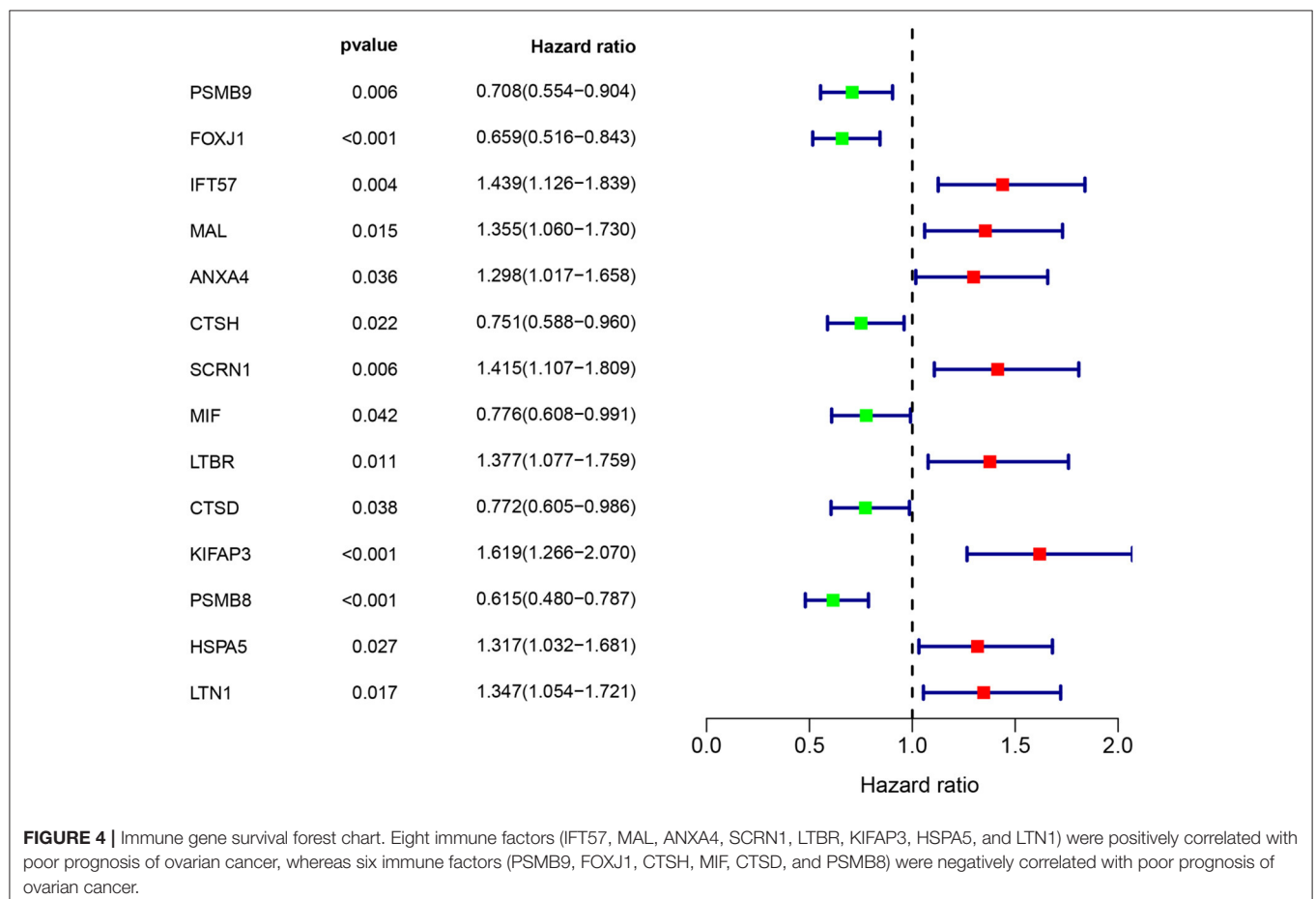
and GSE53963 dataset). An immune regulatory network was depicted involving 63 immune genes and 5 transcription factors. Through bioinformatics research, the current study depicted potential regulatory relationships among immune genes and transcription factors. Fourteen immune genes were identified as independent prognostic factors by multivariate survival analysis. Kaplan-Meier survival curves showed that these 14 prognostic genes were closely related to the prognosis of ovarian cancer patients. These 14 prognostic genes were used to develop a prognostic nomogram for ovarian cancer. Moreover, two artificial intelligence predictive tools were developed for precise individual mortality risk prediction in ovarian cancer. Based on a random survival forest algorithm, a multitask logistic regression algorithm, and a Cox survival regression algorithm, the current artificial intelligence survival predictive system provided three individual mortality risk predictive curves for the evaluation and improvement of individualised medical decisions.

In the current study, 1,308 differentially expressed genes (including 337 differential immune genes) were identified

TABLE 2 | Information of prognostic immune genes.

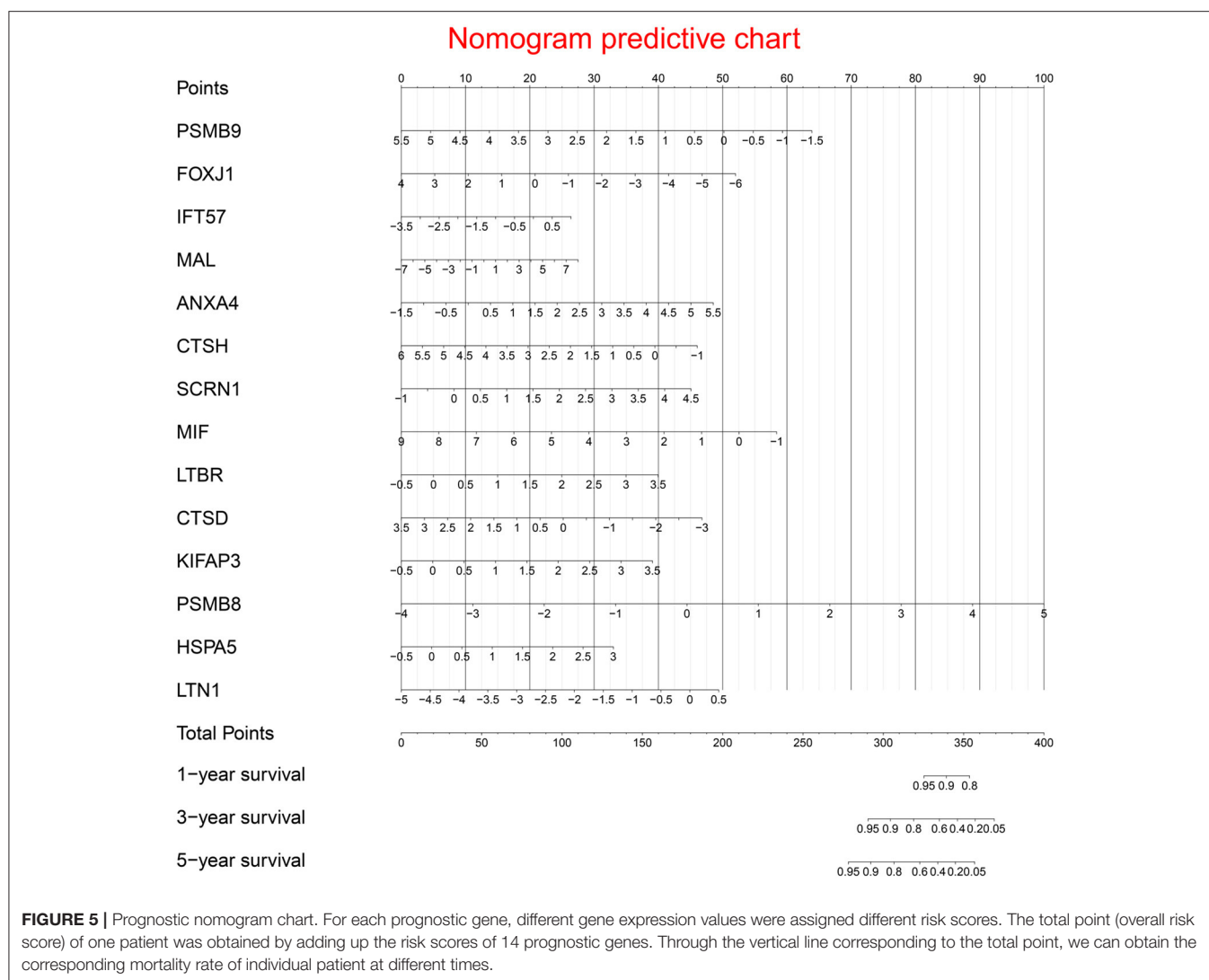
Immune gene	Univariate analysis			Multivariate analysis			
	HR	95% CI	P-value	Coefficient	HR	95% CI	P-value
PSMB9 (High/Low)	0.708	0.554–0.904	0.006	−0.472	0.624	0.449–0.867	0.005
FOXJ1 (High/Low)	0.659	0.516–0.843	<0.001	−0.268	0.765	0.672–0.871	<0.001
IFT57 (High/Low)	1.439	1.126–1.839	0.004	0.303	1.354	1.014–1.807	0.040
MAL (High/Low)	1.355	1.060–1.730	0.015	0.095	1.099	1.019–1.186	0.014
ANXA4 (High/Low)	1.298	1.017–1.658	0.036	0.357	1.430	1.123–1.820	0.004
CTSH (High/Low)	0.751	0.588–0.960	0.022	−0.339	0.712	0.565–0.898	0.004
SCRN1 (High/Low)	1.415	1.107–1.809	0.006	0.422	1.525	1.126–2.065	0.006
MIF (High/Low)	0.776	0.608–0.991	0.042	−0.301	0.740	0.602–0.910	0.004
LTBR (High/Low)	1.377	1.077–1.759	0.011	0.515	1.674	1.190–2.354	0.003
CTSD (High/Low)	0.772	0.605–0.986	0.038	−0.371	0.690	0.523–0.912	0.009
KIFAP3 (High/Low)	1.619	1.266–2.070	<0.001	0.503	1.653	1.189–2.298	0.003
PSMB8 (High/Low)	0.615	0.480–0.787	<0.001	0.574	1.775	1.165–2.703	0.008
HSPA5 (High/Low)	1.317	1.032–1.681	0.027	0.485	1.625	1.094–2.415	0.016
LTN1 (High/Low)	1.347	1.054–1.721	0.017	0.463	1.588	1.051–2.400	0.028

HR, hazard ratio; CI, confidence interval. The medians of gene expression values were used as cut-off values to stratify gene expression values into high expression group (as value 1) and low expression group (as value 0).



by differential expression analysis. Compared with normal ovarian tissues, these differentially expressed genes showed high expression or low expression in tumour tissues, suggesting

that these differentially expressed genes might be related to the biological characteristics and clinical process of OC. Further univariate Cox and multivariate Cox regression analyses



identified 84 and 14 prognostic immune genes, respectively, suggesting that these 14 prognostic immune genes might be closely related to the prognosis of OC patients. Functional enrichment analysis showed that the 84 genes were mainly related to the regulation of immune inflammation and were enriched in leukocyte migration, cell chemotaxis, regulation of protein serine/threonine kinase activity, and regulation of MAP kinase activity.

The immune regulatory network further indicated the potential regulatory relationship among 63 immune genes and 5 transcription factors, suggesting that these immune genes and transcription factors might play a potential role in the regulatory mechanism of the tumour immune environment. Previous studies have provided supporting evidence for the potential mechanisms of these five transcription factors regarding tumour growth, progression and prognosis. There is a close relationship between GATA3 and poor prognosis of high-grade serous ovarian carcinoma (25). GATA3 positivity is associated with poor prognosis of pancreatic ductal adenocarcinoma (26). High

expression of GATA3 is associated with good prognosis of ER+ breast cancer (27). IRF4 might activate the Notch-Akt signalling pathway in non-small cell lung cancer (28). Higher expression of IRF4+ Tregs was related to poor prognosis for different cancers (29). IRF4 was an independent prognostic factor for node-negative breast cancer (30). MYH11 positively modulated the immune-related gene GLP2R in colon adenocarcinoma (31). MYH11 positively regulated GSTM5, PTGIS, ENPP2, and P4HA3 (32). GATA4 inhibits tumour growth by affecting the assembly of tumour suppressor enhancement modules (33). Overexpression of GATA4 can protect human granulosa cell tumours from apoptosis induced by TRAIL *in vitro* (34).

Different research teams have established valuable survival prediction models for ovarian cancer based on different research cohorts and modelling methods. Previous prognostic models provided mortality curves for two classes of patients with different clinical characteristics (7, 8) but did not provide mortality curves for individual patients. He et al. constructed a prognostic model based on 10 RNA-binding proteins for

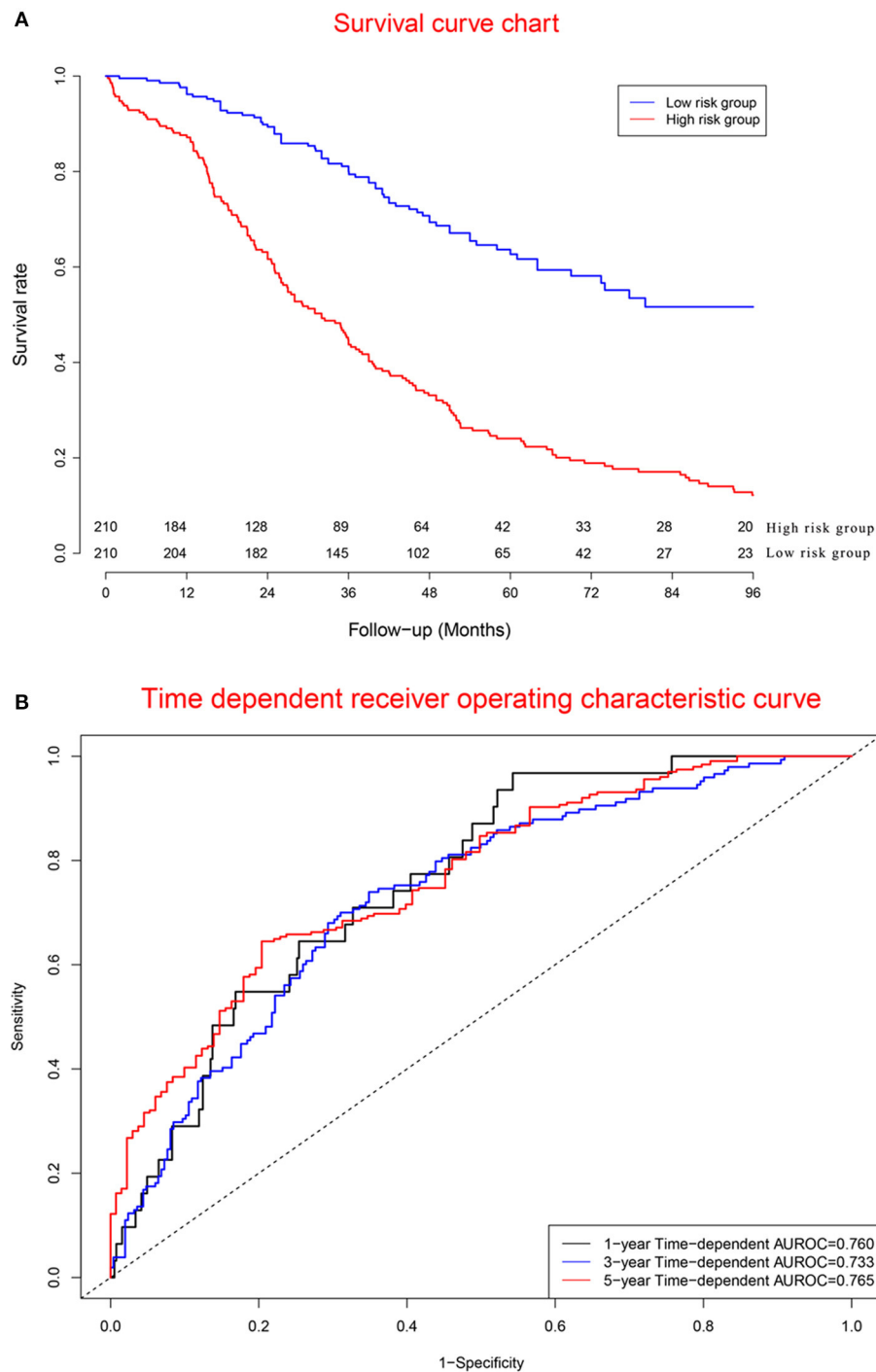


FIGURE 6 | Clinical performance in model cohort: **(A)**. Survival curves for high risk group and low risk group; **(B)**. Time-dependent receiver operating characteristic curves. The mortality rate in the high risk group was significantly higher than that in the low risk group. Concordance indexes were 0.760, 0.733, and 0.765 for 1-, 3-, and 5-year survival, indicating that the prognostic model has a good predictive value for the prognosis of ovarian cancer patients.

ovarian cancer (35). However, the calculation formula of this model is so complex that it is difficult for patients to calculate their personal risk score. Bing et al. constructed a novel model by merging three previous models selected by the integrated

P-value method, providing a new idea for the establishment of a prognostic model (36). However, this theoretically feasible method has not been applied in clinical research because it involves the fusion of multiple prognostic models. Tang et al.

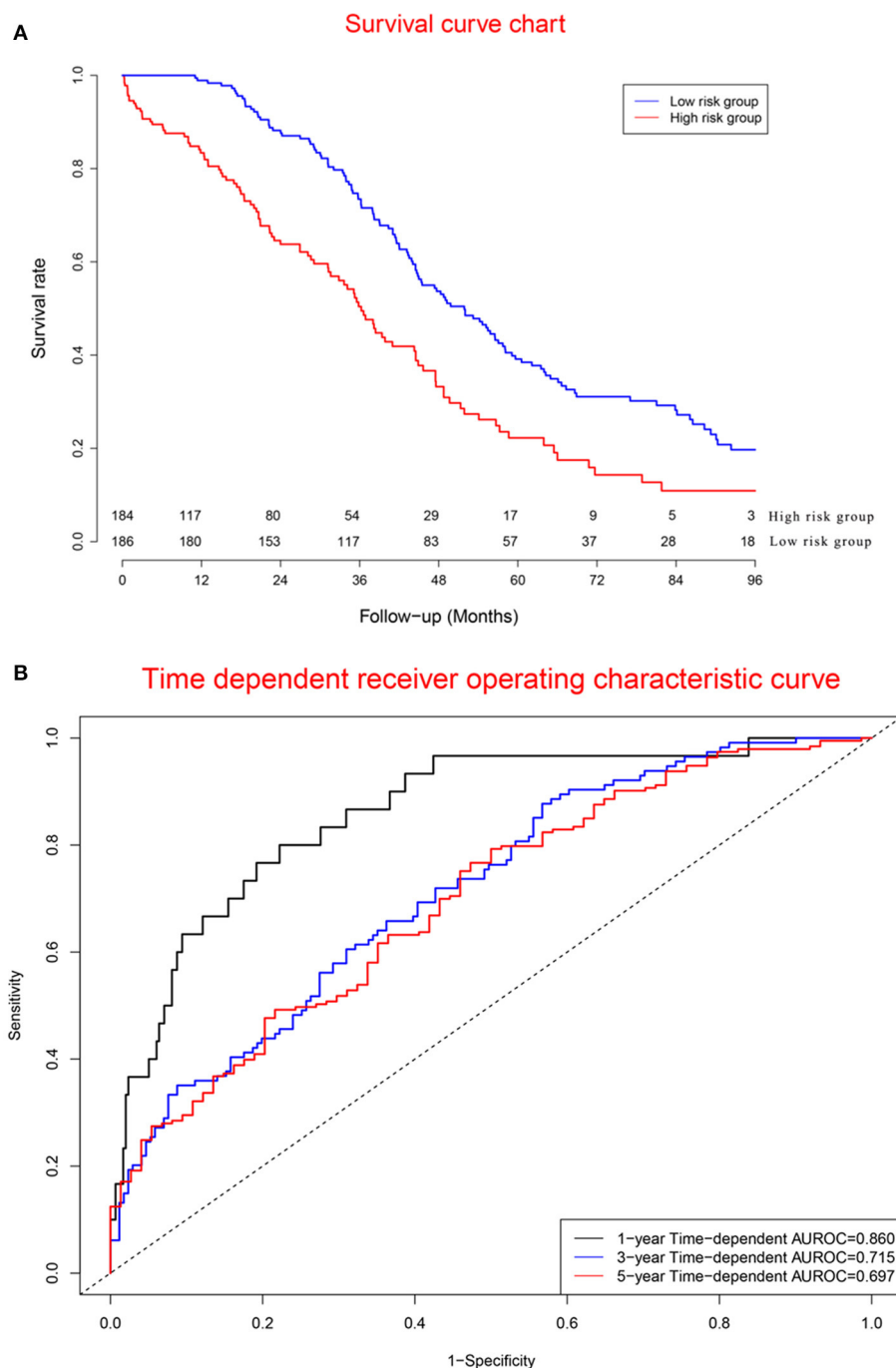


FIGURE 7 | Clinical performance in validation cohort: **(A)**, Survival curves for high risk group and low risk group; **(B)**, Time-dependent receiver operating characteristic curves. The mortality rate in the high risk group was significantly higher than that in the low risk group. Concordance indexes were 0.860, 0.715, and 0.679 for 1-, 3-, and 5-year survival, respectively **(B)**, indicating that the prognostic model has a good predictive value for the prognosis of ovarian cancer patients.

presented an eight-mRNA prognostic model for ovarian cancer (37), providing a valuable predictive model for clinical practise. If the above models can provide a simple calculation tool, it will be more helpful to provide convenient survival prediction information for patients with ovarian cancer. In fact, every

cancer patient cares only for her or his own individual mortality after diagnosis. Due to the considerable clinical heterogeneity of tumours, clinicians observe large differences in clinical prognosis among different cancer patients. Therefore, it is of great significance to predict the individual mortality risk of

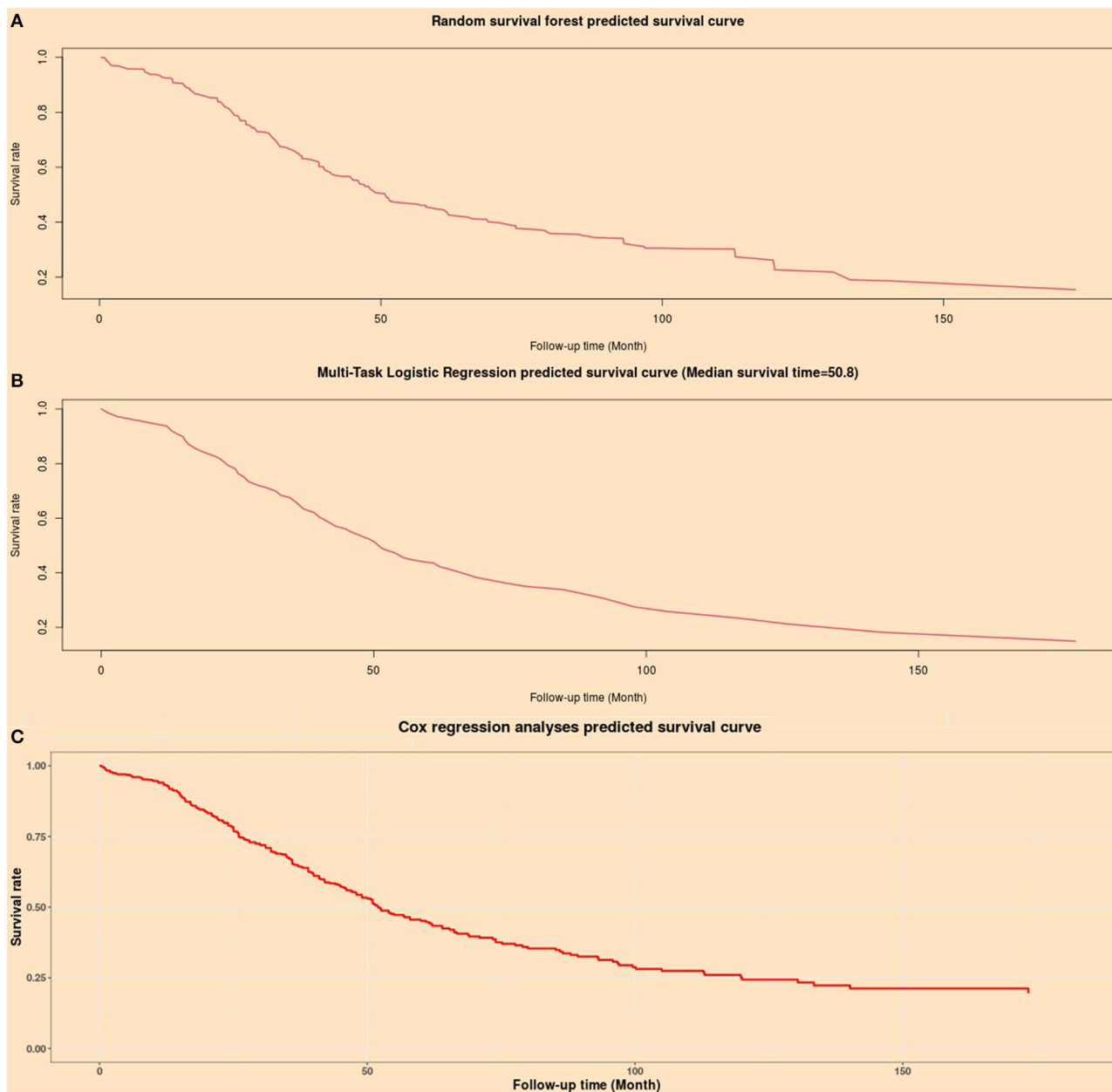


FIGURE 8 | Individual mortality risk predictive curves based on artificial intelligence algorithms. **(A)** Random survival forest model; **(B)** Multitask logistic regression model; **(C)** Cox proportional hazard regression model.

cancer patients. The emergence of big data and advanced algorithms has laid a solid foundation for artificial intelligence research. Different artificial intelligence algorithms have been used to improve clinical diagnosis and prognostic prediction (11–13). Based on the artificial intelligence algorithms provided in previous studies, the current study developed an artificial intelligence survival prediction system. The current artificial intelligence survival prediction system provides three individual mortality risk predictive curves according to different artificial intelligence algorithms. These artificial intelligence algorithms

are not widely used in clinical research because of the complexity of calculation. To the best of our knowledge, our team is the first to introduce various artificial intelligence algorithms for tumour prognosis research. Our study showed that artificial intelligence algorithms have great application value and superiority in predicting the individual mortality risk for cancer patients and are worth further research and application. The tumour immune microenvironment is reportedly related to oncogenesis and prognosis (7, 38). The current study revealed the potential association of tumour-infiltrating immune cells and immune

TABLE 3 | Results of Cox regression analyses.

	Univariate analysis			Multivariate analysis			
	HR	95% CI	P-value	Coefficient	HR	95% CI	P-value
Model cohort (n = 420)							
AJCC Stage (3–4/1–2)	1.913	0.783–4.679	0.155	1.000	2.718	1.111–6.646	0.028
AJCC Grade (3–4/1–2)	1.417	1.066–1.882	0.016	0.108	1.114	0.832–1.492	0.468
Prognostic model (High/Low)	2.988	2.294–3.893	<0.001	1.105	3.019	2.302–3.959	<0.001
Validation cohort (n = 370)							
AJCC Stage (3–4/1–2)	1.999	0.887–4.505	0.095	0.734	2.084	0.913–4.756	0.081
AJCC Grade (3–4/1–2)	1.209	0.808–1.812	0.356	0.046	1.048	0.695–1.579	0.824
Prognostic model (High/Low)	1.915	1.466–2.500	<0.001	0.658	1.930	1.472–2.531	<0.001

AJCC, the American Joint Committee on Cancer; HR, hazard ratio; CI, confidence interval. The median of Prognostic model scores was used as the cut-off value to stratify gastric cancer patients into high risk group and low risk group.

genes with tumour prognosis. Compared with several previous predictive models for the prognosis of OC patients (14, 39), our precision medical predictive tools were more valuable in providing individual mortality risk prediction at different time points.

The TISIDB database was used to explore the biological processes of immune genes. The top biological processes of proteasome subunit beta 9 (PSMB9) were immune response-activating signal transduction, the immune response-regulating signalling pathway, and the immune response-activating cell surface receptor signalling pathway. The top biological processes of Forkhead box J1 (FOXJ1) were adaptive immune responses, leucocyte-mediated immunity, humoral immune response mediated by circulating immunoglobulin, and lymphocyte-mediated immunity. The top biological processes of mal, T-cell differentiation protein (MAL) were the extrinsic apoptotic signalling pathway via death domain receptors, regulation of apoptotic signalling pathway, and the extrinsic apoptotic signalling pathway. The top biological processes of annexin A4 (ANXA4) were interleukin-8 production, regulation of interleukin-8 production, and negative regulation of interleukin-8 production. The top biological processes of cathepsin H (CTSH) were T cell-mediated immunity, lymphocyte-mediated immunity, leucocyte-mediated immunity, and adaptive immune response. The top biological processes of macrophage migration inhibitory factor (MIF) were negative regulation of immune system process, B cell homeostasis, regulation of immune effect or process, and lymphocyte homeostasis. The top biological processes of lymphotoxin beta receptor (LTBR) were myeloid dendritic cell activation, leucocyte differentiation, response to tumour necrosis factor, and response to molecules of bacterial origin. The top biological processes of cathepsin D (CTSD) were autophagy, antigen processing and presentation of exogenous antigen, antigen processing and presentation of exogenous peptide antigen via MHC class II. The top biological processes of kinesin-associated protein 3 (KIFAP3) were antigen processing and presentation, antigen processing and presentation of peptide antigen via MHC class II, and antigen processing and presentation of exogenous antigen. The top biological processes of proteasome

subunit beta 8 (PSMB8) were immune response-activating signal transduction, innate immune response-activating signal transduction, and the immune response-regulating cell surface receptor signalling pathway.

PSMB9, FOXJ1, IFT57, MAL, ANXA4, CTSH, SCRNI, MIF, LTBR, CTSD, KIFAP3, PSMB8, HSPA5, and LTN1 were recognised as independent risk factors by multivariate Cox analyses, suggesting that these 14 prognostic immune genes might have potential effects on the occurrence, progression and prognosis of tumours. NANOG controls cell migration and invasion by regulating FOXJ1 expression in ovarian cancer (15). FOXJ1 promoted tumour growth in bladder cancer (16). Highly expressed FOXJ1 promoted the proliferation and invasiveness of laryngeal squamous cell carcinoma cells (17). High expression of MAL was associated with poor survival of advanced ovarian cancer (40). Overexpression of the MAL gene was used to predict chemoresistance and poor prognosis in serous ovarian cancer patients (18). High expression of MAL promoted metastasis in colorectal cancer (24). Ikaros inhibited the proliferation of tumour cells by downregulating the expression of ANXA4 in hepatocellular carcinoma (23). Knockdown of SCRNI significantly reduced tumour cell growth in colorectal cancer (19). EIF expression was associated with overall survival in patients with ovarian cancer (20). The KIFAP3 gene is highly expressed at the mRNA and protein levels in breast cancer (41). miR-451a inhibited cancer growth and induced apoptosis of papillary thyroid cancer by targeting PSMB8 (41). The CpG mutation of PSMB9 is related to the recurrence or drug resistance of ovarian cancer after chemotherapy (42). High expression of PSMB8 and PSMB9 is related to the five-year survival of ovarian cancer (43). High expression of MIF is correlated with poor overall survival of ovarian cancer (44). HSPA5 inhibits the growth of epithelial ovarian cancer cells through G1 phase arrest (45). High expression of CD5L promoted proliferation and the antiapoptotic response in hepatocellular carcinoma cells by binding to HSPA5 (46).

CD4T helper cells can inhibit the transformation of immunosuppressive regulatory T cells in ovarian cancer (41). Regulatory T cells were positively correlated with ovarian cancer (20). An increased CD8/regulatory T cell ratio suggests good

prognosis for ovarian cancer (47). Dendritic cell immunotherapy could stimulate antitumour T cell immunity and improve the prognosis of cancer patients (21). Interleukin 10 regulates Toll-like receptor-mediated dendritic cell activation in ovarian cancer (22). IL-15 enhanced natural killer cell function in ovarian cancer patients (13). A low lymphocyte-to-monocyte ratio was related to poor survival in ovarian cancer (48). Mast cell infiltration with high mean vessel density indicated favourable prognosis in ovarian cancer (49). Macrophage secretory proteins induce ovarian cancer proliferation through the JAK2/STAT3 pathway (50). M1 macrophages induce ovarian cancer cell metastasis through the activation of NF- κ B (51). Small extracellular vesicles could inhibit the T cell response and promote the growth of ovarian cancer cells (51). Artesunate induced apoptosis of ovarian cancer cells by microRNA-142 (52). Mature neutrophils inhibited T cell immunity in ovarian cancer patients (50). Regulatory T cells inhibit CD8 T cell function through the IL-10 pathway (53). ISG15 induced CD8 T cells and inhibited the progression of ovarian cancer (54). TGF- β 1 induces CD8 Tregs through the p38 MAPK pathway in ovarian cancer (55). CD4 T helper cells inhibit the transformation of immunosuppressive regulatory T cells (56). CD4 T cells induce the host immune response through dendritic cells in patients with MHC class II-negative ovarian cancer (57).

Advantages: First, the current study developed two artificial intelligence predictive tools that provided individual mortality risk prediction at different time points and were valuable for optimising individual treatment decisions. Second, the current artificial intelligence survival predictive system provided three individual mortality risk predictive curves based on three artificial intelligence algorithms. Different artificial intelligence algorithms provided more reliable and valuable prognostic predictions for ovarian cancer than conventional prognostic models.

Shortcomings: First, because study datasets from public databases did not include information on surgical treatment, radiotherapy, biological targeting therapy, etc., the current study failed to assess the impact of these important clinical variables on survival. Second, from the perspective of model validity and extensibility, the sample size of the current research was relatively small for prognosis, which might weaken the validity of the research conclusions. Large, prospective sample studies can provide more convincing clinical evidence for the current study. Third, as non-parametric algorithms, artificial intelligence algorithms are complex to perform, and their calculation processes cannot be expressed by simple equations, restricting artificial intelligence algorithms as the mainstream methods for prognostic studies. Fourth, the current study constructed an immune regulatory network and revealed potential regulatory associations among immune genes and transcription factors. However, the role and mechanism of immune genes and transcription factors in tumorigenesis, growth and prognosis need to be elucidated by further study.

In conclusion, the current study identified 1,307 differentially expressed genes and 337 differentially expressed immune genes in ovarian cancer patients. Multivariate Cox analyses identified

fourteen prognostic immune biomarkers for ovarian cancer. The current study constructed an immune regulatory network involving 63 immune genes and 5 transcription factors, revealing potential regulatory associations among immune genes and transcription factors. The current study developed a prognostic model to predict the prognosis of ovarian cancer patients. The current research further developed two artificial intelligence predictive tools for ovarian cancer, which are available at https://zhangzhiqiao8.shinyapps.io/Smart_Cancer_Survival_Predictive_System_17_OC_F1001/ and https://zhangzhiqiao8.shinyapps.io/Gene_Survival_Subgroup_Analysis_17_OC_F1001/. The artificial intelligence survival predictive system can improve individualised treatment decision-making.

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

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ZZ, TH, PW, JL, and LH: conceptualisation, methodology, resources, investigation, data curation, formal analysis, validation, software, project administration, and supervision. ZZ and PW: writing and visualisation. ZZ: funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

The current research was funded by the Medical Science and Technology Foundation of Guangdong Province (A2016450 and B2018237).

ACKNOWLEDGMENTS

We would like to thank Dr. Gary S. Collins (University of Oxford), Dr. Manali Rupji (Emory University), and Mrs. Qingmei Liu for help and support in the development of precision medicine tools.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.587496/full#supplementary-material>

REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2018) 68:394–424. doi: 10.3322/caac.21492
- Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. *Gut.* (2017) 66:683–91. doi: 10.1136/gutjnl-2015-310912
- Zeng J, Cai X, Hao X, Huang F, He Z, Sun H, et al. LncRNA FUND2P4 down-regulation promotes epithelial-mesenchymal transition by reducing E-cadherin expression in residual hepatocellular carcinoma after insufficient radiofrequency ablation. *Int J Hyperthermia.* (2018) 34:802–11. doi: 10.1080/02656736.2017.1422030
- Zhong X, Long Z, Wu S, Xiao M, Hu W. LncRNA-SNHG7 regulates proliferation, apoptosis and invasion of bladder cancer cells assurance guidelines. *J BUON.* (2018) 23:776–81.
- Shi X, Zhao Y, He R, Zhou M, Pan S, Yu S, et al. Three-lncRNA signature is a potential prognostic biomarker for pancreatic adenocarcinoma. *Oncotarget.* (2018) 9:24248–59. doi: 10.18632/oncotarget.24443
- Huang Y, Xiang B, Liu Y, Wang Y, Kan H. LncRNA CDKN2B-AS1 promotes tumor growth and metastasis of human hepatocellular carcinoma by targeting let-7c-5p/NAP1L1 axis. *Cancer Lett.* (2018) 437:56–66. doi: 10.1016/j.canlet.2018.08.024
- Pagès F, Galon J, Dieu-Nosjean MC, Tartour E, Sautès-Fridman C, Fridman WH. Immune infiltration in human tumors: a prognostic factor that should not be ignored. *Oncogene.* (2010) 29:1093–102. doi: 10.1038/ncr.2009.416
- Domingues P, González-Tablas M, Otero, Pascual D, Miranda D, Ruiz L, et al. Tumor infiltrating immune cells in gliomas and meningiomas. *Brain Behav Immun.* (2016) 53:1–5. doi: 10.1016/j.bbi.2015.07.019
- Zhang Z, Li J, He T, Ouyang Y, Huang Y, Liu Q, et al. The competitive endogenous RNA regulatory network reveals potential prognostic biomarkers for overall survival in hepatocellular carcinoma. *Cancer Sci.* (2019) 110:2905–23. doi: 10.1111/cas.14138
- Zhang Z, Ouyang Y, Huang Y, Wang P, Li J, He T, et al. Comprehensive bioinformatics analysis reveals potential lncRNA biomarkers for overall survival in patients with hepatocellular carcinoma: an on-line individual risk calculator based on TCGA cohort. *Cancer Cell Int.* (2019) 19:174. doi: 10.1186/s12935-019-0890-2
- Tran WT, Jerzak K, Lu FI, Klein J, Tabbarah S, Lagree A, et al. Personalized breast cancer treatments using artificial intelligence in radiomics and pathomics. *J Med Imaging Radiat Sci.* (2019) 50(Suppl. 2):S32–41. doi: 10.1016/j.jmir.2019.07.010
- Nir G, Karimi D, Goldenberg SL, Fazli L, Skinnider BF, Tavassoli P, et al. Comparison of artificial intelligence techniques to evaluate performance of a classifier for automatic grading of prostate cancer from digitized histopathologic images. *JAMA Netw Open.* (2019) 2:e190442. doi: 10.1001/jamanetworkopen.2019.0442
- Enshaei A, Robson CN, Edmondson RJ. Artificial intelligence systems as prognostic and predictive tools in ovarian cancer. *Ann Surg Oncol.* (2015) 22:3970–75. doi: 10.1245/s10434-015-4475-6
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
- Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, et al. ImmPort: disseminating data to the public for the future of immunology. *Immunol Res.* (2014) 58:234–9. doi: 10.1007/s12026-014-8516-1
- Mei S, Meyer CA, Zheng R, Qin Q, Wu Q, Jiang P, et al. Cistrome cancer: a web resource for integrative gene regulation modeling in cancer. *Cancer Res.* (2017) 77:e19–22. doi: 10.1158/0008-5472.CAN-17-0327
- Jia Q, Wu W, Wang Y, Alexander PB, Sun C, Gong Z, et al. Local mutational diversity drives intratumoral immune heterogeneity in non-small cell lung cancer. *Nat Commun.* (2018) 9:5361. doi: 10.1038/s41467-018-07767-w
- Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* (2017) 18:248–62. doi: 10.1016/j.celrep.2016.12.019
- Xu H, Gu X, Tadesse MG, Balasubramanian R. A modified random survival forests algorithm for high dimensional predictors and self-reported outcomes. *J Comput Graph Stat.* (2018) 27:763–72. doi: 10.1080/10618600.2018.1474115
- Nasejje JB, Mwambi H. Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Res Notes.* (2017) 10:459. doi: 10.1186/s13104-017-2775-6
- Alaeddini A, Hong SH. A multi-way multi-task learning approach for multinomial logistic regression*. An application in joint prediction of appointment miss-opportunities across multiple clinics. *Methods Inform Med.* (2017) 56:294–307. doi: 10.3414/ME16-01-0112
- Bisaso KR, Karungi SA, Kiragga A, Mukonzo JK, Castelnovo B. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. *BMC Med Inform Decis Mak.* (2018) 18:77. doi: 10.1186/s12911-018-0659-x
- Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol.* (2018) 18:24. doi: 10.1186/s12874-018-0482-1
- Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Ann Rev Public Health.* (1999) 20:145–57. doi: 10.1146/annurev.publhealth.20.1.145
- El-Arabey AA, Denizli M, Kanlikilic P, Bayraktar R, Ivan C, Rashed M, et al. GATA3 as a master regulator for interactions of tumor-associated macrophages with high-grade serous ovarian carcinoma. *Cell Signal.* (2020) 68:109539. doi: 10.1016/j.cellsig.2020.109539
- Agostini-Vulaj D, Bratton LE, Dunne RF, Cates JMM, Zhou Z, Findeis-Hosey JJ, et al. Incidence and significance of GATA3 positivity in pancreatic ductal adenocarcinoma and cholangiocarcinoma. *Appl Immunohistochem Mol Morphol.* (2020) 28:460–63. doi: 10.1097/PAI.0000000000000764
- Fararjeh AS, Tu SH, Chen LC, Liu YR, Lin YK, Chang HL, et al. The impact of the effectiveness of GATA3 as a prognostic factor in breast cancer. *Hum Pathol.* (2018) 80:219–30. doi: 10.1016/j.humpath.2018.06.004
- Qian Y, Du Z, Xing Y, Zhou T, Chen T, Shi M. Interferon regulatory factor 4 (IRF4) is overexpressed in human non-small cell lung cancer (NSCLC) and activates the Notch signaling pathway. *Mol Med Rep.* (2017) 16:6034–40. doi: 10.3892/mmr.2017.7319
- Alvisi G, Brummelman J, Puccio S, Mazza EM, Tomada EP, Losurdo A, et al. IRF4 instructs effector Treg differentiation and immune suppression in human cancer. *J Clin Invest.* (2020) 130:3137–50. doi: 10.1172/JCI130426
- Heimes AS, Madjar K, Edlund K, Battista MJ, Almstedt K, Gebhard S, et al. Prognostic significance of interferon regulating factor 4 (IRF4) in node-negative breast cancer. *J Cancer Res Clin Oncol.* (2017) 143:1123–31. doi: 10.1007/s00432-017-2377-7
- Sun YL, Zhang Y, Guo YC, Yang ZH, Xu YC. A prognostic model based on the immune-related genes in colon adenocarcinoma. *Int J Med Sci.* (2020) 17:1879–96. doi: 10.7150/ijms.45813
- Sun YL, Zhang Y, Guo YC, Yang ZH, Xu YC. A prognostic model based on six metabolism-related genes in colorectal cancer. *Biomed Res Int.* (2020) 2020:5974350. doi: 10.1155/2020/5974350
- Lu F, Zhou Q, Liu L, Zeng G, Ci W, Liu W, et al. A tumor suppressor enhancing module orchestrated by GATA4 denotes a therapeutic opportunity for GATA4 deficient HCC patients. *Theranostics.* (2020) 10:484–97. doi: 10.7150/thno.38060
- Kyrönlähti A, Kauppinen M, Lind E, Unkila-Kallio L, Butzow R, Klefström J, et al. GATA4 protects granulosa cell tumors from TRAIL-induced apoptosis. *Endocr Relat Cancer.* (2010) 17:709–17. doi: 10.1677/ERC-10-0041
- He C, Huang F, Zhang K, Wei J, Hu K, Liang M. Establishment and validation of an RNA binding protein-associated prognostic model for ovarian cancer. *J Ovarian Res.* (2021) 14:27. doi: 10.1186/s13048-021-00777-1
- Bing Z, Yao Y, Xiong J, Tian J, Guo X, Li X, et al. Novel model for comprehensive assessment of robust prognostic gene signature in ovarian cancer across different independent datasets. *Front Genet.* (2019) 10:931. doi: 10.3389/fgene.2019.00931

37. Tang W, Li J, Chang X, Jia L, Tang Q, Wang Y, et al. Construction of a novel prognostic-predicting model correlated to ovarian cancer. *Biosci Rep.* (2020) 40:BSR20201261. doi: 10.1042/BSR20201261
38. Gough MJ, Crittenden MR. Immune system plays an important role in the success and failure of conventional cancer therapy. *Immunotherapy.* (2012) 4:125–8. doi: 10.2217/imt.11.157
39. Cheng C, Wang Q, Zhu M, Liu K, Zhang Z. Integrated analysis reveals potential long non-coding RNA biomarkers and their potential biological functions for disease free survival in gastric cancer patients. *Cancer Cell Int.* (2019) 19:123. doi: 10.1186/s12935-019-0846-6
40. Berchuck A, Iversen ES, Luo J, Clarke JP, Horne H, Levine DA, et al. Microarray analysis of early stage serous ovarian cancers shows profiles predictive of favorable outcome. *Clin Cancer Res.* (2009) 15:2448–55. doi: 10.1158/1078-0432.CCR-08-2430
41. Hsieh E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ Cardiovasc Qual Outcomes.* (2011) 4:39–45. doi: 10.1161/CIRCOUTCOMES.110.939371
42. Zeller C, Dai W, Steele NL, Siddiq A, Walley AJ, Wilhelm-Benartzi CS, et al. Candidate DNA methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling. *Oncogene.* (2012) 31:4567–76. doi: 10.1038/ncr.2011.611
43. Fejzo MS, Chen HW, Anderson L, McDermott MS, Karlan B, Konecny GE, et al. Analysis in epithelial ovarian cancer identifies KANSL1 as a biomarker and target gene for immune response and HDAC inhibition. *Gynecol Oncol.* (2021) 160:539–46. doi: 10.1016/j.ygyno.2020.11.008
44. Tas F, Karabulut S, Serilmez M, Ciftci R, Duranyildiz D. Serum levels of macrophage migration-inhibitory factor (MIF) have diagnostic, predictive and prognostic roles in epithelial ovarian cancer patients. *Tumour Biol.* (2014) 35:3327–31. doi: 10.1007/s13277-013-1438-z
45. Sethi G, Pathak HB, Zhang H, Zhou Y, Einarson MB, Vathipadiekal V, et al. An RNA interference lethality screen of the human druggable genome to identify molecular vulnerabilities in epithelial ovarian cancer. *PLoS ONE.* (2012) 7:e47086. doi: 10.1371/journal.pone.0047086
46. Ruyssinck J, van der Herten J, Houthoofd R, Ongenaes F, Couckuyt I, Gadeyne B, et al. Random survival forests for predicting the bed occupancy in the intensive care unit. *Comput Math Methods Med.* (2016) 2016:7087053. doi: 10.1155/2016/7087053
47. Hamidi O, Poorolajal J, Farhadian M, Tapak L. Identifying important risk factors for survival in kidney graft failure patients using random survival forests. *Iran J Public Health.* (2016) 45:27–33.
48. Shi M, Xu G. Development and validation of GMI signature based random survival forest prognosis model to predict clinical outcome in acute myeloid leukemia. *BMC Med Genomics.* (2019) 12:90. doi: 10.1186/s12920-019-0540-5
49. Wang H, Liu D, Yang J. Prognostic risk model construction and molecular marker identification in glioblastoma multiforme based on mRNA/microRNA/long non-coding RNA analysis using random survival forest method. *Neoplasma.* (2019) 66:459–69. doi: 10.4149/neo_2018_181008N746
50. Adham D, Abbasgholizadeh N, Abazari M. Prognostic factors for survival in patients with gastric cancer using a random survival forest. *Asian Pacific J Cancer Prev.* (2017) 18:129–34. doi: 10.22034/APJCP.2017.18.1.129
51. Wang H, Li G. A selective review on random survival forests for high dimensional data. *Quant Biosci.* (2017) 36:85–96. doi: 10.22283/qbs.2017.36.2.85
52. Wang H, Shen L, Geng J, Wu Y, Xiao H, Zhang F, et al. Prognostic value of cancer antigen–125 for lung adenocarcinoma patients with brain metastasis: a random survival forest prognostic model. *Sci Rep.* (2018) 8:5670. doi: 10.1038/s41598-018-23946-7
53. Liang C, Zhang Y, Zhang Y, Li R, Wang Z, Wei Z, et al. The prognostic value of LINC01296 in pan-cancers and the molecular regulatory mechanism in hepatocellular carcinoma: a comprehensive study based on data mining, bioinformatics, and *in vitro* validation. *Oncotargets Ther.* (2019) 12:5861–85. doi: 10.2147/OTT.S205853
54. Kontos CK, Papadopoulos IN, Scorilas A. Quantitative expression analysis and prognostic significance of the novel apoptosis-related gene BCL2L12 in colon cancer. *Biol Chem.* (2008) 389:1467–75. doi: 10.1515/BC.2008.173
55. Malietzis G, Lee GH, Bernardo D, Blakemore AI, Knight SC, Moorghen M, et al. The prognostic significance and relationship with body composition of CCR7-positive cells in colorectal cancer. *J Surg Oncol.* (2015) 112:86–92. doi: 10.1002/jso.23959
56. Tampakis A, Tampaki EC, Nonni A, Tsourouflis G, Posabella A, Patsouris E, et al. L1CAM expression in colorectal cancer identifies a high-risk group of patients with dismal prognosis already in early-stage disease. *Acta Oncol.* (2019) 59:55–9. doi: 10.1080/0284186X.2019.1667022
57. Liang L, Zhao K, Zhu JH, Chen G, Qin XG, Chen JQ. Comprehensive evaluation of FKBP10 expression and its prognostic potential in gastric cancer. *Oncol Rep.* (2019) 42:615–28. doi: 10.3892/or.2019.7195

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

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



Bioactive Molecules of Tea as Potential Inhibitors for RNA-Dependent RNA Polymerase of SARS-CoV-2

Vijay Kumar Bhardwaj^{1,2,3}, Rahul Singh^{1,2}, Jatin Sharma^{1,2}, Vidya Rajendran²,
Rituraj Purohit^{1,2,3*} and Sanjay Kumar²

¹ Structural Bioinformatics Lab, CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT), Palampur, India,

² Biotechnology Division, CSIR-IHBT, Palampur, India, ³ Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

OPEN ACCESS

Edited by:

D. Thirumal Kumar,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Umashankar Vetrivel,
Indian Council of Medical Research
(ICMR), India
Pranshu Sahgal,
Dana-Farber Cancer Institute,
United States
Atrayee Bhattacharya,
Dana-Farber Cancer Institute,
United States

*Correspondence:

Rituraj Purohit
rituraj@ihbt.res.in;
riturajpurohit@gmail.com

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 22 March 2021

Accepted: 16 April 2021

Published: 31 May 2021

Citation:

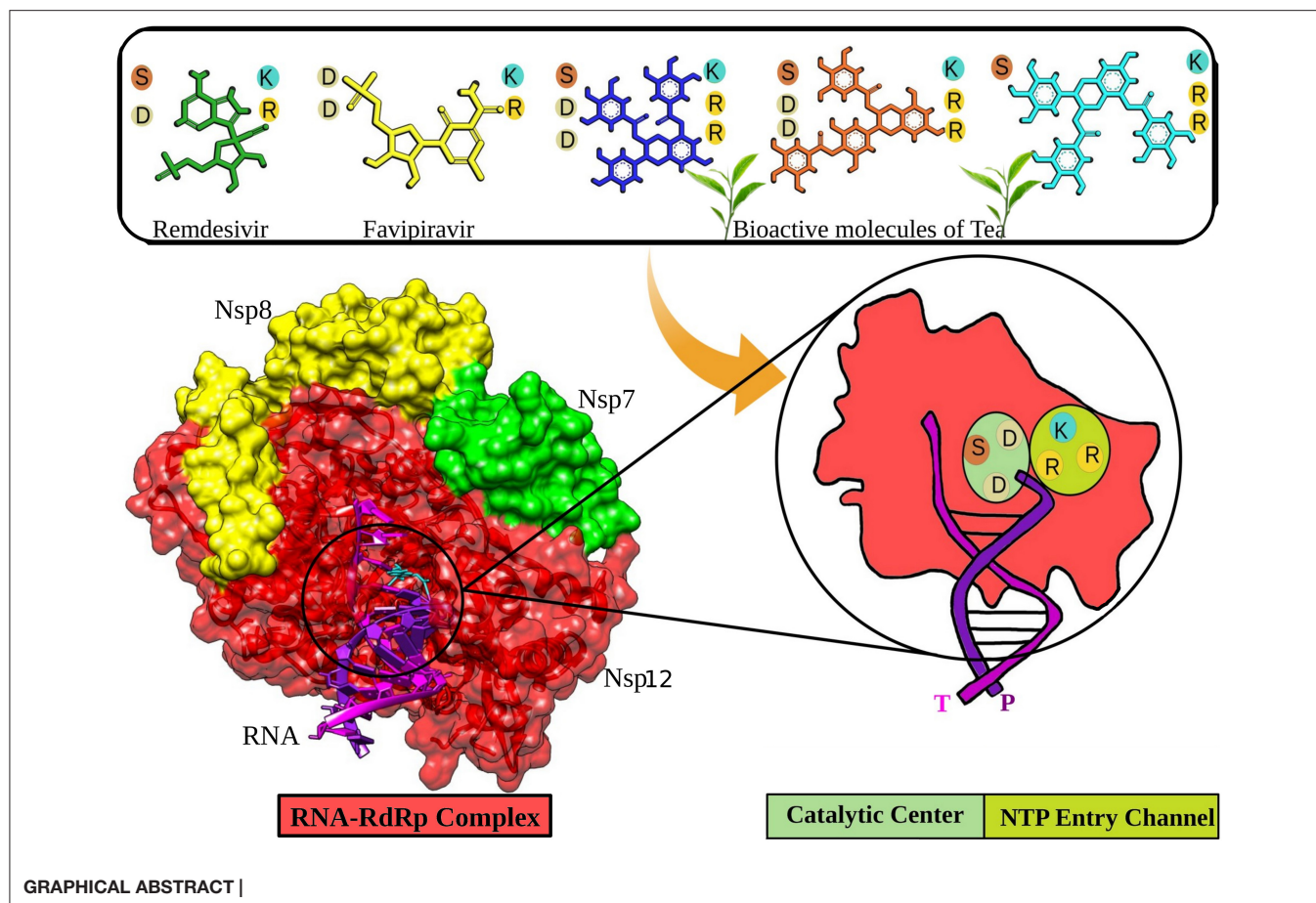
Bhardwaj VK, Singh R, Sharma J,
Rajendran V, Purohit R and Kumar S
(2021) Bioactive Molecules of Tea as
Potential Inhibitors for
RNA-Dependent RNA Polymerase of
SARS-CoV-2. *Front. Med.* 8:684020.
doi: 10.3389/fmed.2021.684020

The coronavirus disease (COVID-19), a worldwide pandemic, is caused by the severe acute respiratory syndrome-corona virus-2 (SARS-CoV-2). At this moment in time, there are no specific therapeutics available to combat COVID-19. Drug repurposing and identification of naturally available bioactive molecules to target SARS-CoV-2 are among the key strategies to tackle the notorious virus. The enzyme RNA-dependent RNA polymerase (RdRp) performs a pivotal role in replicating the virus. RdRp is a prime target for Remdesivir and other nucleotides analog-based antiviral drugs. In this study, we showed three bioactive molecules from tea (epicatechin-3,5-di-O-gallate, epigallocatechin-3,5-di-O-gallate, and epigallocatechin-3,4-di-O-gallate) that showed better interaction with critical residues present at the catalytic center and the NTP entry channel of RdRp than antiviral drugs Remdesivir and Favipiravir. Our computational approach to identify these molecules included molecular docking studies, followed by robust molecular dynamics simulations. All the three molecules are readily available in tea and could be made accessible along with other medications to treat COVID-19 patients. However, these results require validation by further *in vitro* and *in vivo* studies.

Keywords: RNA-RdRp, bioactive molecules, SARS-CoV-2, tea, COVID-19

INTRODUCTION

Recently, a major threat to humanity has emerged in the form of a novel coronavirus (CoV), causing a disease that is regarded as coronavirus disease 2019 (COVID-19) (1, 2). Taxonomically, this virus hails to the Coronaviridae family, which contains the enveloped positive-sense RNA virus of four major groups, alpha, beta, gamma, and delta (3, 4). Among these, Severe Acute Respiratory Syndrome (SARS) CoV and Middle East Respiratory Syndrome (MERS) CoV from the beta group are highly pathogenic to humans and develop symptoms like common cold, fever, and respiratory problems (5, 6). Previous outbreaks of the CoV in humans were reported in 2002 and 2012, which involved the SARS and MERS CoV, respectively (7, 8). Due to the absence of a specific treatment protocol, they had to be controlled via several public health measures (4, 9). The COVID-19 disease is provoked by a new CoV, named SARS-CoV-2. As compared to the other CoVs, SARS-CoV-2 has an uplifted human-to-human transmission rate, which gives a rationale for its extensive spread (10, 11).



The non-structural protein 12 (nsp12), also called the RNA-dependent RNA polymerase (RdRp), performs a significant function in the replication and transcription cycles of SARS-CoV-2 by catalyzing the synthesis of the viral RNA, making it one of the most critical targets for viral inhibition (1, 12). The nsp12 is likely to be assisted by cofactors like the nsp7 and nsp8 (13). The RdRp structure comprises a polymerase domain ranging from residue S367 to F920 that resembles a cupped “right hand” and a nidovirus-unique N-terminal extension domain ranging from residues D60 to R249, also called the NiRAN domain (14, 15). The two domains interact via the interface domain ranging from residues A250 to R365 (12). It also comprises the fingers subdomain (residues L366–A581 and K621–G679), the palm domain (residues T582–P620 and residues T680–Q815), and the thumb domain (residues H816–E920) (12, 14). There is an N-terminal β -hairpin (residues D29 to K50) between the palm subdomain and the NiRAN domain, which assists in the overall stabilization of the structure (12). The region from residue A4 to R118 comprises two helices and five antiparallel β -strands (9, 12). Another β -strand is present in the region between residues N215 and D218. It interacts with a strand in the region V96 to A100, thereby providing stability to the conformation by forming a compact and firm β -barrel architecture (12). The RdRp mediates a template-directed RNA synthesis part of the viral life cycle

where the template entry, the nucleoside triphosphate (NTP) entrance, and the nascent strand exit pathway converge into a central cavity, which is all positively charged (12, 16). The NTP entry channel is demarcated via hydrophilic motif F having K545, R553, and R555 residues (12). The RNA template enters the active site from a channel between motifs F and G, where motif E and the thumb subdomain hold the template strand. The active site is mainly constituted of motifs A and C, held up by motifs B and D (9).

The worldwide spread of SARS-CoV-2 and the rising statistics emphasize the importance of identifying drug candidates, which can act as potent antivirals to control the growing pandemic. RdRp is a promising target for inhibition, firstly, due to its critical involvement in the viral life cycle; secondly, it conserved the nature of its structure and sequence across several RNA viruses; and lastly, due to the missing homologs in the host (4, 12). Nucleotide analogs (NAs) include Remdesivir (adenosine analog), which has already proven to be effective against several viral diseases and has also been reported to inhibit SARS-CoV-2 by controlling its proliferation (17, 18). NAs upon entry tend to acquire an active 5'-triphosphate, which challenges the endogenous nucleotides to get incorporated in the viral RNA by acting as an alternate substrate for the RdRp (1). Remdesivir also works similarly by benefiting from the low fidelity of the RdRp,

thereby preventing viral proliferation by chain termination (19, 20). Analysis of the Remdesivir-mediated inhibition state of the virus illustrates conformational changes in the residues D760, D761, and D618. This allows the phosphate group of the inhibitor to interact with allosteric residue R555 (12).

In this study, a dataset of bioactive molecules of tea were screened and compared to Remdesivir and Favipiravir for their inhibitory potential against the RdRp of SARS-CoV-2. Tea is consumed by more than half of the world's population. The proof underpinning the health benefits of tea is rapidly growing with each new research published in the scientific literature. A plethora of studies have reported advantageous effects of habitual tea consumption against several types of cancers, cardiovascular diseases, diabetes, and arthritis (21). Apart from it, bioactive tea molecules are promising compounds in manifesting antiviral activities. They exhibit antiviral activity against a broad spectrum of human viruses, including HIV, herpes simplex virus, influenza, hepatitis B, and hepatitis C (22, 23). These compounds are even effective against Zika, Chikungunya, and Dengue viruses (23). Recent computational and experimental investigations documented the potent antiviral activities of bioactive tea molecules against multiple vital proteins, including the main protease (Mpro), non-structural protein 15 (Nsp15), spike, and RdRp of SARS-CoV-2 (24–28). The main objectives of the study were to analyze the interaction pattern and binding affinity of selected bioactive molecules and FDA-approved antiviral drugs with the active pocket of SARS-CoV-2 RdRp and, furthermore, to rank and suggest topmost molecules on the basis of their potential to hinder the replication process of SARS-CoV-2.

MATERIALS AND METHODS

Datasets

The crystallographic RdRp-RNA structure complex (PDB Id: 7BV2) was obtained from the protein data bank with a resolution of 2.50 Å (29). The chain A (nsp12) contains 951 amino acids. The protein complex also constitutes of primer and template RNA strands of length 20 and 30 nucleotides, respectively. A set of bioactive molecules of Tea (24) was prepared for molecular docking and MD simulation studies. The structures of Remdesivir and Favipiravir in their active forms were obtained from PubChem (30).

Molecular Docking

A set of bioactive molecules from tea along with Remdesivir and Favipiravir was docked into the active site of RdRp of SARS-CoV-2. Discovery Studio's CDOCKER algorithm was utilized to carry out molecular docking. CDOCKER is CHARMM-based semiflexible docking engine (31). The resilient conformation section grabbed by ligand molecules searched employing high-temperature kinetics. The optimization at the binding site of each conformation is achieved by using the simulated annealing method to obtain reliable docking outcomes. The CDOCKER parameters were kept on default. The number of starting random conformations and the number of rotated ligand orientations to refine for each of the conformations for 1,000 dynamics steps were set to 10. Moreover, for annealing refinement, the number

of heating steps was 2000, while the number of cooling sets was set to 5,000. The distance to consider Pi-cation, Pi-Pi, and Pi-alkyl interactions was set to 5, 6, and 5.5 Å, respectively. A radius of 8.0 Å was assigned centering the ligand in the active site that contains all the active residues participating in the binding of the ligand to the RdRp–RNA SARS-CoV-2 protein. The 3D structures of all the bioactive molecules were prepared using the Discovery Studio package (32). Furthermore, for energy minimization, we have used CHARMM force field and DFT protocols (33). The built molecules were then read in Discovery Studio.

Molecular Dynamics Simulations

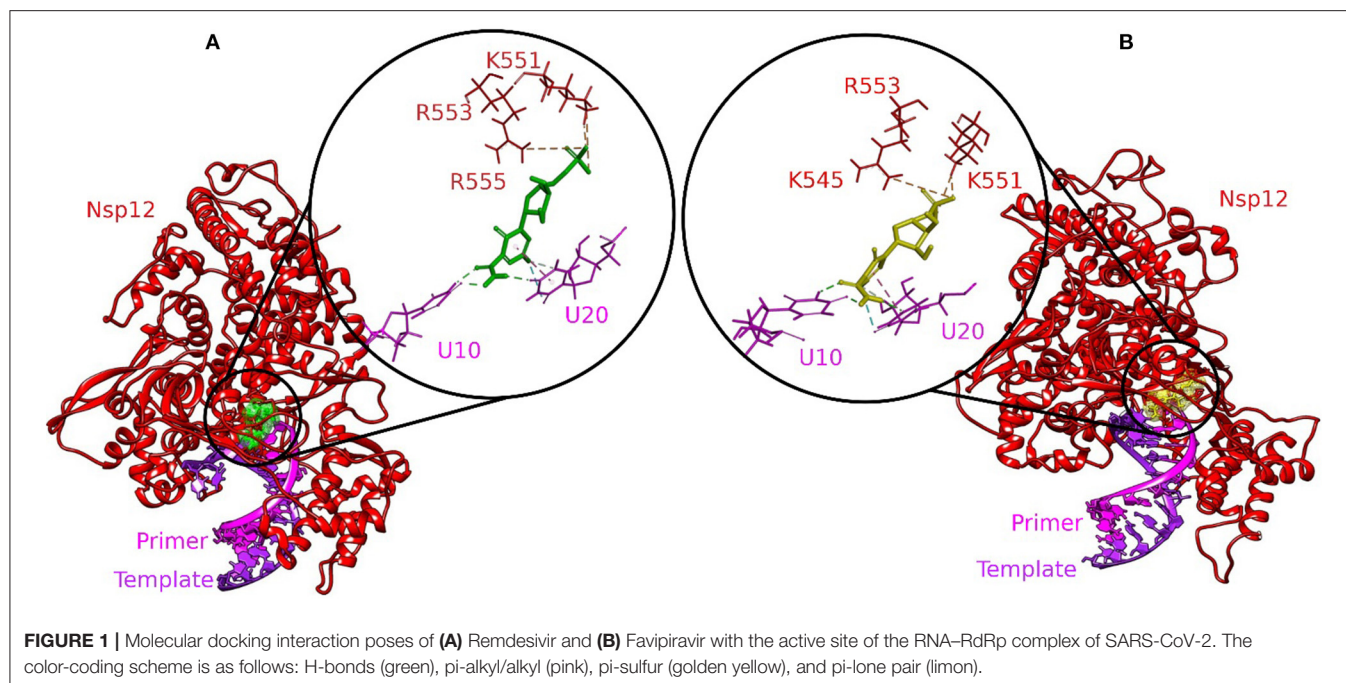
A 100-ns MD simulation was performed for all the selected complexes using the GROMACS 4.6.7 suite (34). A large protein size of RdRp (951 amino acids) along with RNA increases the complexity of all atomic simulation setups by several folds. The protein topology for the protein complexes has been derived from the CHARMM27 force field, while ligand topologies were prepared by employing the PRODRG server. Every protein complex system was solvated with a simple point charge (SPC) water model. Each system was neutralized by attaching chloride ions, accompanied by energy minimization with the steepest descent method of integration. After minimization, the protein was equilibrated for 10 ns at 300 K in NVT as well as NPT ensemble (35). Finally, MD simulation was performed at a temperature of 300 K for 100 ns under periodic boundary state and the time constant of 1.0 ps for coupling. The constant pressure and temperature (1/atm/300 K) were managed through Berendsen Coupling Algorithm17 with a time constant of 0.2 ps for heat-bath coupling (36). The SHAKE algorithm was used during simulation to maintain the length of the bond involving the hydrogen bond. The free binding energies of the selected complexes were calculated with g_mmpbsa software in GROMACS 4.6.7. The MM-PBSA method was applied to the calculation of the binding free energies (37). It can be calculated via the following equation:

$$\Delta G_{\text{binding}} = G_{\text{Complex}} - [G_{\text{Protein}} + G_{\text{Ligand}}] \quad (1)$$

Here, $\Delta G_{\text{binding}}$ delineates the binding free energy of the protein–ligand complexes; G_{Protein} and G_{Ligand} delineate the overall free energies of the protein and ligand molecule. The generated trajectories of MD simulations were then practiced to construct the graph for root mean square deviation (RMSD), RMSD conformational clustering, and hydrogen bond; “gmh rms,” “gmh cluster,” and “gmh hbond” scripts of GROMACS were employed to interpret the yield trajectory data.

RESULTS AND DISCUSSION

The availability of a high-resolution crystallographic structure of the RdRp–RNA complex has unlocked the pathway for the development of potential antivirals targeting the particular protein of SARS-CoV-2. Many studies around the world have suggested that high intake of foods rich in bioactive molecules has beneficial impact on human health and may mitigate the possibility of various human ailments, such as diabetes,



cancer, Alzheimer's disease, cataracts, stroke, and age-related functional disorders (38). Bioactive molecules are rich in structural diversity and provide a large area of chemical space for the exploration of possible target sites. In our previous study, the bioactive molecules Oolonghomobisflavan-A, Theasinensin-D, and Theaflavin-3-O-gallate of tea showed better binding than the FDA-approved drugs to the active site of the main protease of SARS-CoV-2 (24). Herein, we screened a dataset of bioactive molecules from tea to check and compare their affinity toward the active site of the RdRp-RNA complex of SARS-CoV-2. The top three tea bioactive molecules screened in this study were epicatechin-3,5-di-O-gallate, epigallocatechin-3,4-di-O-gallate, and epigallocatechin-3,5-di-O-gallate.

Molecular Docking

Molecular docking is an exemplary tool to identify the intermolecular framework of ligand-protein, protein-nucleic acid, and protein-protein complexes. The RdRp-RNA complex of SARS-CoV-2 was docked with a set of bioactive molecules from tea and FDA-approved repurposed drugs Remdesivir and Favipiravir. All the molecules were ranked according to their binding scores generated by the CDOCKER protocol of Discovery Studio (**Supplementary Table 1**). The binding poses of Remdesivir and Favipiravir within the active site of the RdRp-RNA complex were shown in **Figure 1**. Remdesivir binds to the active site by interacting with the residues of RdRp and RNA nucleotides of both the primer and template strand. The Uracil at position 20 of the primer strand is bound to Remdesivir by two hydrogen bonds and Pi-alkyl interactions. Remdesivir also formed hydrogen bonds with Uracil at position 10 of the template strand. Additionally, Remdesivir was stabilized in the binding site by a Pi-Sulfur interaction with residues Arg555, Lys551, and Arg553 of

RdRp. Furthermore, residues Val557, Lys545, Asp760, Cys622, Asp623, Thr680, Ser759, Asn691, Thr687, and Ala688 of RdRp were involved in van der Waals interactions with Remdesivir. Favipiravir occupied the central pocket of the RdRp-RNA complex and formed a hydrogen bond with Uracil at position 20 of the primer RNA strand. Two hydrogen bonds were formed between Favipiravir and Uracil at position 10 of the template RNA strand. Residues Lys545 and Lys551 were involved in Pi-Sulfur interaction with Favipiravir. Residues Arg555, Val557, Ser6682, Asp761, Asp760, and Adenine at position 11 of the template strand were involved in van der Waals interactions.

Three molecules from tea displayed stronger binding with the active site of the RdRp-RNA complex in terms of CDOCKER interaction energy (**Table 1**). The docking poses with the most favorable interaction patterns are shown in **Figure 2**. The molecule epicatechin-3,5-di-O-gallate formed three hydrogen bonds and a Pi-anion interaction with the Uracil at position 20 of the primer RNA strand. The molecule was further stabilized within the active pocket by eight hydrogen bonds with residues Lys545, Asp623, Asp425, Asn691, Ser759, Ser682, and Ser814. Residues Asp623 and Thr556 were involved in the formation of Pi-anion and Pi-Lone Pair interactions, respectively. Many other residues of protein RdRp along with the Adenine and Uracil of the template RNA strand at positions 11 and 10 showed van der Waals interactions. The second molecule from tea, epigallocatechin-3,5-di-O-gallate, formed two hydrogen bonds with Uracil at position 20 of primer RNA. It interacted with residues Ser682, Asp623, Arg553, Arg555, Lys545, Ile548, Asp760, and Ser759 via nine hydrogen bonds. Residues Arg555 and Ala547 were also involved in Pi-alkyl interactions. Furthermore, the binding of epigallocatechin-3,5-di-O-gallate to the active pocket of the RdRp-RNA complex

TABLE 1 | Selected tea bioactive molecules and FDA-approved drugs based on -CDOCKER interaction energy.

Molecules	-CDOCKER interaction energy
Epigallocatechin-3,5-di-O-gallate	79.4086
Epicatechin-3,5-di-O-gallate	78.0801
Epigallocatechin-3,4-di-O-gallate	74.4848
Favipiravir	74.1136
Remdesivir	33.9374

was enhanced by van der Waals interactions by residues Cys622, Thr680, Ser682, Arg624, Val557, Ser549, Ser814, Asp761, Ala688, Thr687, and Asn691. The molecule epigallocatechin-3,4-di-O-gallate also showed higher binding potential than Favipiravir and Remdesivir at the active site of the RdRp–RNA complex. The primer strand Uracil at position 10 interacted via two hydrogen bonds, while the Adenine at position 20 of the same RNA strand formed one hydrogen bond with epigallocatechin-3,4-di-O-gallate. The residues Arg836, Arg555, Ser759, Asn691, Asp623, Ser682, Asp452, Arg553, and Lys545 of the RdRp protein showed 10 hydrogen bonds with epigallocatechin-3,4-di-O-gallate. Moreover, three Pi-anion interactions (Arg836, Cys622, Asp623), two Pi-alkyl interactions (Arg555 and Val557), and a Pi Lone Pair interaction were also observed at the active site.

Remdesivir is a nucleotide analog that occupied the central position of the catalytic active site and formed a covalent bond with the primer RNA strand, and terminates replication by non-obligate RNA chain termination (12). However, studies contrary to these results showed the addition of more nucleotides to the RNA strand even after the incorporation of Remdesivir resulting in delayed chain termination (20, 39). The catalytic center of RdRp protein is composed of residues Ser759, Asp760, and Asp761. This site is conserved in most viral RdRps (12, 29). Residues Lys545, Arg553, and Arg555 contribute to the formation of the NTP entry channel (12). The FDA-approved drugs Remdesivir and Favipiravir formed weaker van der Waals and Pi-Sulfur interaction with the residues of the catalytic center and the NTP entry channel. However, all the three selected tea molecules formed stronger hydrogen bonds with most of these residues. Residues Asn691, Ser682, and Asp623 impart specificity to RNA replication over DNA strand by recognizing the RNA specific 2'-OH group (40). Our selected molecules from tea formed stronger hydrogen bonds with residues Asn691, Ser682, and Asp623 as compared to weaker van der Waals interactions formed by Remdesivir and Favipiravir. Stronger interactions with RNA recognition residues and other residues involved in the formation of the catalytic center and the NTP entry channel would ensure the destabilization of the incoming RNA in the active site and hence halt/meddle with the process of viral replication. Furthermore, MD simulations were conducted to substantiate the molecular docking results and explore the dynamics of ligand–protein interactions at the catalytic site of the RdRp–RNA complex.

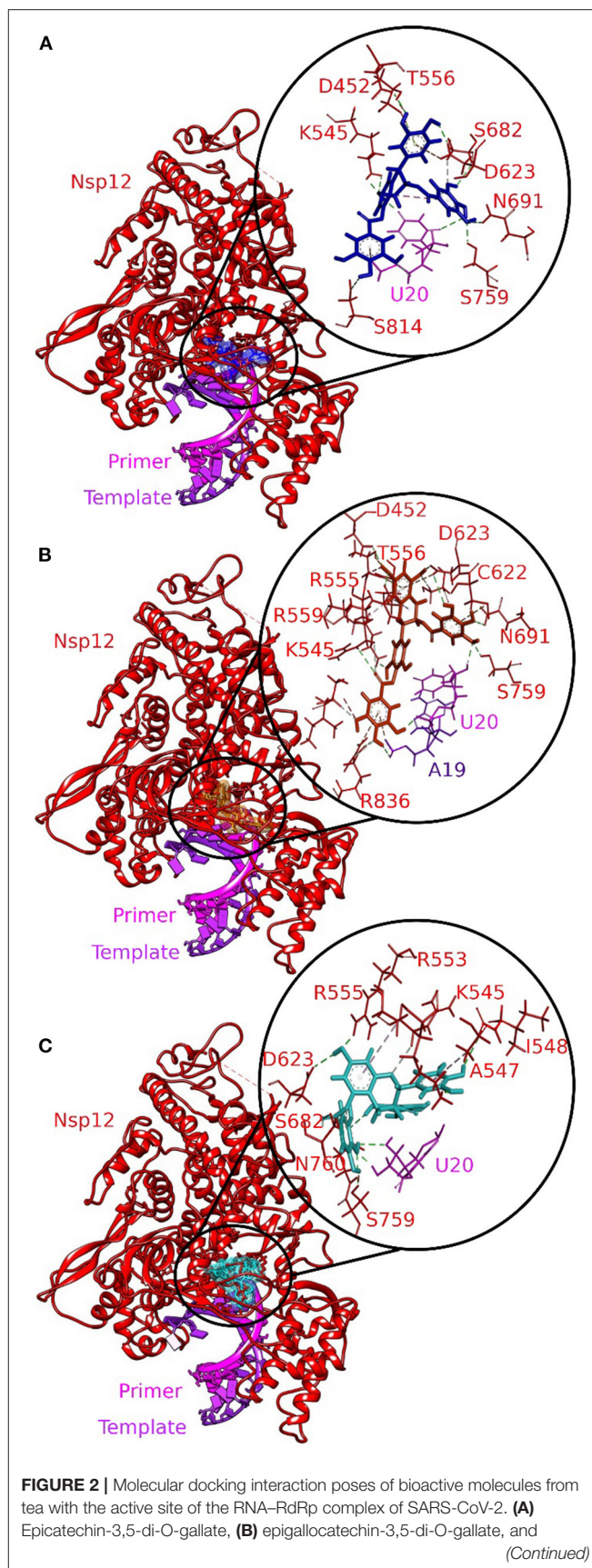


FIGURE 2 | (C) epigallocatechin-3,4-di-O-gallate. The color-coding scheme is as follows: H-bonds (green), pi-alkyl/alkyl (pink), pi-sulfur (golden yellow), and pi-lone pair (limon).

Molecular Dynamics Simulations

The basic understanding of how biological macromolecules function requires an awareness of molecular structure and dynamics (41). MD simulations establish fundamental links based on experimental and theoretical evidences between structure and dynamics, allowing the investigation of conformational energy landscape available to biological macromolecules (42, 43). In our previous studies, we showed the potential of bioactive tea molecules to inhibit the Mpro (24) and nsp15 (25) of SARS-CoV-2. Moreover, in a different study, we demonstrated the inhibitory potential of acridinedione analogs to inhibit the Mpro of SARS-CoV-2 (44). A recent study employing MD simulations suggested potential molecules to target the RdRp of SARS-CoV-2 (45). The RdRp–RNA complexes with Remdesivir, Favipiravir, and three selected bioactive molecules of tea were subjected to explicit MD simulations. The MD results were analyzed by using different multiscale computational methods.

Structural Stability of RdRp–Ligand–RNA Complexes

The RMSD is a classical technique for the analysis of MD results. It is a popular measure for analyzing the structural stability of protein structures. The RMSDs of all the C- α atoms of selected protein complexes were calculated, as depicted in **Figure 3**. All the protein structures deviated < 0.35 nm during the simulation. The RdRp–RNA complex with Remdesivir had the highest deviation with an average RMSD value of \sim 0.32 nm. The complex with Favipiravir deviated at a lower trajectory than Remdesivir after 25 ns until the end of the simulation. All the three RdRp–RNA complexes having selected tea molecules (epicatechin-3,5-di-O-gallate, epigallocatechin-3,4-di-O-gallate, and epigallocatechin-3,5-di-O-gallate) showed relatively lower deviations than Remdesivir. The average RMSD values of epicatechin-3,5-di-O-gallate and epigallocatechin-3,5-di-O-gallate were 0.27 and 0.29 nm, respectively. After initial deviations from the starting structure for the first 25 ns, the RMSD trajectories of the simulated complexes stabilized and showed convergence in the second half of the simulation. These results indicated that the structural stability of RdRp–RNA complexes with bioactive molecules of tea was comparable to that of Favipiravir and was more stable than Remdesivir. Additionally, lower RMSD values also showed that all the simulated structures were able to reproduce correct binding poses as generated by molecular docking studies.

Dynamics of Simulated Complexes

The ensemble clustering of simulation data is a conclusive and effectual method of analyzing the structural flexibility of concerned protein systems (46). The MD trajectories of all the selected complexes for 5 ns (45–50 ns) of the simulation

period were extracted and subjected to clustering analysis. The clustering was done on three different combinations of the receptor and the ligand to explore the dynamics of ligand interactions with individual receptors and the whole protein (RdRp)–nucleic acid (active site RNA) complex. The RMSD clustering results are shown in **Figure 4**. In RdRp–ligand complexes, Favipiravir formed the least number of clusters with an average RMSD of 0.130 nm, while Remdesivir and epigallocatechin-3,4-di-O-gallate formed five clusters each with an average RMSD of 0.131 and 0.130 nm, respectively. In RNA–ligand complexes, Favipiravir showed only two clusters with an average RMSD of 0.109 nm. Among the bioactive molecules of tea, epigallocatechin-3,4-di-O-gallate showed only three clusters, while the rest of the two molecules formed six clusters each while interacting with RNA. Remdesivir formed 12 clusters with an average RMSD of 0.160 nm. Similarly, in protein–RNA–ligand complexes, Remdesivir formed the most number of clusters (nine clusters) followed by epicatechin-3,5-di-O-gallate (eight clusters), epigallocatechin-3,5-di-O-gallate (eight clusters), epigallocatechin-3,4-di-O-gallate (six clusters), and Favipiravir (four clusters). The average RMSD of all the clusters was below 0.140 nm. These results showed that the selected bioactive tea molecules were more stable than Remdesivir and almost comparable to Favipiravir in clustering analysis. Furthermore, to visualize the effect of structural fluctuations on intermolecular interactions, we analyzed the hydrogen bond formations between RdRp–ligand and RNA–ligand complexes.

Analysis of Intermolecular Hydrogen Bonds

Hydrogen bonds are commonly considered as mediators of protein–ligand binding and also promote the binding affinity of a ligand by displacing the water molecules bound to protein into the bulk solvent. We calculated the number of hydrogen bonds formed by the selected bioactive molecules of tea and standard drugs with both the RdRp and RNA of SARS-CoV-2. Remdesivir and Favipiravir formed an average of four and five hydrogen bonds, respectively, with the RdRp of SARS-CoV-2. Epicatechin-3,5-di-O-gallate formed the most number of hydrogen bonds during the simulation with the residues of the RdRp of SARS-CoV-2. The average number of hydrogen bonds formed during the simulation by epicatechin-3,5-di-O-gallate was 7, with few conformations formed up to 10 hydrogen bonds (**Figure 5A**). Epigallocatechin-3,4-di-O-gallate stabilized in the binding pocket by forming an average of six hydrogen bonds with the RdRp of SARS-CoV-2. The third selected molecule epigallocatechin-3,5-di-O-gallate for the first 25 ns showed an average of five hydrogen bonds, while for the next 25 ns, the average number of hydrogen bonds was five with RdRp. Similarly, epicatechin-3,5-di-O-gallate, epigallocatechin-3,4-di-O-gallate, and epigallocatechin-3,5-di-O-gallate formed more hydrogen bonds with RNA at the active site of RdRp than Remdesivir and Favipiravir. The average number of hydrogen bonds formed between epicatechin-3,5-di-O-gallate and RNA was two, while the highest number of hydrogen bonds was four. Both epigallocatechin-3,4-di-O-gallate and epigallocatechin-3,5-di-O-gallate showed up to five hydrogen bonds with the active

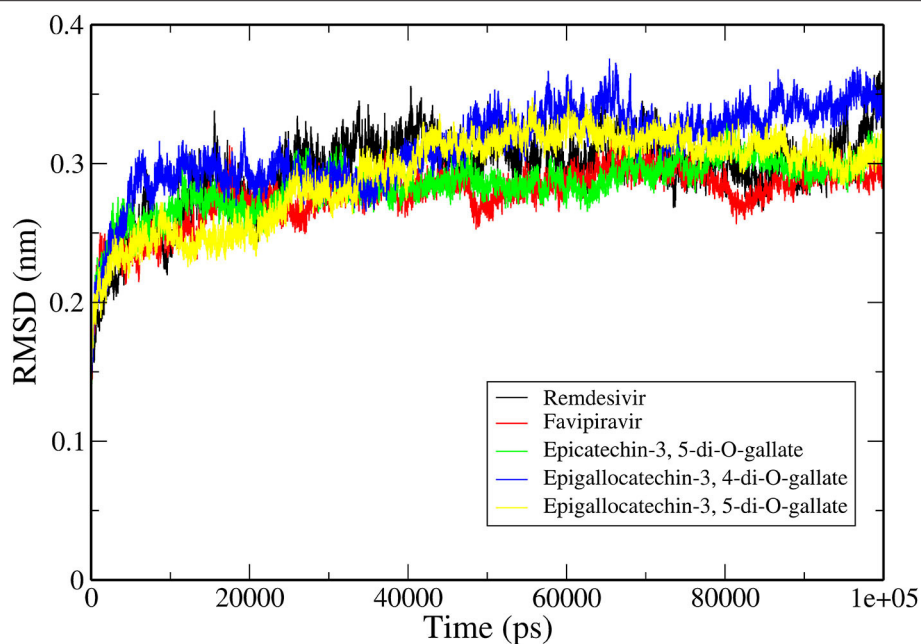


FIGURE 3 | Backbone RMSDs are shown as a function of time for the RNA-RdRp complex of SARS-CoV-2 and ligands.

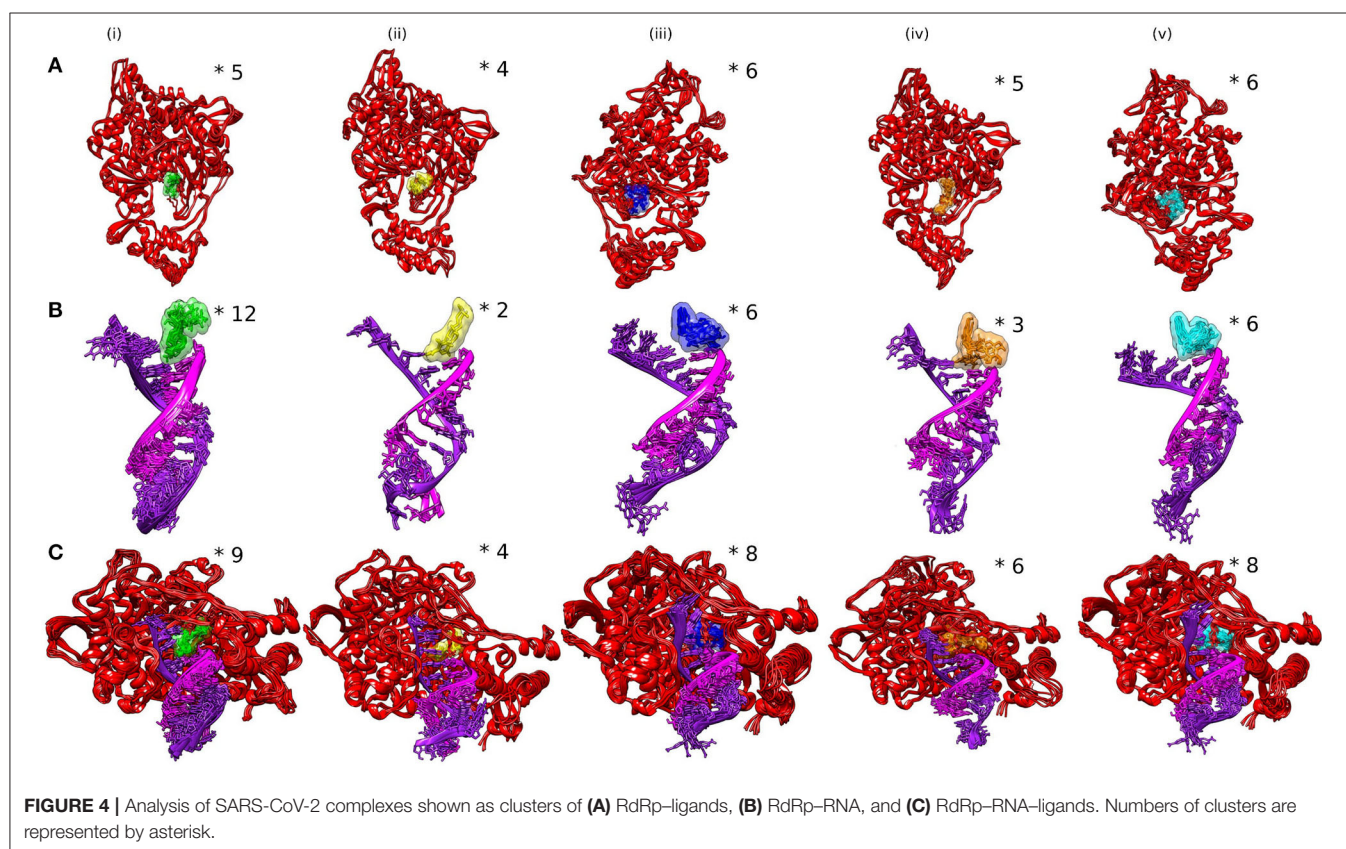


FIGURE 4 | Analysis of SARS-CoV-2 complexes shown as clusters of (A) RdRp-ligands, (B) RdRp-RNA, and (C) RdRp-RNA-ligands. Numbers of clusters are represented by asterisk.

site RNA (**Figure 5B**). The standard drugs Remdesivir and Favipiravir were able to form an average of one and two hydrogen bonds, respectively, with the active site RNA. All the three

selected bioactive molecules of tea formed a greater number of hydrogen bonds within the active site of the RdRp of SARS-CoV-2 than Remdesivir and Favipiravir throughout the simulation

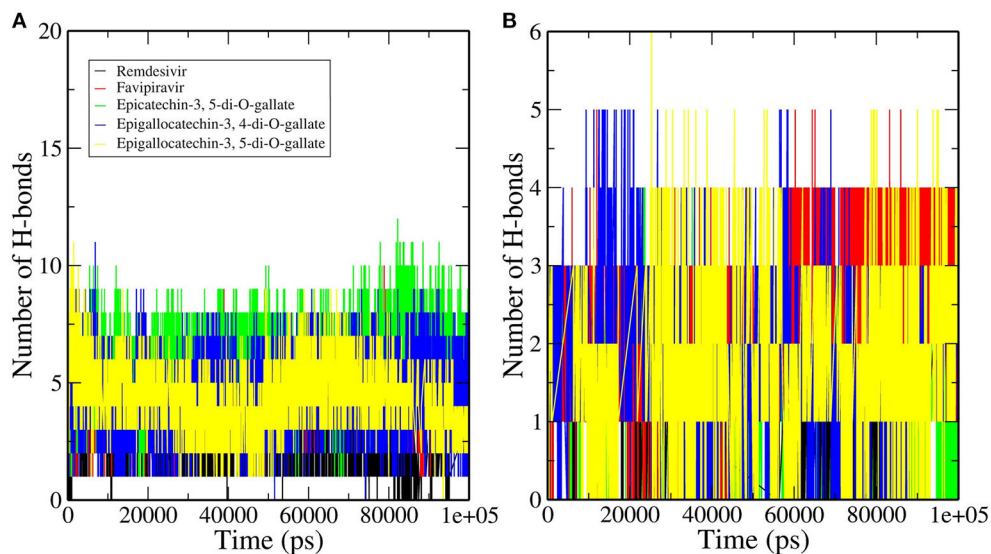


FIGURE 5 | Hydrogen bond profiles of the RNA-RdRp complex of SARS-CoV-2 having (A) RdRp and (B) RNA.

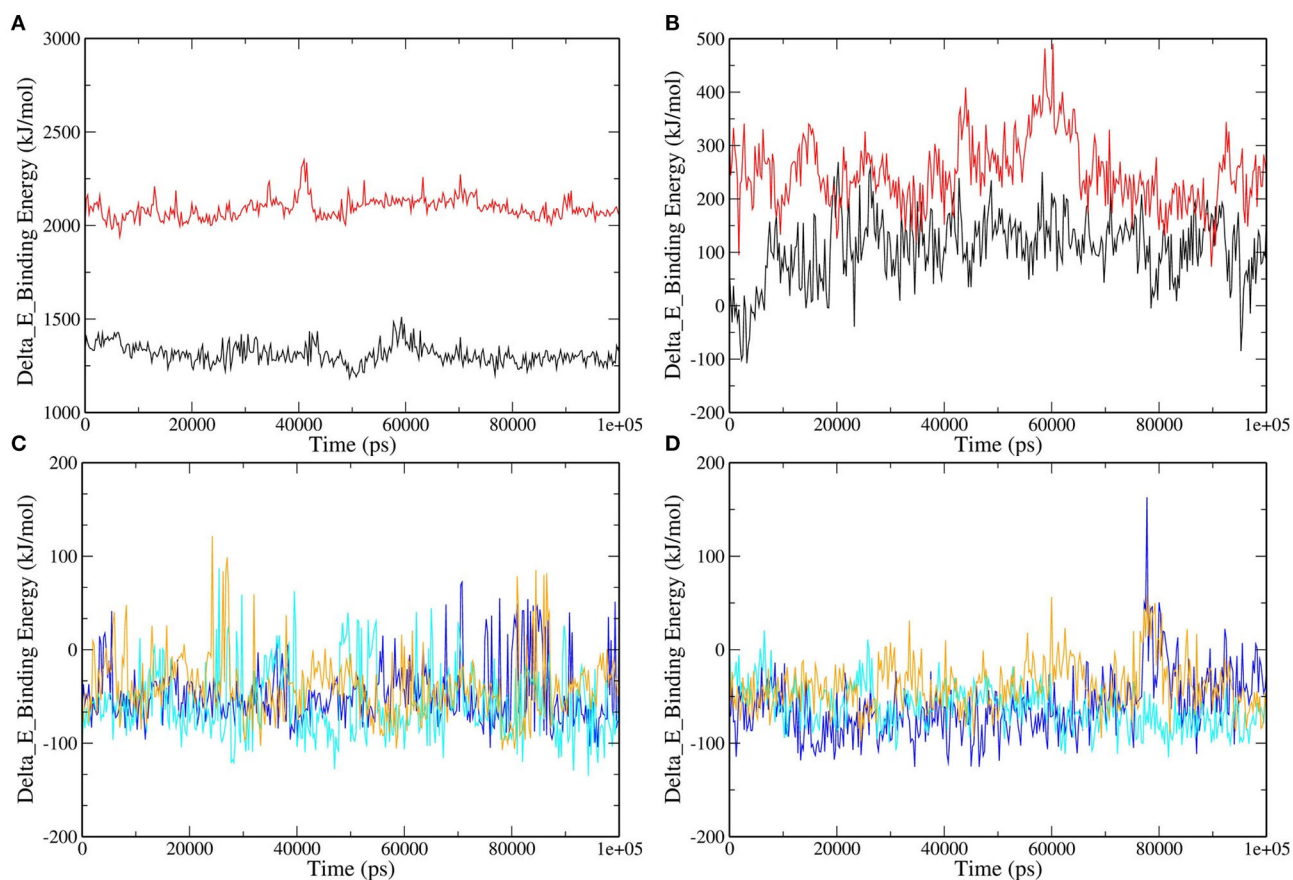


FIGURE 6 | Binding free energy represented graphically for RdRp protein and color scheme as follows: (A) RNA-Remdesivir (black), RNA-Favipiravir (red). (B) RdRp-Remdesivir (black), RdRp-Favipiravir (red). (C) RNA-epicatechin-3,5-di-O-gallate (blue), RNA-epigallocatechin-3,5-di-O-gallate (cyan), RNA-epigallocatechin-3,4-di-O-gallate (orange). (D) RdRp-epicatechin-3,5-di-O-gallate (blue), RdRp-epigallocatechin-3,5-di-O-gallate (cyan), RNA-epigallocatechin-3,4-di-O-gallate (orange).

TABLE 2 | Binding free energy (MM-PBSA) calculations for selected complexes.

RNA-RdRp-Ligand Complexes	ΔE binding (kJ/mol)	ΔE polar solvation (kJ/mol)	SASA (kJ/mol)	ΔE Electrostatic (kJ/mol)	ΔE Van der Waal (kJ/mol)
RdRp-Epicatechin-3,5-di-O-gallate	−61.461	209.138	−19.313	−87.878	−163.408
RdRp-Epigallocatechin-3,5-di-O-gallate	−38.515	280.167	−18.915	−139.544	−160.222
RdRp-Epigallocatechin-3,4-di-O-gallate	−62.278	249.433	−22.516	−110.399	−178.796
RdRp-Remdesivir	107.346	437.201	−12.356	−248.406	−69.093
RdRp-Favipiravir	248.378	977.029	−12.653	−657.208	−58.790
RNA-Epicatechin-3,5-di-O-gallate	−47.148	66.063	−7.478	−41.142	−64.592
RNA-Epigallocatechin-3,5-di-O-gallate	−39.857	58.538	−7.632	−27.821	−62.942
RNA-Epigallocatechin-3,4-di-O-gallate	−56.695	99.860	−10.168	−55.757	−90.630
RNA-Favipiravir	2091.712	−18.286	−6.780	2168.615	−51.572
RNA-Remdesivir	1311.862	−30.025	−7.045	1387.624	−40.383

period. These results were further confirmed by extracting RdRp–ligand–RNA complex trajectories at different time intervals and visualizing the stability of selected molecules in the active site of RdRp (**Supplementary Figure 1**). The bioactive molecules of tea were tightly bound to the active site of RdRp by forming a greater number of hydrogen bonds and other non-covalent interactions than Remdesivir and Favipiravir. Furthermore, an efficient and reliable method of calculating the binding free energy and its contributors was employed to compare bioactive tea molecules with Remdesivir and Favipiravir.

The Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) Analysis

MD simulations present a glimpse of a ligand's stability in the binding region of a concerned protein. By implementing MM-PBSA calculations, we assessed the binding free energy of selected RdRp–ligand complexes. The observation was made with the extraction of the RdRp–ligand complex scripts from MD simulations. The binding free energy can acceptably illustrate the durability of the linking ligand receptor, which is an integral aspect of drug development. The binding energies of the ligands with RdRp and RNA were compared. Remdesivir and Favipiravir showed positive binding energy with both RdRp and RNA. Moreover, the binding free energy was decomposed into electrostatic, SASA, van der Waals, and polar solvation energies. The lesser the binding energy, the more reliable the ligand–protein binding. The favorable contribution of SASA, electrostatic, and van der Waals energies was devoted to the binding of RdRp and RNA with our selected molecules from tea. In contrast, electrostatic energy was positive for standard molecules in RNA but favorable in RdRp, as shown in **Figure 6** and **Table 2**. Approximately all atoms inside a macromolecule convey a partial charge, and thus, molecules striving for molecular classification interact via electrostatic interactions. It is believed that these interactions assist two leading roles: to control the molecules toward their binding style and to generate unique interactions within the active site (47). It is assumed that the positive values of the electrostatic associations destabilize the interaction and thus diminish the affinity. By contrast, in

binding free energy, the polar solvation energy participated generously to increase the total energy. Van der Waals energy's augmentation to the overall binding free energy was higher upon the electrostatic contribution energy. The higher (−ve) binding energy is responsible for potential binding. These results bestow higher binding energy for all the selected tea molecules as compared to Remdesivir and Favipiravir.

CONCLUSION

The RdRp is an attractive target for the development of specific inhibitors for SARS-CoV-2. The FDA-approved drugs Remdesivir and Favipiravir showed promising results in curing COVID-19 patients. In this study, a dataset of bioactive molecules from tea was screened to analyze its interaction profiles within the active site of the RdRp–RNA complex. The molecules epicatechin-3,5-di-O-gallate, epigallocatechin-3,4-di-O-gallate, and epigallocatechin-3,5-di-O-gallate formed stronger hydrogen bonds with the key residues involved in the recognition of RNA for replication, the catalytic center, and the NTP entry channel. These residues showed weaker van der Waals interactions with Remdesivir and Favipiravir. Both the selected molecules also showed the most favorable binding energies during robust MD simulations than the standard drug molecules. The bioactive molecules of tea also target the Mpro and nsp15 of SARS-CoV-2 as shown by our previous reports. The results our previous study along with the present findings disclose the ability of bioactive molecules of tea to target multiple proteins of SARS-CoV-2 (Mpro, nsp15, and RdRp). These bioactive molecules could be quickly made available in formulations along with other antiviral therapies to rapidly cure COVID-19 patients. These *in silico* results, however, require validation by experimental studies.

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

RP conceived of and designed the study. VKB, RS, JS, and VR analyzed and interpreted the data. RS, VKB, VR, RP, and SK critically revised it for important intellectual content. All authors gave final approval of the version to be published.

ACKNOWLEDGMENTS

We gratefully acknowledged the Director, CSIR-Institute of Himalayan Bioresource Technology, Palampur for providing the facilities to carry out this work. RP gratefully acknowledged

the Board of Research in Nuclear Sciences, Department of Atomic Energy, Mumbai, India for financial support vide letter No: 37(1)/14/26/2015/BRNS. The CSIR support in the form of projects MLP0201 for bioinformatics studies and HCP0007 for providing project fellowship is highly acknowledged. Also update IHBT communication number to 4844.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.684020/full#supplementary-material>

REFERENCES

- Gordon CJ, Tchesnokov EP, Woolner E, Perry JK, Feng JY, Porter DP, et al. Remdesivir is a direct-acting antiviral that inhibits RNA-dependent RNA polymerase from severe acute respiratory syndrome coronavirus 2 with high potency. *J Biol Chem.* (2020) 295:6785–97. doi: 10.1074/jbc.RA120.013679
- WHO. *COVID-19 Situation Report 29*. Geneva: World Health Organization (2020).
- Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet.* (2020) 395:565–74. doi: 10.1016/S0140-6736(20)30251-8
- Shannon A, Le NTT, Selisko B, Eydoux C, Alvarez K, Guillemot JC, et al. Remdesivir and SARS-CoV-2: structural requirements at both nsp12 RdRp and nsp14 exonuclease active-sites. *Antiviral Res.* (2020) 178:104793. doi: 10.1016/j.antiviral.2020.104793
- Phadke M, Saunik S. COVID-19 treatment by repurposing drugs until the vaccine is in sight. *Drug Dev Res.* (2020) 81:541–3. doi: 10.1002/ddr.21666
- Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med.* (2020) 382:727–33. doi: 10.1056/NEJMoa2001017
- Peiris JSM, Lai ST, Poon LLM, Guan Y, Yam LYC, Lim W, et al. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet.* (2003) 361:1319–25. doi: 10.1016/S0140-6736(03)13077-2
- Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N Engl J Med.* (2012) 367:1814–20. doi: 10.1056/NEJMoa1211721
- Kirchdoerfer RN, Ward AB. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nat Commun.* (2019) 10:2342. doi: 10.1038/s41467-019-10280-3
- Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet.* (2020) 395:514–23. doi: 10.1016/S0140-6736(20)30154-9
- Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet.* (2020) 395:507–13. doi: 10.1016/S0140-6736(20)30211-7
- Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, et al. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science.* (2020) 368:779–82. doi: 10.1126/science.abb7498
- Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, et al. One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proc Natl Acad Sci USA.* (2014) 111:E3900–9. doi: 10.1073/pnas.1323705111
- McDonald SM. RNA synthetic mechanisms employed by diverse families of RNA viruses. *Wiley Interdiscip Rev RNA.* (2013) 4:351–67. doi: 10.1002/wrna.1164
- Lehmann KC, Gulyaeva A, Zevenhoven-Dobbe JC, Janssen GMC, Ruben M, Overkleeft HS, et al., Kravchenko AA, Leontovich AM, et al. Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic Acids Res.* (2015) 43:8416–34. doi: 10.1093/nar/gkv838
- Gong P, Peersen OB. Structural basis for active site closure by the poliovirus RNA-dependent RNA polymerase. *Proc Natl Acad Sci USA.* (2010) 107:22505–10. doi: 10.1073/pnas.1007626107
- Wang M, Cao R, Zhang L, Yang X, Liu J, Xu M, et al. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) *in vitro*. *Cell Res.* (2020) 30:269–71. doi: 10.1038/s41422-020-0282-0
- Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First case of 2019 novel coronavirus in the United States. *N Engl J Med.* (2020) 382:929–36. doi: 10.1056/NEJMoa2001191
- Jordan PC, Liu C, Raynaud P, Lo MK, Spiropoulou CF, Symons JA, et al. Initiation, extension, and termination of RNA synthesis by a paramyxovirus polymerase. *PLoS Pathog.* (2018) 14:e1006889. doi: 10.1371/journal.ppat.1006889
- Tchesnokov EP, Feng JY, Porter DP, Götte M. Mechanism of inhibition of ebola virus RNA-dependent RNA polymerase by remdesivir. *Viruses.* (2019) 11:326. doi: 10.3390/v11040326
- Khan N, Mukhtar H. Tea and health: studies in humans. *Curr Pharm Des.* (2013) 19:6141–7. doi: 10.2174/1381612811319340008
- Kumar Bhardwaj V, Purohit R, Kumar S. Himalayan bioactive molecules as potential entry inhibitors for the human immunodeficiency virus. *Food Chem.* (2021) 347:128932. doi: 10.1016/j.foodchem.2020.128932
- Xu J, Xu Z, Zheng W. A review of the antiviral role of green tea catechins. *Molecules.* (2017) 22:1337. doi: 10.3390/molecules22081337
- Bhardwaj VK, Singh R, Sharma J, Rajendran V, Purohit R, Kumar S. Identification of bioactive molecules from tea plant as SARS-CoV-2 main protease inhibitors. *J Biomol Struct Dyn.* (2020) 1–10. doi: 10.1080/07391102.2020.1766572
- Sharma J, Kumar Bhardwaj V, Singh R, Rajendran V, Purohit R, Kumar S. An *in-silico* evaluation of different bioactive molecules of tea for their inhibition potency against non structural protein-15 of SARS-CoV-2. *Food Chem.* (2020) 346:128933. doi: 10.1016/j.foodchem.2020.128933
- Du A, Zheng R, Disoma C, Li S, Chen Z, Li S, et al. Epigallocatechin-3-gallate, an active ingredient of traditional Chinese medicines, inhibits the 3CLpro activity of SARS-CoV-2. *Int J Biol Macromol.* (2021) 176:1–12. doi: 10.1016/j.ijbiomac.2021.02.012
- Ohgitani E, Shin-Ya M, Ichitani M, Kobayashi M, Takihara T, Kawamoto M, et al. Significant inactivation of SARS-CoV-2 by a green tea catechin, a catechin-derivative and galloylated theaflavins *in vitro*. *bioRxiv [Preprint]*. (2020). doi: 10.1101/2020.12.04.412098
- Liu J, Bodnar BH, Meng F, Khan A, Wang X, Luo G, et al. Epigallocatechin gallate from green tea effectively blocks infection of SARS-CoV-2 and new variants by inhibiting spike binding to ACE2 receptor. *bioRxiv [Preprint]*. (2021). doi: 10.1101/2021.03.17.435637

29. Yin W, Mao C, Luan X, Shen DD, Shen Q, Su H, et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science*. (2020) 368:1499–504. doi: 10.1126/science.abc1560
30. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* (2019) 47:D1102–9. doi: 10.1093/nar/gky1033
31. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem.* (1983) 4:187–217. doi: 10.1002/jcc.540040211
32. Studio D. Dassault systemes BIOVIA, discovery studio modelling environment, release 4.5. *Accelrys Softw Inc.* (2015) 98–104.
33. Zheng J, Frisch MJ. Efficient geometry minimization and transition structure optimization using interpolated potential energy surfaces and iteratively updated hessians. *J Chem Theory Comput.* (2017) 13:6424–32. doi: 10.1021/acs.jctc.7b00719
34. Abraham MJ, van der Spoel D, Lindahl E, Hess B, Apol E, Berendsen HJC, et al. *GROMACS User Manual version 5.0.5* (2015). Available online at: <http://www.gromacs.org/>
35. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: fast, flexible, and free. *J Comput Chem.* (2005) 26:1701–18. doi: 10.1002/jcc.20291
36. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. *J Comput Chem.* (1997) 18:1463–72. doi: 10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H
37. Kumari R, Kumar R, Lynn A. G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model.* (2014) 54:1951–62. doi: 10.1021/ci500020m
38. Siriwardhana N, Kalupahana NS, Cekanova M, LeMieux M, Greer B, Moustaid-Moussa N. Modulation of adipose tissue inflammation by bioactive food compounds. *J Nutr Biochem.* (2013) 24:613–23. doi: 10.1016/j.jnutbio.2012.12.013
39. Nishimura K, Sano M, Ohtaka M, Furuta B, Umemura Y, Nakajima Y, et al. Development of defective and persistent Sendai virus vector: a unique gene delivery/expression system ideal for cell reprogramming. *J Biol Chem.* (2011) 286:4760–71. doi: 10.1074/jbc.M110.183780
40. Hillen HS, Kokic G, Farnung L, Dienemann C, Tegunov D, Cramer P. Structure of replicating SARS-CoV-2 polymerase. *Nature*. (2020) 584:154–6. doi: 10.1038/s41586-020-2368-8
41. Karplus M, McCammon JA. Molecular dynamics simulations of biomolecules. *Nat Struct Biol.* (2002) 9:646–52. doi: 10.1038/nsb0902-646
42. Hansson T, Oostenbrink C, Van Gunsteren WF. Molecular dynamics simulations. *Curr Opin Struct Biol.* (2002) 12:190–6. doi: 10.1016/S0959-440X(02)00308-1
43. Karplus M, Kuriyan J. Molecular dynamics and protein function. *Proc Natl Acad Sci USA.* (2005) 102:6679–85. doi: 10.1073/pnas.0408930102
44. Bhardwaj VK, Singh R, Das P, Purohit R. Evaluation of acridinedione analogs as potential SARS-CoV-2 main protease inhibitors and their comparison with repurposed anti-viral drugs. *Comput Biol Med.* (2021) 128:104117. doi: 10.1016/j.compbiomed.2020.104117
45. Parvez MSA, Karim MA, Hasan M, Jaman J, Karim Z, Tahsin T, et al. Prediction of potential inhibitors for RNA-dependent RNA polymerase of SARS-CoV-2 using comprehensive drug repurposing and molecular docking approach. *Int J Biol Macromol.* (2020) 163:1787–97. doi: 10.1016/j.ijbiomac.2020.09.098
46. Bhardwaj VK, Purohit R. A new insight into protein-protein interactions and the effect of conformational alterations in PCNA. *Int J Biol Macromol.* (2020) 148:999–1009. doi: 10.1016/j.ijbiomac.2020.01.212
47. Li L, Wang L, Alexov E. On the energy components governing molecular recognition in the framework of continuum approaches. *Front Mol Biosci.* (2015) 2:5. doi: 10.3389/fmolb.2015.00005

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

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



Saudi Familial Hypercholesterolemia Patients With Rare *LDLR* Stop Gain Variant Showed Variable Clinical Phenotype and Resistance to Multiple Drug Regimen

Zuhier Ahmed Awan^{1,2†}, Omran M. Rashidi^{1,3,4†}, Bandar Ali Al-Shehri^{3,4}, Kaiser Jamil⁵, Ramu Elango^{3,4†}, Jumana Y. Al-Aama^{3,4}, Robert A. Hegele^{6*}, Babajan Banaganapalli^{3,4*} and Noor A. Shaik^{3,4*}

OPEN ACCESS

Edited by:

George Priya Doss C.,
VIT University, India

Reviewed by:

Basharat Ahmad Bhat,
University of Otago, New Zealand
Maruthi Prasad E.,
China Medical University, Taiwan

*Correspondence:

Robert A. Hegele
hegele@roberts.ca
Babajan Banaganapalli
bbabajan@kau.edu.sa
Noor A. Shaik
nshaik@kau.edu.sa

[†]These authors share first authorship

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 13 April 2021

Accepted: 31 May 2021

Published: 25 June 2021

Citation:

Awan ZA, Rashidi OM, Al-Shehri BA, Jamil K, Elango R, Al-Aama JY, Hegele RA, Banaganapalli B and Shaik NA (2021) Saudi Familial Hypercholesterolemia Patients With Rare *LDLR* Stop Gain Variant Showed Variable Clinical Phenotype and Resistance to Multiple Drug Regimen. *Front. Med.* 8:694668. doi: 10.3389/fmed.2021.694668

¹ Department of Clinical Biochemistry, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, ² Department of Genetics, Al Borg Medical Laboratories, Jeddah, Saudi Arabia, ³ Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, ⁴ Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia, ⁵ Department of Genetics, Bhagwan Mahavir Medical Research Center (BMMRC), Hyderabad, India, ⁶ Departments of Medicine and Biochemistry, Schulich School of Medicine and Dentistry, Roberts Research Institute, Western University, London, ON, Canada

Familial hypercholesterolemia (FH), a well-known lipid disease caused by inherited genetic defects in cholesterol uptake and metabolism is underdiagnosed in many countries including Saudi Arabia. The present study aims to identify the molecular basis of severe clinical manifestations of FH patients from unrelated Saudi consanguineous families. Two Saudi families with multiple FH patients fulfilling the combined FH diagnostic criteria of Simon Broome Register, and the Dutch Lipid Clinic Network (DLCN) were recruited. LipidSeq, a targeted resequencing panel for monogenic dyslipidemias, was used to identify causative pathogenic mutation in these two families and in 92 unrelated FH cases. Twelve FH patients from two unrelated families were sharing a very rare, pathogenic and founder *LDLR* stop gain mutation i.e., c.2027delG (p.Gly676Alafs*33) in both the homozygous or heterozygous states, but not in unrelated patients. Based on the variant zygosity, a marked phenotypic heterogeneity in terms of LDL-C levels, clinical presentations and resistance to anti-lipid treatment regimen (ACE inhibitors, β -blockers, ezetimibe, statins) of the FH patients was observed. This loss-of-function mutation is predicted to alter the free energy dynamics of the transcribed RNA, leading to its instability. Protein structural mapping has predicted that this non-sense mutation eliminates key functional domains in *LDLR*, which are essential for the receptor recycling and LDL particle binding. In conclusion, by combining genetics and structural bioinformatics approaches, this study identified and characterized a very rare FH causative *LDLR* pathogenic variant determining both clinical presentation and resistance to anti-lipid drug treatment.

Keywords: familial hypercholesterolemia, genetic diagnosis, monogenic diseases, consanguineous populations, *LDLR* pathogenic mutations

INTRODUCTION

FH (OMIM 143890) is a relatively common metabolic disease in which patients demonstrate life-long elevation of plasma low-density lipoprotein (LDL) cholesterol (1). If left untreated, modified LDL particles enter arterial wall macrophages contributing to plaque formation, particularly within coronary arteries leading to premature development of coronary heart disease (CHD) (2). Cholesterol-laden macrophages lead to formation of not only atheromatous plaques, but also extensor tendon xanthomas (e.g., Achilles and fingers), xanthelasmata (yellow deposit underneath the skin of upper and lower eyelids), and arcus cornealis (cholesterol ring accumulating at the edge of the cornea) (3). Since FH is asymptomatic in the initial stages, most FH patients do not realize their illness until the onset of symptomatic atherosclerotic cardiovascular disease in their forties or fifties, which is sometimes fatal (4). The overall prevalence of FH in the Gulf region is estimated to be $\sim 0.43\%$ (1/232), however, in Saudi Arabia, prevalence of FH is not yet established due to the dearth of local FH clinical registries, epidemiological studies, and population genetic screening programs (5, 6). Diagnosis rates of FH are quite high among individuals who have a positive family history of premature CHD or hypercholesterolemia (7).

FH is caused by defective hepatic uptake of LDL receptor (R) mediated LDL-C particle degradation processes. About 30–60% of clinically diagnosed FH patients have a single copy of a pathogenic mutation (8). Majority of the clinically diagnosed FH patients ($\sim 80\%$) have a mono-allelic loss-of-function (LoF) variant in *LDLR* gene, while the rest are LoF variants in the receptor-binding domain of the *APOB* gene or a gain-of-function (GoF) variant in the *PCSK9* gene. A very small proportion of FH have biallelic LoF mutations in *LDLRAP1* which normally assists in LDL receptor internalization by liver cells. However, ~ 30 – 70% of clinically diagnosed FH patients are negative for *LDLR*, *APOB*, or *PCSK9* pathogenic mutations. Few have very rare LoF mutations in secondary FH genes, including *ABCG5*, *ABCG8*, *LIPA*, or *APOE*, while many other patients with hypercholesterolemia carry several common genetic variants (also called single nucleotide polymorphisms), which collectively act to influence the serum LDL-C concentration (9).

FH typically shows either autosomal dominant (HeFH; heterozygous FH) or autosomal recessive (HoFH; homozygous FH) mode of inheritance based on one or two copies of pathogenic variants in *LDLR*, *APOB* or *PCSK9* genes (10, 11). In most of the studied populations, HoFH affects 1 in 160,000–300,000 individuals, while the HeFH affects out 1 in 250–300 individuals (8). There is evidence for higher prevalence rates of HeFH in founder subpopulations like Saudi Arabians, in which consanguineous marriages are practiced as part of a social norm. Furthermore, FH is underdiagnosed all around the world, with $<5\%$ of affected individuals in many countries being identified as having FH (12). Owing to the limited data describing the genetic and phenotypic characteristics of hypercholesterolemia among Saudi patients (13–15) this study aims at identifying the inherited basis of FH in two consanguineous families from Saudi Arabia. In this study, we show that LipidSeq targeted

resequencing panel for monogenic dyslipidemias, can effectively detect FH causative *LDLR* founder variant (c.2027delG) in genetically isolated populations like Saudi Arabians.

MATERIALS AND METHODS

Recruitment of FH Patients and Their Families

The institutional Ethics Committee for Human Research of King Abdulaziz University Hospital (KAUH) gave the approval to conduct the present study according to standard international guidelines. This study has recruited FH patients from Genetic Dyslipidemia and Familial Hypercholesterolemia clinic at the King Abdulaziz University Hospital, Jeddah, Saudi Arabia. Initially two families with multiple members, fulfilling the combined FH diagnostic criteria of Simon Broome Register, and the Dutch Lipid Clinic Network (DLCN), were identified. In Simon Broome criteria for FH diagnosis, points are assigned for cholesterol concentrations, clinical characteristics, molecular diagnosis, and family history, which include risk of fatal heart disease (16). Although the Simon Broome Register criteria consider the molecular diagnosis as evidence for definite FH, the DLCN requires that at least one other criterion be met in addition to molecular diagnosis (17). All the affected individuals from these families underwent detailed physical examinations and their full family history was collected. Laboratory investigations for multiple parameters including Plasma lipid profile (LDL-C, HDL-C, Triglyceride and Total Cholesterol), blood glucose, thyroid function, and liver function were measured by a homogenous enzymatic assay. Clinical geneticist revisited the medical data of patients, interviewed them, drew three generation pedigree charts and enrolled the remaining relatives of the patient families. We have also recruited 92 unrelated FH patients following the DLCN criteria. Approximately 5 mL of blood sample (in EDTA vacutainers) was collected from all individuals after explanation of the study, along with risks and potential benefits. All the participants have signed the informed consent.

Genotyping DNA Preparation

Genomic DNA from peripheral blood cells was isolated using the standard protocols supplied by commercial extraction kits. DNA's quality and quantity were assessed with Nanodrop spectrophotometer and DNA integrity was checked with 1% agarose gel. DNA dilutions at starting concentration of 2 ng/ μ L were prepared with help of a Qubit 2.0 fluorometer.

Targeted DNA Resequencing With LipidSeq

The DNA samples of the index cases and other members from both families were sequenced on LipidSeq, a targeted resequencing panel for monogenic dyslipidemias, at London Regional Genomics Center, London, Ontario, Canada (www.lrgc.ca). This LipidSeq resequencing panel can scan pathogenic mutations in 73 genes and 185 single nucleotide polymorphisms (SNPs) associated with dyslipidemia and other metabolic disorders (18). A latest article has reviewed the utility of LipidSeq technology in successfully diagnosing the monogenic

dyslipidemias and metabolic disorders (19). The full details of DNA library preparation, sequencing, sequence alignment and variant calling (10-fold coverage and 20% read frequency) are described in the original publication (18). VarSeq software was used to annotate and prioritize the variants and for identifying the FH potential variants. Different nucleotide sequence-based prediction algorithms, such as CADD, SIFT and PolyPhen-2 which assess pathogenicity of variants were used to filter the likely deleterious variants (20). The minor allele frequency (MAF) of the variants was determined based on the data available like SHGP, 1,000 genomes, ExAC, ESP and GnomAD databases. From this data, we picked up the extremely rare (MAF is <0.01) mutation occurring in coding regions or splicing regions of the FH causative disease genes and validated its presence in remaining family members and unrelated FH cases.

Sanger Sequencing

The LipidSeq identified a potential FH causative variant was validated in index case, family members and unrelated FH cases, using the Sanger sequencing method. In brief, initially oligonucleotide primers (forward primer; 5'-CCCAACCTTGA AACCTCCTTGTGGAAA-3' and reverse primer; 5'-CCATTG ACAGATGAGCAGAGAG-3') spanning the potential mutation location were designed, followed by PCR reactions with dNTP and ddNTP mixture, and bidirectional sequencing in an automatic DNA sequencing machine. The sequence reads were analyzed with help of Bioedit program and nucleotide numbering of the mutations was done considering A of ATG code of mRNA sequence as the first nucleotide. Variant segregation in the family was determined by careful analysis of variant status in each family member.

Computational Functional Analysis of Pathogenic Mutations

Functional Analysis of Pathogenic Mutation on RNA Structure

Studying the impact of pathogenic variants on the RNA secondary structures gives hint about its possible functional consequences. Thus, we used a RNA fold (<http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi>) prediction tool which estimates the back bone traces and minimum free energy (MFE) value differences on the optimal secondary structure of RNA molecule (21). This tool intakes the native and wildtype RNA sequences in FASTA format and uses Mccaskill's algorithm (22) for computing the probabilities of base pairing matrix, partition function, and structure of centroid molecules. The output of RNA folding is an interactive string representation RNA secondary structure and a mountain plot showing the folding of energy differences between native and mutant sequences. MFE differences between native and mutant RNA structures were compared to estimate the effect of pathogenic variant on their secondary structural features.

Functional Analysis of Pathogenic Mutation on Protein Structure

Studying the impact of pathogenic variants on protein structure provides insight into the complex dynamics of genotype vs.

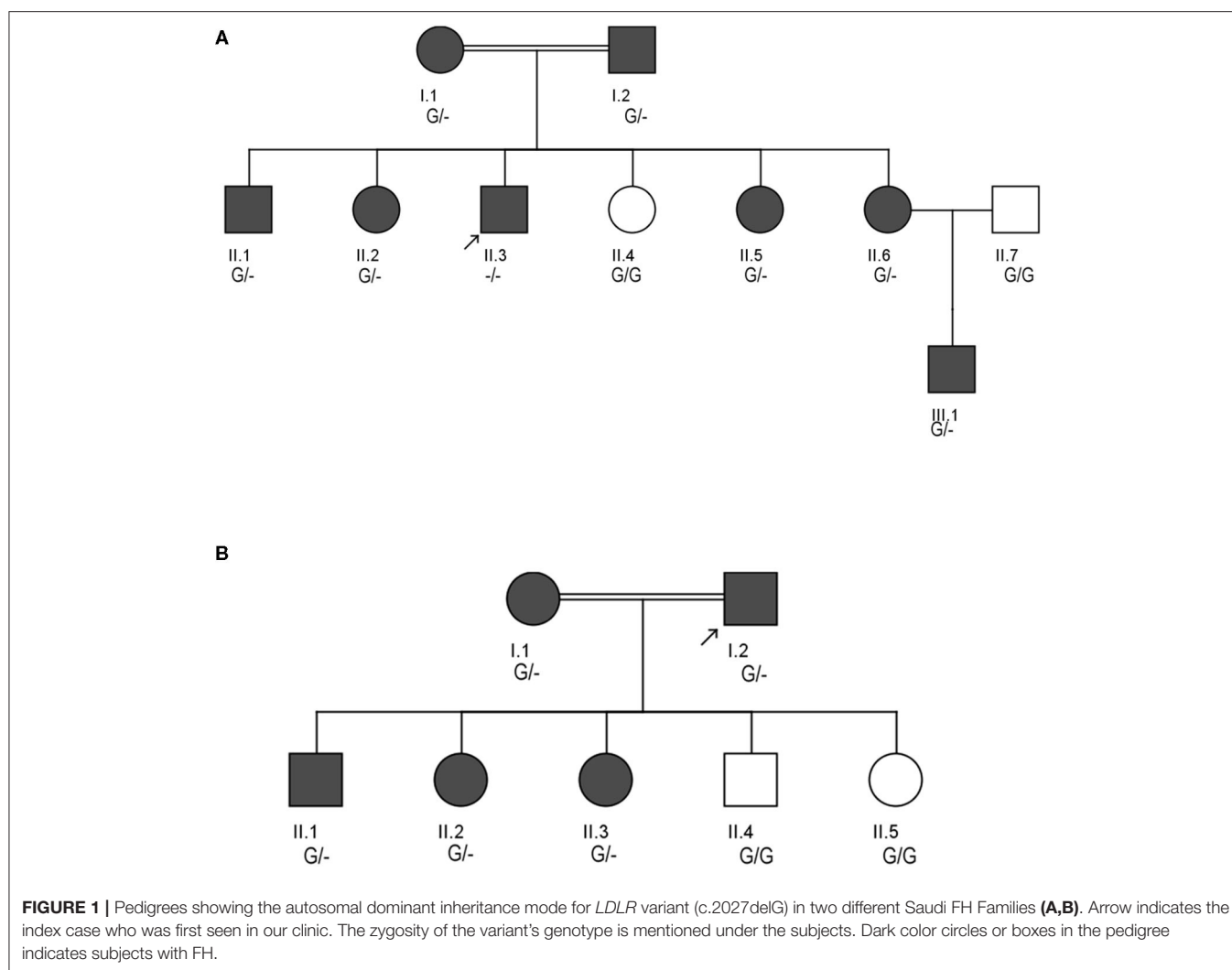
protein phenotype and structure-function relationships. In this study, we retrieved the x-ray crystallography solved tertiary structure of the query protein from Protein Data Bank (PDB). The construction of missing structural regions basing on original crystal structure coordinates was simulated through *ab-initio* method using I-Tasser webserver (23, 24). The full length tertiary model was subsequently processed for energy minimization and stereochemical assessment steps as described in our recent publication (25). We subsequently created mutant form of FH candidate protein by providing the mutant amino acid sequence and followed the similar steps involved in native protein structure modeling. The built 3D models were provided as an input to PDBSUM for examining the variant induced protein phenotype changes at secondary structure level. PyMOL software was used for visualizing and examining the salt bridges in, which the 3D models built. Stability changes induced by the variants on *LDLR* structure were estimated with help of DUET webserver (26).

RESULTS

Case Presentations in the FH Families Family A

Family A is a native Saudi Arabian family from the North-Western region (**Figure 1A**). The index case (II.3) was clinically diagnosed as FH patient at the age of nine and later presented to our clinic when he was 34 years old for his lipid management. His clinical examination revealed signs of severe hyperlipidemia. All classical manifestations of FH were present including bilateral large Achilles tendon xanthomas, huge cholesterol depositions around both mid-thighs, severe bilateral eye xanthelasma and corneal arcus, and bilateral multiple extensor tendon xanthomas on hands. His biochemical profile revealed on an average high level of total cholesterol (15.18 ± 1.33 mmol/L), LDL-C (12.98 ± 2.08 mmol/L) and normal triglycerides (0.81 ± 0.16 mmol/L) (**Table 1**). He had a past history of hospitalization after chest discomfort and shortness of breath, during which electrocardiogram generated ischemic changes were noticed. A computerized tomography (CT) of the entire aorta was performed, which showed extensive atherosclerotic calcifications in the thoracic aorta, abdominal aorta and into the iliac vessels (**Figure 2**).

At the time of his visit to our lipid clinic, II.3 was on the following regimen; dual antiplatelet agents, angiotensin converting enzyme (ACE) inhibitors, β -blockers, ezetimibe 10 mg/day and intensive statin treatment with rosuvastatin 40 mg/day, without showing any signs of improvement in his lipid profile (and even when PCSK9 inhibitors were given). He has been undergoing a bimonthly LDL- apheresis therapy in the Cardiovascular Prevention and Rehabilitation Unit of a major referral hospital in Saudi Arabia over the last 10 years. The patient reported his strict adherence to diet and medications as per the physician's instructions. With each apheresis session, his LDL cholesterol level drops by 70 to 83 percent, but within 1 week, returns back to the pre-apheresis level.



Pedigree analysis of the index case suggested positive family history of dyslipidemia, consistent with an autosomal dominant mode of inheritance. The biochemical findings of his parents (I.1 and I.2), five siblings including two brothers (II.1 and II.3), and 3 sisters (II.2, II.6 and II.6) were consistent with an FH diagnosis. However, one sister (II.4) has showed a healthy lipid profile and normal clinical features. Clinical and medication details of family members are shown in **Table 1**. Furthermore, proband has reported that three of his grandparents (both maternal and paternal), two maternal uncles, three paternal uncles, and three siblings suffered from cardiovascular complications including multiple myocardial infarctions (MI). Of the three paternal uncles, two had percutaneous coronary interventions with insertion of coronary stents; one underwent coronary artery bypass grafting (CABG) in his forties.

Family B

The second consanguineous family comes from the Southern region of Saudi Arabia (**Figure 1B**). The proband (I.2) was

referred to our clinic after undergoing CABG surgery together with replacement of two valves. His past medical history revealed that he was hypercholesterolemic since early adulthood, has undergone cardiac catheterization four times, and had multiple stent placements at the ages of 39 (1 stent), 42 (4 stents), and 44 (1 stent) due to coronary artery narrowing. At the age of 47, the patient was admitted for an open-heart surgery to perform CABG to improve blood flow and oxygen supply to the heart. Clinical examination did not reveal the presence of severe physical signs including the absence of Achilles and tendon xanthomas. The only physical finding was the presence of mild corneal arcus. At the time of his presentation at the lipid clinic, his on-treatment lipid measurements were as follows; total cholesterol 3.86 mmol/dL, LDL-C 2.69 mmol/DL, and triglyceride level 1.13 mmol/L (**Table 1**). Despite receiving combination of lipid lowering drugs i.e., ezetimibe 10 mg daily, and evolocumab subcutaneous injections 140 mg/mL once every 2 weeks, there was no improvement in his blood LDL cholesterol, which ranged between (2.47–3.09 mmol/L).

TABLE 1 | Clinical and biochemical characteristics of FH families studied in this investigation.

Family	Member	Genotype	LDL-C (mmol/L)	Reference range	TC (mmol/L)	Reference range	TG (mmol/L)	Reference range	Clinical phenotype
			Mean \pm Standard deviation		Mean \pm standard deviation		Mean \pm standard deviation		
A	I.1	G/–	8.3 \pm 1.98	0–3.57 (mmol/L)	11.0 \pm 1.01	0–5.20 (mmol/L)	0.76 \pm 0.29	0.3–2.30 (mmol/L)	Diabetes, hyperlipidemia, bilateral tendon xanthomas, history of atherosclerosis myocardial infarctions (MI).
	I.2	G/–	9.7 \pm 1.32		13.7 \pm 1.37		0.86 \pm 0.41		Hyperlipidemia, bilateral tendon xanthomas, history of peripheral atherosclerosis in both legs and history of cardiovascular disease.
	II.1	G/–	11.4 \pm 0.54		12.8 \pm 0.91		0.66 \pm 0.54		Hyperlipidemia, history of MI.
	II.2	G/–	10.1 \pm 1.11		11.9 \pm 1.11		0.78 \pm 0.37		Hyperlipidemia, history of MI.
	II.3	–/–	13.4 \pm 2.11		15.5 \pm 1.24		0.82 \pm 0.19		Statin resistance, bilateral xanthelasma, corneal arcus, bilateral tendon xanthomas, Achilles tendon xanthomas, severe and huge cholesterol depositions around both mid-thighs.
	II.4	G/G	–		–		–		–
	II.5	G/–	7.8 \pm 0.98		14.6 \pm 0.93		0.77 \pm 0.33		Hyperlipidemia
	II.6	G/–	8.2 \pm 1.22		12.8 \pm 0.87		0.87 \pm 0.23		Hyperlipidemia
B	I.1	G/–	7.6 \pm 2.40		9.6 \pm 1.33		0.75 \pm 0.56		Hyperlipidemia, history of aortic atherosclerosis
	I.2*	G/–	2.2 \pm 1.08		3.6 \pm 1.24		1.69 \pm 0.55		Sever aortic stenosis, multiple MI events and CABG surgery.
	II.1	G/–	4.7 \pm 1.51		6.3 \pm 1.81		1.73 \pm 1.01		Hyperlipidemia, chronic angina and severe chest pain.
	II.2	G/–	3.8 \pm 1.41		5.5 \pm 1.46		0.86 \pm 0.05		Diabetes, hyperlipidemia, chronic angina and severe chest pain.
	II.3	G/–	4.8 \pm 2.26		6.3 \pm 2.42		1.08 \pm 0.09		Hyperlipidemia, chronic angina and severe chest pain.

*On-treatment lipid measurements; G/G, homozygote, G/–, heterozygote, –/–, homozygote for LDLR, c.2027delG mutation.

The clinical screening, biochemical investigations and pedigree analysis of this family were consistent with an autosomal dominant mode of inheritance. As per biochemistry reports, spouse of the index case (I.1), elder son (II.1) and two daughters (II.2 and II.3) were also dyslipidemic. However, his younger son (II.4) and younger daughter (II.5) were healthy and free from any symptoms related to dyslipidemia (**Table 1**). The index case (I.2) and his wife (I.1) both have reported that their mothers have died before the age of 60 due to myocardial infarction (MI) and other heart associated related complications. Moreover, the elder sister of the index case (I.1) and younger brother of the spouse (I.2) were reported to have had open-heart surgery before their fifties due to severe MI after they had cardiac catheterizations initially at the age of 25. All these cardiac events in the family and elevated blood lipid profiles strongly

suggests premature atherosclerosis which is consistent with severe heterozygous or homozygous FH.

Genetic Analysis

The LipidSeq data of both families were analyzed for pathogenic mutations in *LDLR*, *APOB*, *PCSK9*, *ARH*, *APOE*, *ABCG5*, *ABCG8*, and *LIPA* owing to their known involvement in FH. Out of all FH candidate genes screened, only one a rare pathogenic c. 2027delG (g.11231084delG) variant localized to exon 14 of the *LDLR* gene, which is positioned on chromosome 19 p13.2 was noticed in 8 affected individuals in family A and five individuals in family B. This deletion mutation results in a frameshift in coding sequence of the *LDLR* gene and subsequently substitutes the native amino acid glycine to variant alanine at 676th position, followed by 33 nonsense residues, leading eventually

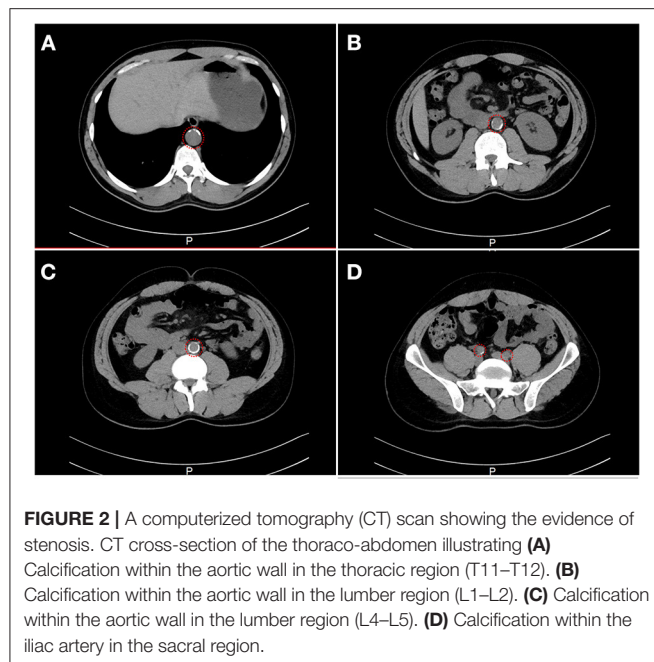


FIGURE 2 | A computerized tomography (CT) scan showing the evidence of stenosis. CT cross-section of the thoraco-abdomen illustrating (A) Calcification within the aortic wall in the thoracic region (T11–T12). (B) Calcification within the aortic wall in the lumber region (L1–L2). (C) Calcification within the aortic wall in the lumber region (L4–L5). (D) Calcification within the iliac artery in the sacral region.

to a premature stop gain signal to truncate the *LDLR* protein (UniProtKB - P01130) at 709th amino acid (G676AfsX33). This variant is expected to result in a prematurely truncated protein which likely undergoes nonsense-mediated protein decay (Figure 3).

This *LDLR* variant is listed in both dbSNP (ID: rs875989937) and ClinVar (ID: 226383). The rare prevalence of this variant is also ascertained through its absence in Exome Aggregation Consortium (ExAC), 1,000 genomes, EURO-WABB (LOVD), Greater Middle East (GME) Variome and Saudi Human Genome Program (SHGP) databases. This variant has a very low frequency of 0.0001 (six homozygotes and one heterozygote out of 4,706 individuals) in Saudi population as per the Saudi Human Genome Program database. According to variant interpretation standards and guidelines set out by American College of Medical Genetics and Genomics (ACMG), this variant is very strongly predicted to be pathogenic because it's a null variant in coding region of the *LDLR* gene whose loss of function (LOF) is a well-known mechanism for FH. Moreover, this mutation has been reported to fully segregate with FH in few Saudi families (27, 28) and is also listed in Human Genome Mutation Database (HGMD) to cause FH. Sanger sequencing results have confirmed the autosomal dominant mode of inheritance of variant in FH patients from both families. The distribution of mutation in family A is as follows; both parents (I.1 and I.2), four siblings (one brother- II.1 and three sisters- II.2, II.5 and II.6) were heterozygote carriers of the variant. Whereas, the index case (II.3) was homozygote for the c. 2027delG variant. As described in previous section, this index case has presented with severe clinical features of FH, including Achilles tendon xanthomas. However, the younger sister (II.4) of index case was homozygous for the reference G allele. In family B, the proband (I.2), his

wife (I.1), elder son (II.1) and two daughters (II.2 and II.3) were heterozygotes, whereas his younger son (II.4) and younger daughter (II.5) were homozygote carriers of the reference G allele. The genotyping results in both families corroborate with the biochemical and clinical findings. This frameshift deletion variant was found to be completely absent in an unrelated 92 FH cases tested in this study, which suggests a strong possibility that c.2027delG of the *LDLR* gene is a potential FH founder mutation in Saudi patients. Furthermore, this variant was seen to be located in evolutionarily highly conserved region of gene sequence across different species like *Gorilla*, *Paniscus*, *Pongo abelli*, *Callithrix jachchus*, *Microcebus murinus*, *Theropithecus gelada*, *Macaca fascicularis*, *Macaca mulatta* etc.

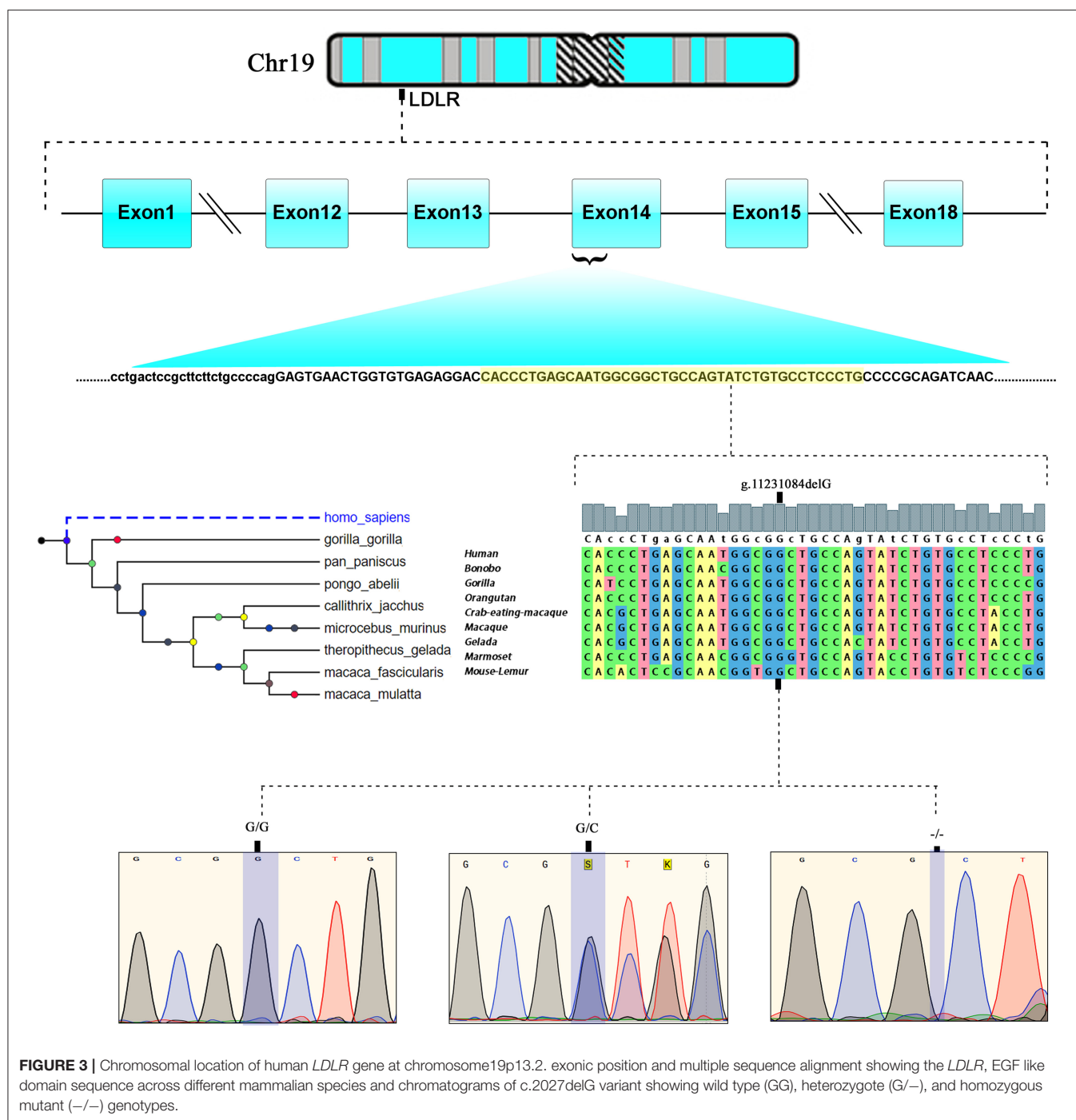
Computational Functional Analysis of Pathogenic Mutation

Functional Impact of Pathogenic Mutation on RNA Structure

The minimum free energy (MFE) calculation of *LDLR* centroid structures revealed that mutant mRNA molecule of *LDLR* (c.2027delG) possesses a relatively lower stability of secondary structures with -50.70 kcal/mol compared to native *LDLR* mRNA molecules (MEF was -52.80 kcal/mol). Hence, it is assumed that the lower stability of mRNA with c.2027delG is likely to affect the mRNA folding pattern and tertiary structure formation (Figure 4).

3-Dimensional (3D) Protein Modeling and Secondary Structure Analysis

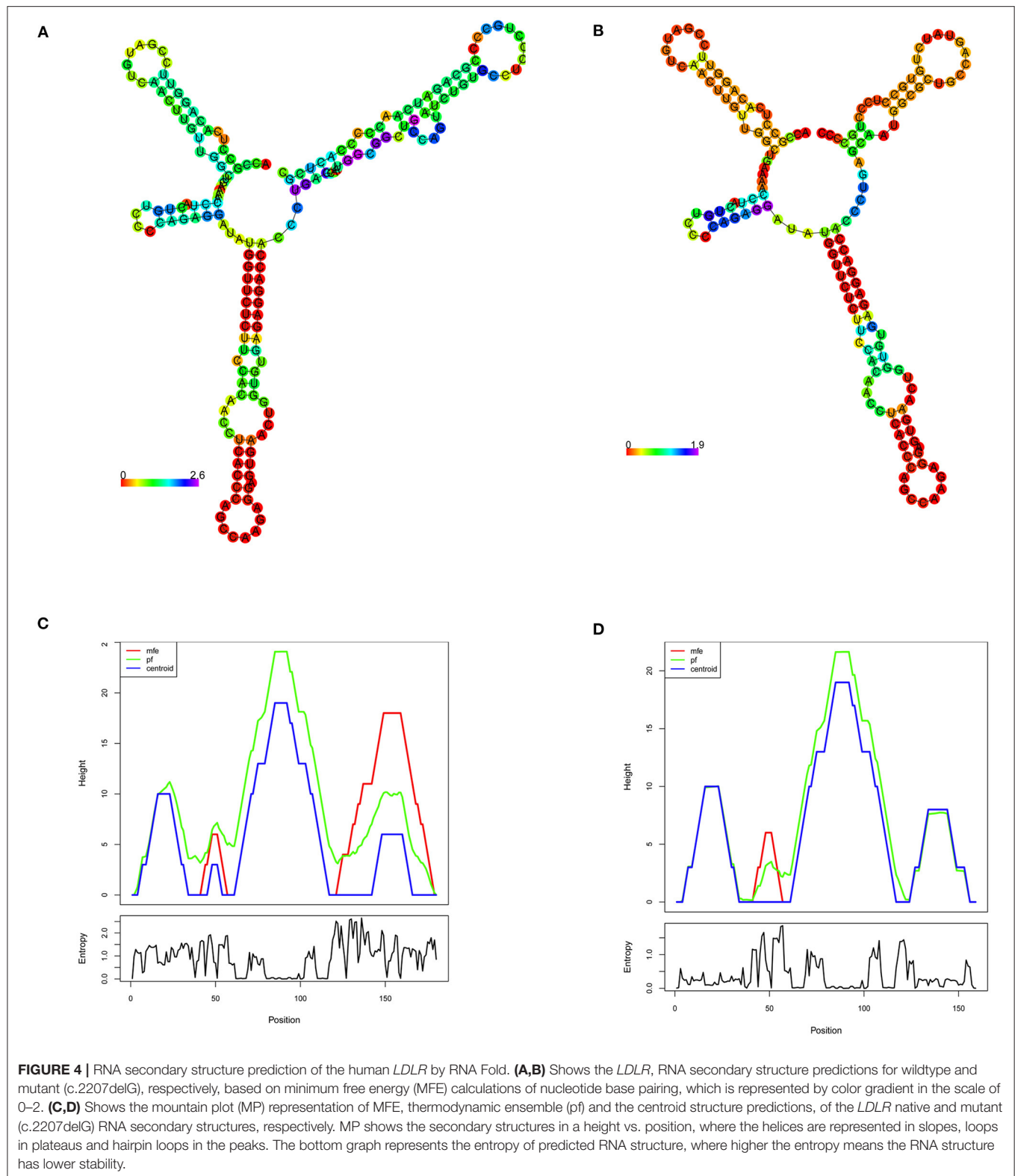
The BLASTP program search for *LDLR* protein (860 aa) identified that the PDB sourced experimentally solved structure (3M0C: chain C) has 91% (1–715 aa) of amino acid sequence coverage. However, the remaining 9% of the sequence spanning 786–860th amino acids is not yet solved. The structure of *LDLR* spans over LD repeat domain (20–311), EGF like domain (314–712), oligosaccharide linked sugars (700–758), membrane domain (residues 759–781) and cytoplasmic domain (811–860). Hence, the missing chain portions from EGF like and cytoplasmic domains were simulated by using I-Tasser, which predicted 5 probable models. The best fit *LDLR* model was selected based on confidence (1.25), template modeling (0.54 ± 0.12) and root mean square deviation scores (4.5 ± 2.8). The built protein models were subsequently energy minimized and taken as reference in constructing mutant model, which were later used to predict the effect of variant on secondary structural features. The native *LDLR* secondary structure is characterized by 3 α -helices, 11 sheets, 164 β -strands (11 β -sheets, 22 β -hairpins, 20 β -bulges, 77 β -turns, 34 β -pleated strands), 12 loops (12 γ -turns) and 243 other components like disulphide bridges. The G676A missense variant is localized to 3rd helix, does not change the secondary structure conformation as such, but truncation of the protein at Asp 707 residue eliminates/skips 34th β -strand and 12th loop spanning from 714 to 860th amino acid toward C-terminal region of the protein (Figure 5).



DISCUSSION

FH occurs in two clinical forms, namely homozygous or heterozygous, depending upon the gene dosage of the variant alleles, i.e., bi-allelic and mono-allelic, respectively (29). In the current study, we identified both homozygous (one patient) and several heterozygous FH patients (12 patients) from two unrelated Saudi Bedouin families bearing a pathogenic c.207delG (p.G676Afs*33) stop gain mutation. This mutation

was reported as c.206delG in 5 among 4 HoFH patients and 14 HeFH patients belonging to different tribes from Saudi Arabia (27, 28). So far, five different *LDLR* mutations (p.D445*, p.R471R, p.G676Afs*33, p.Y419D, p.W577*) were identified in 27 FH patients from five studies from Saudi Arabia (Table 2). Hence, most likely c.207delG is the founder FH mutation in Saudi population, where both inter- and intra- consanguineous marriages among tribal communities is a normal practice. A few other FH founder mutations in the *LDLR* gene have been



previously identified among French Canadians from Quebec Province (33–35), Finnish from Finland (36) and Dutch from Netherlands (37).

In the current study, we noticed that a clinically severe patient with a phenotype resembling HoFH (II.3) from family A, has inherited two copies *LDLR* c.2027delG stop gain

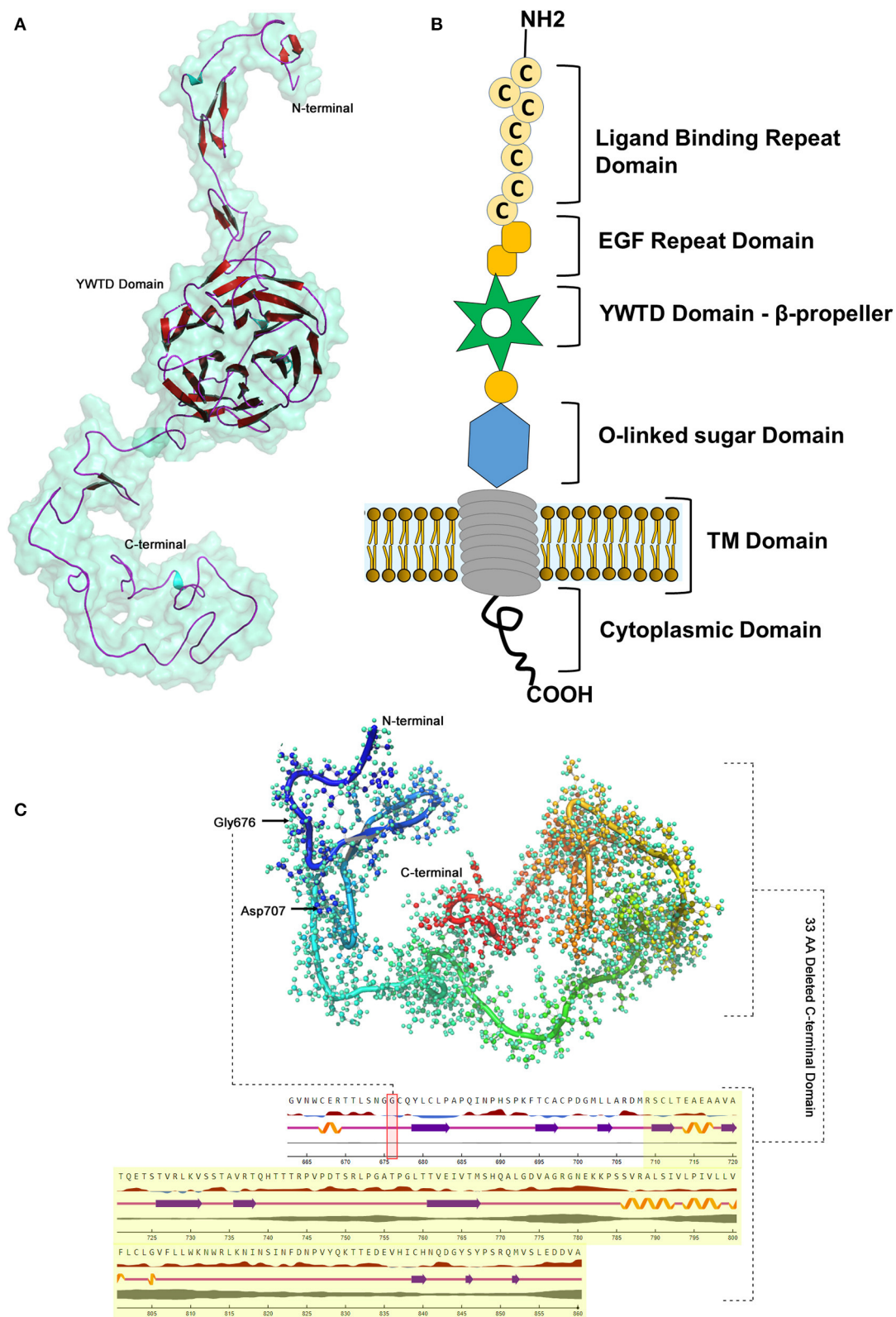


FIGURE 5 | *LDLR* protein structure visualization. **(A)** 3D structural representation of the protein molecule. **(B)** Functional domains distribution. **(C)** The p.G676Afs*33 variant effect on the secondary structure organization on the protein.

TABLE 2 | *LDLR* variants reported in Saudi FH patients.

S.No	Variant ID	Nucleotide change	Amino acid change	Exon	Variant Nature	No of Patients	FH Zygosity	Reference
1	–	c.1332dupA	p.D445*	9	Missense	2	HoFH	(30)
	rs5930	c.1413A>G	p.R471R	10	Silent	2	HoFH	
2	–	c.2026delG	p.G676Afs*33	14	Frameshift	1	HeFH	(27)
3	–	c.2027delG	p.G676Afs*33	14	Frameshift	11	HeFH	(28)
					Frameshift	9	HoFH	
	rs5930	c.1413A>G	p.R471R	10	Silent	8	HeFH	
4	–	c.1255T>G	p.Y419D	9	Missense	1	HoFH	(31)
5	–	c.1731G>T	p.W577*	12	Missense	3	HeFH	(32)
6	–	c.2027delG	p.G676Afs*33	14	Frameshift	11	HeFH	Present Study
					Frameshift	1	HoFH	

mutation from his HeFH parents (i.e., bi-allelic), although one defective copy is sufficient enough to develop the disease in heterozygotes. The severe clinical signs observed in this patient could undoubtedly be the result of an extremely compromised receptor capacity, a null variant with <2% functional activity, therefore homozygosity would explain their particularly severe biochemical and clinical phenotype. The LDL-C levels in HoFH due to bi-allelic null variants can increase by 4 to 10-fold from normal (38). It is notable that this HoFH patient despite having severe coronary stenoses, has not yet reported any cardiovascular events like myocardial infarctions (MI). Although this HoFH patient has presented with severe atherosclerosis, we speculate that early disease diagnosis (9 years), continuous medications, regular clinical monitoring and lifestyle modifications, would have averted the possibility of severe or fatal cardiac event. Literature review suggests that untreated HoFH patients by their 2nd decade of life present with CVD due to the development of advance atherosclerotic plaques and stenosis in blood vessels (38). A clinical assessment and follow-up study of 39 HoFH patients under <16 years of age reported cardiovascular events in 88% of the subjects (39). The HoFH patients show worse prognosis even on maximum treatment doses of lipid lowering drugs these patients show LDL-C levels >7.8 mmol/L. Our HoFH patient has also manifested tendon xanthomas which are known to be formed by huge cholesterol depositions in the tendons and joints that may account for pain and disability (38). Arcus cornealis, are bright zone of cholesterol deposits around the rim of the cornea before the age of 40 and is an additional clinical feature observed in HoFH patients (40).

Most FH patients from both family (A and B) are heterozygous and carry one copy of the *LDLR*, c.2027delG stop mutation, which could have been inherited from either of their parents following an autosomal dominant mode of inheritance. These patients demonstrate 2/3rd reduction in LDL clearance rate, which subsequently elevates the circulating LDL-C by 2 to 3-folds (5–10 mmol/L; 200–400 mg/L) (10, 29, 41). FH patients manifest the disease in their adulthood, spanning 3–7th decades of their life (10, 41). We have observed that the majority of our HeFH patients have presented with MI before their 60th birthday. These clinical signs point us toward chronic deposits of cholesterol that induce arterial atherosclerotic damages. Careful

physical examination in childhood often could prompts the early clinical diagnosis of HoFH. The undiagnosed and untreated HeFH patients have a very high risk of (10 to 20-folds) of developing premature coronary artery disease (CAD) (42), while the risk in untreated HoFH can be 100 to 200-fold increased from normal (38). The variable expressivity of FH can be attributed to modifier variants in *LDLRAP1*, *EPHX2*, *ABCG5*, *ABCG8*, *LIPA*, or *APOE* genes or polygenic risk variants (2, 43, 44). Early intervention to control the high LDL-C levels is clearly beneficial in reducing the cardiovascular events among young FH patients (45). The genetic testing of *LDLR*, c.2027delG variant in extended family members of both family A and B could potentially offer an advantage of early identification of FH cases, planning lifelong lipid lowering therapy, genetic counseling, and prenatal diagnosis (46).

The *LDLR* allelic makeup determines the molecular diagnosis (i.e., heterozygous vs. homozygous) and in turn determines the severity of clinical manifestation in FH. HoFH patients who have lost most or all receptor function, show very high circulating LDL-C levels and manifest cholesterol deposits in the body compared to HeFH patients who can still maintain 50% of functional receptors (47–49). Deleterious *LDLR* mutations are known to either eliminate or considerably reduce the *LDLR* function (50). This is true for the c.2027delG (p.G676Afs*33) frame shift mutation identified in this study, that introduces a protein termination codon (PTC) at 33rd codon downstream, and leads to the truncation of the *LDLR* protein at the Asp709th residue located in *EGF like domain*. The truncated protein will be 152 amino acid (aa) shorter than the native *LDLR* and lacks 5 aa residues from the *EGF like domain*, 58 aa residues of oligosaccharide linked sugars domain and 22 aa residues of the receptor transmembrane domain. Spanning between 314 and 712 aa is the *EGF like domain* the largest *LDLR* protein domain where more than 50% of FH causative mutations are reported highlighting its functional importance. *EGF* homology domain controls the release of lipoproteins in low pH environment and take part in receptor recycling (51). It is therefore fair to assume that the truncated *LDLR* may undergo degradation that may reduce the *LDLR* protein level and subsequently interfere with the receptor assembly in those individuals who harbor the mutation.

Indeed, the mRNAs with PTC may potentially activates nonsense mediated mRNA decay (NMD) (52) or leave a markedly truncated LDLR protein (at variant residue 709) which may eventually undergo degradation by ubiquitin mediated proteasomal pathway (53). The correlation between *LDLR* variant zygosity and its effect on protein expression is shown through functional biology experiments on lymphocytes obtained from FH patients with nonsense mutations (W23X, S78X, E207X, and W541X) (52). Human mRNAs with PTC were reported to reduce both mRNA abundance and stability (54). Our free energy-based RNA stability investigation predicted that the *LDLR*, p.G676Afs*33 variant destabilizes the mRNA structure, and eventually affect its folding pattern. Free energy changes are also affect the protein structure stability for disease causative stop gain mutations (55). Therefore, we predict that p.G676Afs*33 is a loss-of-function (LoF) variant or null allele leading to abolition of LDLR protein synthesis, hence it belongs to *LDLR* class 1 category of mutations (56).

Recent advances in molecular and cell biology present an alternative therapeutic opportunity to counter the genetic defects (PTC mutations) by pharmacological modulation (NMD or proteasomal degradation inhibitors) and help to control disease pathogenesis (54, 57, 58). In theory, HoFH patients who are unresponsive to statins might benefit from the use of pharmacological modulators like aminoglycosides in restoring partial sized LDLR molecules arising from stop gain mutations. In experiments, treating lymphocytes bearing different nonsense FH mutations with different translation modulator drugs, Holla *et al* has successfully demonstrated the increased mRNA levels of *LDLR* and LDL-C clearance (52). There are numerous other treatments in development for severe HeFH and HoFH, including inhibitors of angiotensin like 3 protein (ANGPTL3), long-acting inhibitors of proprotein convertase subtilisin/kexin type 9 (PCSK9) in addition to gene therapies (59).

In conclusion, this study reports a very rare, pathogenic and FH founder *LDLR* stop gain variant i.e. c.2027delG (p.Gly676Alafs*33) in 12 FH patients belonging to two different Saudi families. Founder FH mutations concentrate due to low genetic variability in genetically isolated populations. Nevertheless, identifying FH founder confer advantages like targeted screening, early genetic diagnosis, genetic counseling and adoption of effective anti-lipid treatment strategies (i.e., LDL-apheresis) to prevent the cardiovascular disease burden among the FH patients. Based on the variant zygosity of the stop

gain variant, we have noticed marked phenotypic heterogeneity in terms of LDL-C levels and clinical presentations of the FH patients. This loss-of-function variant was predicted to alter the free energy dynamics of RNA molecule hence its stability. Protein structure mapping has predicted that this variant eliminates key *LDLR* functional domains and eventually undergoes degradation. However, future functional biology studies are required to study the effect of c.2027delG variant on LDL-C clearance in the body.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because (a) participants' refusal to store or distribute the genomic data in the public domain and (b) as per the local Institutional Ethics committee approval and Saudi national policy on genomic data sharing in the public domain outside the country. Requests to access the datasets should be directed to NS or BB.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics Committee for Human Research of King Abdulaziz University Hospital. Written informed consent to participate in this study was provided by the participant and legal guardian/next of kin of the children below 16 years.

AUTHOR CONTRIBUTIONS

ZA, BB, RE, JA-A, and NS: conceptualization. OR, NS, and BB: data curation. RH, ZA, and OR: formal analysis. NS: funding acquisition and project administration. OR, RH, BA-S, NS, and BB: methodology. BB: software. ZA, RH, NS, BB, and RE: supervision. OR and KJ: validation. BB: visualization. ZA, OR, RE, BB, NS, and RH: writing original draft and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research work was funded by Institutional Fund Projects under grant no (IFPRC-059-140-2020). Therefore, authors gratefully acknowledge technical and financial support from the Ministry of Education and King Abdulaziz University, Jeddah, Saudi Arabia.

REFERENCES

- Bouhairie VE, Goldberg AC. Familial hypercholesterolemia. *Cardiol Clin*. (2015) 33:169–79. doi: 10.1016/j.ccl.2015.01.001
- Trinder M, Paquette M, Cermakova L, Ban MR, Hegele RA, Baass A, et al. Polygenic contribution to low-density lipoprotein cholesterol levels and cardiovascular risk in monogenic familial hypercholesterolemia. *Circ Genom Precis Med*. (2020) 13:515–23. doi: 10.1161/CIRCGEN.120.002919
- Taylor B, Cheema A, Soslowsky L. Tendon pathology in hypercholesterolemia and familial hypercholesterolemia. *Curr Rheumatol Rep*. (2017) 19:76. doi: 10.1007/s11926-017-0704-2
- Singh S, Bittner V. Familial hypercholesterolemia—epidemiology, diagnosis, and screening. *Curr Atheroscler Rep*. (2015) 17:482. doi: 10.1007/s11883-014-0482-5
- Bamimore MA, Zaid A, Banerjee Y, Al-Sarraf A, Abifadel M, Seidah NG, et al. Familial hypercholesterolemia mutations in the Middle Eastern and North African region: a need for a national registry. *J Clin Lipidol*. (2015) 9:187–94. doi: 10.1016/j.jacl.2014.11.008
- Alallaf F, Fa HN, Alnefaie M, Almaymuni A, Rashidi OM, Alhabib K, et al. The spectrum of Familial Hypercholesterolemia (FH) in Saudi Arabia: prime time for patient FH registry. *Open Cardiovasc Med J*. (2017) 11:66–75. doi: 10.2174/1874192401711010066

7. Turgeon RD, Barry AR, Pearson GJ. Familial hypercholesterolemia: review of diagnosis, screening, and treatment. *Can Fam Physician*. (2016) 62:32–7.
8. Berberich AJ, Hegele RA. The complex molecular genetics of familial hypercholesterolaemia. *Nat Rev Cardiol*. (2019) 16:9–20. doi: 10.1038/s41569-018-0052-6
9. Paththinige CS, Sirisena ND, Dissanayake V. Genetic determinants of inherited susceptibility to hypercholesterolemia - a comprehensive literature review. *Lipids Health Dis*. (2017) 16:103. doi: 10.1186/s12944-017-0488-4
10. Parihar RK, Razaq M, Saini G. Homozygous familial hypercholesterolemia. *Indian J Endocrinol Metab*. (2012) 16:643–5. doi: 10.4103/2230-8210.98032
11. McGowan MP, Hosseini Dehkordi SH, Moriarty PM, Duell PB. Diagnosis and treatment of heterozygous familial hypercholesterolemia. *J Am Heart Assoc*. (2019) 8:e013225. doi: 10.1161/JAHA.119.013225
12. Nordestgaard BG, Benn M. Genetic testing for familial hypercholesterolaemia is essential in individuals with high LDL cholesterol: who does it in the world? *Eur Heart J*. (2017) 38:1580–3. doi: 10.1093/eurheartj/ehx136
13. Mahzari M, Zarif H. Homozygous familial hypercholesterolemia (HoFH) in Saudi Arabia and two cases of lomitapide use in a real-world setting. *Adv Ther*. (2021) 38:2159–69. doi: 10.1007/s12325-021-01720-y
14. Albakheet N, Al-Shawi Y, Bafaqeh M, Fatani H, Orz Y, Shami I. Familial hypercholesterolemia with bilateral cholesterol granuloma: a case series. *Int J Surg Case Rep*. (2019) 62:135–9. doi: 10.1016/j.ijscr.2019.07.018
15. Awan ZA, Bondagji NS, Bamimore MA. Recently reported familial hypercholesterolemia-related mutations from cases in the Middle East and North Africa region. *Curr Opin Lipidol*. (2019) 30:88–93. doi: 10.1097/MOL.0000000000000586
16. Henderson R, O'kane M, McGilligan V, Watterson S. The genetics and screening of familial hypercholesterolaemia. *J Biomed Sci*. (2016) 23:39. doi: 10.1186/s12929-016-0256-1
17. Austin MA, Hutter CM, Zimmern RL, Humphries SE. Genetic causes of monogenic heterozygous familial hypercholesterolemia: a HuGE prevalence review. *Am J Epidemiol*. (2004) 160:407–20. doi: 10.1093/aje/kwh236
18. Johansen CT, Dube JB, Loyzer MN, Macdonald A, Carter DE, McIntyre AD, et al. LipidSeq: a next-generation clinical resequencing panel for monogenic dyslipidemias. *J Lipid Res*. (2014) 55:765–72. doi: 10.1194/jlr.D045963
19. Dron JS, Wang J, McIntyre AD, Iacocca MA, Robinson JF, Ban MR, et al. Six years' experience with LipidSeq: clinical and research learnings from a hybrid, targeted sequencing panel for dyslipidemias. *BMC Med Genomics*. (2020) 13:23. doi: 10.1186/s12920-020-0669-2
20. Wang J, Dron JS, Ban MR, Robinson JF, McIntyre AD, Alazzam M, et al. Polygenic versus monogenic causes of hypercholesterolemia ascertained clinically. *Arterioscler Thromb Vasc Biol*. (2016) 36:2439–45. doi: 10.1161/ATVBAHA.116.308027
21. Lorenz R, Bernhart SH, Honer Zu Siederdisen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*. (2011) 6:26. doi: 10.1186/1748-7188-6-26
22. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*. (1990) 29:1105–19. doi: 10.1002/bip.360290621
23. George Priya Doss C, Nagasundaram N, Chakraborty C, Chen L, Zhu H. Extrapolating the effect of deleterious nsSNPs in the binding adaptability of flavopiridol with CDK7 protein: a molecular dynamics approach. *Hum Genomics*. (2013) 7:10. doi: 10.1186/1479-7364-7-10
24. Thirumal Kumar D, Jerushah Emerald L, George Priya Doss C, Sneha P, Siva R, Charles Emmanuel Jebaraj W, et al. Computational approach to unravel the impact of missense mutations of proteins (D2HGDH and IDH2) causing D-2-hydroxyglutaric aciduria 2. *Metab Brain Dis*. (2018) 33:1699–710. doi: 10.1007/s11011-018-0278-3
25. Ahmed Awan Z, Bima A, Rashidi OM, Jamil K, Khan IA, Almukadi HS, et al. Low resolution protein mapping and KB-R7943 drug-protein molecular interaction analysis of long-QT syndrome linked KCNH2 mutations. *All Life*. (2020) 13:183–93. doi: 10.1080/26895293.2020.1737249
26. Shaik NA, Bokhari HA, Masoodi TA, Shetty PJ, Ajabnoor GMA, Elango R, et al. Molecular modelling and dynamics of CA2 missense mutations causative to carbonic anhydrase 2 deficiency syndrome. *J Biomol Struct Dynam*. (2020) 38:4067–80. doi: 10.1080/07391102.2019.1671899
27. Al-Allaf FA, Athar M, Abduljaleel Z, Taher MM, Khan W, Ba-Hammam FA, et al. Next generation sequencing to identify novel genetic variants causative of autosomal dominant familial hypercholesterolemia associated with increased risk of coronary heart disease. *Gene*. (2015) 565:76–84. doi: 10.1016/j.gene.2015.03.064
28. Al-Allaf FA, Alashwal A, Abduljaleel Z, Taher MM, Siddiqui SS, Bouazzaoui A, et al. Identification of a recurrent frameshift mutation at the LDLR exon 14 (c.2027delG, p.(G676Afs*33)) causing familial hypercholesterolemia in Saudi Arab homozygous children. *Genomics*. (2016) 107:24–32. doi: 10.1016/j.ygeno.2015.12.001
29. Hovingh GK, Davidson MH, Kastelein JJ, O'Connor AM. Diagnosis and treatment of familial hypercholesterolaemia. *Eur Heart J*. (2013) 34:962–71. doi: 10.1093/eurheartj/ehx015
30. Al-Allaf FA, Athar M, Abduljaleel Z, Bouazzaoui A, Taher MM, Own R, et al. Identification of a novel nonsense variant c.1332dup, p.(D445*) in the LDLR gene that causes familial hypercholesterolemia. *Hum Genome Var*. (2014) 1:14021. doi: 10.1038/hgv.2014.21
31. Alnouri F, Athar M, Al-Allaf FA, Abduljaleel Z, Taher MM, Bouazzaoui A, et al. Novel combined variants of LDLR and LDLRAP1 genes causing severe familial hypercholesterolemia. *Atherosclerosis*. (2018) 277:425–33. doi: 10.1016/j.atherosclerosis.2018.06.878
32. Al-Allaf FA, Alashwal A, Abduljaleel Z, Taher MM, Bouazzaoui A, Albalkhail H, et al. Compound heterozygous LDLR variant in severely affected familial hypercholesterolemia patient. *Acta Biochim Pol*. (2017) 64:75–9. doi: 10.18388/abp.2016_1283
33. Simard LR, Viel J, Lambert M, Paradis G, Levy E, Delvin EE, et al. The Delta>15 Kb deletion French Canadian founder mutation in familial hypercholesterolemia: rapid polymerase chain reaction-based diagnostic assay and prevalence in Quebec. *Clin Genet*. (2004) 65:202–8. doi: 10.1111/j.0009-9163.2004.00223.x
34. Leitersdorf E, Tobin EJ, Davignon J, Hobbs HH. Common low-density lipoprotein receptor mutations in the French Canadian population. *J Clin Invest*. (1990) 85:1014–23. doi: 10.1172/JCI114531
35. Bétard C, Kessling AM, Roy M, Chamberland A, Lussier-Cacan S, Davignon J. Molecular genetic evidence for a founder effect in familial hypercholesterolemia among French Canadians. *Hum Genet*. (1992) 88:529–36. doi: 10.1007/BF00219339
36. Lahtinen AM, Havulinna AS, Jula A, Salomaa V, Kontula K. Prevalence and clinical correlates of familial hypercholesterolemia founder mutations in the general population. *Atherosclerosis*. (2015) 238:64–9. doi: 10.1016/j.atherosclerosis.2014.11.015
37. Kusters DM, Huijgen R, Defesche JC, Vissers MN, Kindt I, Hutten BA, et al. Founder mutations in the Netherlands: geographical distribution of the most prevalent mutations in the low-density lipoprotein receptor and apolipoprotein B genes. *Netherlands Heart J*. (2011) 19:175–82. doi: 10.1007/s12471-011-0076-6
38. Cuchel M, Bruckert E, Ginsberg HN, Raal FJ, Santos RD, Hegele RA, et al. Homozygous familial hypercholesterolaemia: new insights and guidance for clinicians to improve detection and clinical management. A position paper from the Consensus Panel on Familial Hypercholesterolaemia of the European Atherosclerosis Society. *Eur Heart J*. (2014) 35:2146–57. doi: 10.1093/eurheartj/ehu274
39. Kolansky DM, Cuchel M, Clark BJ, Paridon S, McCRindle BW, Wiegers SE, et al. Longitudinal evaluation and assessment of cardiovascular disease in patients with homozygous familial hypercholesterolemia. *Am J Cardiol*. (2008) 102:1438–43. doi: 10.1016/j.amjcard.2008.07.035
40. Kim YR, Han KH. Familial hypercholesterolemia and the atherosclerotic disease. *Korean Circ J*. (2013) 43:363–7. doi: 10.4070/kcj.2013.43.6.363
41. Mytilinaoui M, Kyrou I, Khan M, Grammatopoulos DK, Randevas HS. Familial hypercholesterolemia: new horizons for diagnosis and effective management. *Front Pharmacol*. (2018) 9:707. doi: 10.3389/fphar.2018.00707
42. Goldberg AC, Hopkins PN, Toth PP, Ballantyne CM, Rader DJ, Robinson JG, et al. Familial hypercholesterolemia: screening, diagnosis and management of pediatric and adult patients: clinical guidance from the National Lipid Association Expert Panel on Familial Hypercholesterolemia. *J Clin Lipidol*. (2011) 5:S1–8. doi: 10.1016/j.jacl.2011.03.001
43. Kamar A, Khalil A, Nemer G. The digenic causality in familial hypercholesterolemia: revising the genotype-phenotype correlations of the disease. *Front Genet*. (2021) 11:572045. doi: 10.3389/fgene.2020.572045

44. Fahed AC, Khalaf R, Salloum R, Andary RR, Safa R, El-Rassy I, et al. Variable expressivity and co-occurrence of LDLR and LDLRAP1 mutations in familial hypercholesterolemia: failure of the dominant and recessive dichotomy. *Mol Genet Genomic Med.* (2016) 4:283–91. doi: 10.1002/mgg3.203
45. Besseling J, Kindt I, Hof M, Kastelein JJ, Hutten BA, Hovingh GK. Severe heterozygous familial hypercholesterolemia and risk for cardiovascular disease: a study of a cohort of 14,000 mutation carriers. *Atherosclerosis.* (2014) 233:219–23. doi: 10.1016/j.atherosclerosis.2013.12.020
46. Cohen H, Stefanutti C, Di Giacomo S, Morozzi C, Widhalm K, Bjelakovic BB, et al. (2021). Current approach to the diagnosis and treatment of heterozygote and homozygous FH children and adolescents. *Curr Atheroscl Rep.* 23:30. doi: 10.1007/s11883-021-00926-3
47. Goldstein JL, Brown MS. Molecular medicine. The cholesterol quartet. *Science.* (2001) 292:1310–2. doi: 10.1126/science.1061815
48. Dedoussis GV, Skoumas J, Pitsavos C, Choumerianou DM, Genschel J, Schmidt H, et al. FH clinical phenotype in Greek patients with LDLR defective vs. negative mutations. *Eur J Clin Invest.* (2004) 34:402–9. doi: 10.1111/j.1365-2362.2004.01351.x
49. Weiss N, Binder G, Keller C. Mutations in the low-density-lipoprotein receptor gene in German patients with familial hypercholesterolaemia. *J Inherit Metab Dis.* (2000) 23:778–90. doi: 10.1023/A:1026704517598
50. Jiang L, Sun LY, Dai YF, Yang SW, Zhang F, Wang LY. The distribution and characteristics of LDL receptor mutations in China: a systematic review. *Sci Rep.* (2015) 5:17272. doi: 10.1038/srep17272
51. Rudenko G, Deisenhofer J. The low-density lipoprotein receptor: ligands, debates and lore. *Curr Opin Struct Biol.* (2003) 13:683–9. doi: 10.1016/j.sbi.2003.10.001
52. Holla ØL, Kulseth MA, Berge KE, Leren TP, Ranheim T. Nonsense-mediated decay of human LDL receptor mRNA. *Scand J Clin Lab Invest.* (2009) 69:409–17. doi: 10.1080/00365510802707163
53. Li Y, Lu W, Schwartz AL, Bu G. Degradation of the LDL receptor class 2 mutants is mediated by a proteasome-dependent pathway. *J Lipid Res.* (2004) 45:1084–91. doi: 10.1194/jlr.M300482-JLR200
54. Kurosaki T, Maquat LE. Nonsense-mediated mRNA decay in humans at a glance. *J Cell Sci.* (2016) 129:461–7. doi: 10.1242/jcs.181008
55. Kamal NM, Sahly AN, Banaganapalli B, Rashidi OM, Shetty PJ, Al-Aama JY, et al. Whole exome sequencing identifies rare biallelic ALMS1 missense and stop gain mutations in familial Alström syndrome patients. *Saudi J Biol Sci.* (2020) 27:271–8. doi: 10.1016/j.sjbs.2019.09.006
56. Galicia-Garcia U, Benito-Vicente A, Uribe KB, Jebari S, Larrea-Sebal A, Alonso-Estrada R, et al. Mutation type classification and pathogenicity assignment of sixteen missense variants located in the EGF-precursor homology domain of the LDLR. *Sci Rep.* (2020) 10:1727. doi: 10.1038/s41598-020-58734-9
57. Kirimtay K, Temizci B, Gültekin M, Yapici Z, Karabay A. Novel mutations in ATP13A2 associated with mixed neurological presentations and iron toxicity due to nonsense-mediated decay. *Brain Res.* (2020) 1750:147167. doi: 10.1016/j.brainres.2020.147167
58. Pawlicka K, Kalathiya U, Alfaro J. Nonsense-mediated mRNA decay: pathologies and the potential for novel therapeutics. *Cancers (Basel).* (2020) 12:765. doi: 10.3390/cancers12030765
59. Hegele RA, Tsimikas S. Lipid-lowering agents. *Circ Res.* (2019) 124:386–404. doi: 10.1161/CIRCRESAHA.118.313171

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

Copyright © 2021 Awan, Rashidi, Al-Shehri, Jamil, Elango, Al-Aama, Hegele, Banaganapalli and Shaik. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



rs12537 Is a Novel Susceptibility SNP Associated With Estrogen Receptor Positive Breast Cancer in Chinese Han Population

Jingkai Xu^{1,2†}, Guozheng Li^{3,4†}, Mengyun Chen^{2†}, Wenjing Li^{3,4}, Yaxing Wu^{3,4}, Xuejun Zhang², Yong Cui^{1*†} and Bo Zhang^{3,4*†}

¹ Department of Dermatology, China-Japan Friendship Hospital, Beijing, China, ² Department of Dermatology, The First Affiliated Hospital of Anhui Medical University, Hefei, China, ³ School of Life Sciences, Anhui Medical University, Hefei, China, ⁴ Department of Oncology, No. 2 Hospital, Anhui Medical University, Hefei, China

OPEN ACCESS

Edited by:

Thirumal Kumar D,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

J. Francis Borgio,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia
Udhaya Kumar S,
Vellore Institute of Technology, India

*Correspondence:

Yong Cui
wuhucuiyong@vip.163.com
Bo Zhang
alvinbo@163.com

[†]These authors have contributed
equally to this work

[‡]These authors share senior
authorship

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 12 May 2021

Accepted: 21 June 2021

Published: 28 July 2021

Citation:

Xu J, Li G, Chen M, Li W, Wu Y,
Zhang X, Cui Y and Zhang B (2021)
rs12537 Is a Novel Susceptibility SNP
Associated With Estrogen Receptor
Positive Breast Cancer in Chinese Han
Population. *Front. Med.* 8:708644.
doi: 10.3389/fmed.2021.708644

Genetic testing is widely used in breast cancer and has identified a lot of susceptibility genes and single nucleotide polymorphisms (SNPs). However, for many SNPs, evidence of an association with breast cancer is weak, underlying risk estimates are imprecise, and reliable subtype-specific risk estimates are not in place. A recent genome-wide long non-coding RNA (lncRNA) association study in Chinese Han has verified a genetic association between rs12537 and breast cancer. This study is aimed at investigating the association between rs12537 and the phenotype. We collected the clinical information of 5,634 breast cancer patients and 6,308 healthy controls in the early study. And χ^2 test was used for the comparison between different groups in genotype. The frequency of genotypic distribution among SNP rs12537 has no statistically significant correlation with family history ($p = 0.8945$), menopausal status ($p = 0.3245$) or HER-2 ($p = 0.2987$), but it is statistically and significantly correlated with ER ($p = 0.004006$) and PR ($p = 0.01379$). Most importantly, compared to the healthy control, rs12537 variant is significantly correlated with ER positive patients and the p -value has reached the level of the whole genome ($p = 1.66E-08 < 5.00E-08$). Furthermore, we found rs12537 associated gene *MTMR3* was lower expressed in breast cancer tissues but highly methylated. In conclusion, our findings indicate that rs12537 is a novel susceptibility gene in ER positive breast cancer in Chinese Han population and it may influence the methylation of *MTMR3*.

Keywords: breast cancer, rs12537, phenotype, estrogen receptor, *MTMR3*

INTRODUCTION

The burden of breast cancer is increasing worldwide. Among the 19.3 million new cases reported by the GLOBOCAN 2020, breast cancer patients account for 11.7% (1). China is undergoing cancer transition with an increasing burden of breast cancer, and the incidence of breast cancer arrives at 18.41%. In China, female breast cancer patients took up approximately 18% of breast cancer deaths across the world (2). Many sequencing methods such as genome-wide association studies (GWASs), exome and lncRNA sequencing are used to identify SNPs/loci/genes related to the occurrence, development, prognosis and drug resistance of breast cancer (3–7).

Breast cancer is a heterogeneous and polygenic disease, and breast cancer susceptibility SNPs and genes are closely related to molecular subtype and clinical phenotypes (8–10). However, for many SNPs, evidence of an association with cancer is often weak, and accurate estimates of the cancer risks associated with variants are often not available (11). In our previous study, we performed a genome-wide lncRNA association study, and reported a suggestive SNP, rs12537 ($p = 8.84\text{E-}07$), which may be associated with breast cancer susceptibility (12). rs12537 variant was reported to be associated with IgA nephropathy in Han Chinese, and rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) in Egyptian patients (13, 14). Moreover, rs12537 variant was also found associated with significantly increased gastric cancer risk (15, 16). However, the relationship between rs12537 and breast cancer remains unknown.

To investigate the association between rs12537 on 22q12.2 and breast cancer susceptibility as well as the clinical phenotype including familial history, menopausal status, estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER-2) and molecular subtypes of breast cancer patients in Chinese Han population, we conducted a genotype-phenotype analysis to clarify the association of rs12537 with breast cancer phenotypes in Chinese Han population. Moreover, we tried to analyze rs12537 associated genes and breast cancer based on public databases.

MATERIALS AND METHODS

Subjects

We collected the genotyping data from our previous data (including GWAS stage data and replication stage genotyping data) and clinical data (including age of onset, family history, menopausal status, ER, PR and HER-2) of a total of 5,634 patients (12). Immunohistochemical analysis was employed to evaluate the ER, PR and HER-2 status of breast tissue of biopsies. Each case was diagnosed and confirmed by at least two oncologists. And their clinical information was collected by investigators with a comprehensive clinical check-up. We also collected the genotyping data and age of 6,308 healthy controls, and they were clinically determined to be free of breast cancer, other neoplastic disease, systemic disorders, and to have no family history of cancer (including first-, second- and third-degree relatives). All participants provided written informed consent. This study was approved by the institutional ethics committee of each hospital and was conducted according to the Declaration of Helsinki principles.

Statistical Analysis

To identify which phenotypes were associated with the specific SNP rs12537, we performed case-control and case-only analysis to examine the risk conferred by the suggestive SNP on different phenotypes of breast cancer. PLINK1.07 software (developed by Christopher Chang and others) and SPSS16.0 (IBM, <https://www.ibm.com>) were used to perform chi-square test and logistic regression analysis to explore the correlation between rs12537 and breast cancer susceptibility, as well as the different phenotypes of breast cancer. Allele frequency and genotype frequency were calculated by direct counting method,

TABLE 1 | Baseline characteristics of breast cancer patients and healthy controls.

Characteristics	Total Samples	GWAS Samples	Replication Samples
Cases			
Sample size	5,634	1,496	4,138
Mean age at onset (SD)	50.7 ± 11.1	49.3 ± 10.9	51.2 ± 11.2
Mean age (SD)	51.0 ± 11.1	50.0 ± 10.9	51.3 ± 11.1
Familial history of cancer			
Familial (%)	265 (4.70%)	24 (1.60%)	241 (5.82%)
Sporadic (%)	5,369 (95.30%)	1472 (98.40%)	3,897 (94.18%)
Menopausal status	4,223	849	3,374
Premenopausal (%)	2,118 (50.15%)	471 (55.48%)	1,647 (48.81%)
Postmenopausal (%)	2,105 (49.85%)	378 (44.52%)	1,727 (51.19%)
ER	4,263	703	3,560
Positive (%)	2,773 (65.05%)	456 (64.86%)	2,317 (65.08%)
Negative (%)	1,490 (34.95%)	247 (35.14%)	1,243 (34.92%)
PR	4,262	702	3,560
Positive (%)	2,704 (63.44%)	413 (58.83%)	2,291 (64.35%)
Negative (%)	1,558 (36.56%)	289 (41.17%)	1,269 (35.65%)
HER-2	4,254	698	3,556
Positive (%)	1,148 (26.99%)	136 (19.48%)	1,012 (28.46%)
Negative (%)	3,106 (73.01%)	562 (80.52%)	2,544 (71.54%)
Molecular subtypes	4,096	687	3,409
Luminal A breast cancer	1,088 (26.56%)	207 (30.13%)	881 (25.84%)
Lumina B breast cancer	1,415 (34.55%)	244 (35.52%)	1,171 (34.35%)
HER-2 amplified breast cancer	630 (15.38%)	93 (13.54%)	537 (15.75%)
Basal-like breast cancer	963 (23.51%)	143 (20.82%)	820 (24.05%)
Controls			
Sample size	6,308	1,257	5,051
Mean age (SD)	47.4 ± 12.8	37.2 ± 11.9	50.3 ± 10.6

χ^2 significance test was carried out, and the relative risk was evaluated by Odds ratio (OR) and 95% confidence interval (95% CI), with the difference being statistically significant ($p < 0.05$), a remarkable deviation from Hardy-Weinberg equilibrium in the controls ($p > 0.05$) during each stage.

RESULTS

Sample Characteristics

All subjects involved in this study were from our early genome-wide lncRNA association study (12). The clinical features of 5,634 cases and 6,308 controls are intact in this study. The average age of onset of 5,634 female breast cancer patients was 50.7 ± 11.1 . 4.70% (265) of these patients had familial history of cancer, 50.15% (2,118) were diagnosed with premenopausal breast cancer, 65.05% (2,773) were ER positive, 63.44% (2,704) were PR positive, and 26.99% (1,148) were HER-2 positive. For molecular subtypes, data of 1,538 of these patients were missing, 26.56% (1,088), 34.55% (1,415) and 15.38% (630) were lumina A, lumina B and HER-2 amplified breast cancer carriers, respectively, and 23.51% (963) were diagnosed with basal-like

TABLE 2 | The genotypic and allelic frequency of rs12537.

Groups		Allele frequency (%)		p-value	OR (95% CI)	p-hwe
		T	C			
Family history	Positive vs. Control	0.175	0.205	0.1095	0.8226 (0.6475–1.045)	0.6475
	Negative vs. Control	0.1774	0.205	1.44E-06	0.8362 (0.7775–0.8994)	0.7775
	Positive vs. Negative	0.175	0.1774	0.8945	0.9837 (0.7719–1.254)	0.7719
Menopausal status	Positive vs. Control	0.1751	0.205	9.89E-05	0.8231 (0.7462–0.908)	0.7462
	Negative vs. Control	0.1843	0.205	0.008677	0.8764 (0.7941–0.9672)	0.7941
	Positive vs. Negative	0.1751	0.1843	0.3245	0.9393 (0.8292–1.064)	0.8292
ER	Positive vs. Control	0.1664	0.205	1.66E-08	0.7744 (0.7085–0.8464)	0.7085
	Negative vs. Control	0.1938	0.205	0.2052	0.9321 (0.8360–1.0390)	0.836
	Positive vs. Negative	0.1664	0.1938	0.004006	0.8309 (0.7323–0.9427)	0.7323
PR	Positive vs. Control	0.1678	0.205	6.13E-08	0.7822 (0.7155–0.8550)	0.7155
	Negative vs. Control	0.1911	0.205	0.114	0.9163 (0.8222–1.021)	0.8222
	Positive vs. Negative	0.1678	0.1911	0.01379	0.8536 (0.7525–0.9683)	0.7525
HER-2	Positive vs. Control	0.1832	0.205	0.02393	0.8699 (0.7708–0.9818)	0.7708
	Negative vs. Control	0.1728	0.205	1.14E-06	0.8101 (0.7441–0.8819)	0.7441
	Positive vs. Negative	0.1832	0.1728	0.2987	1.0740 (0.9388–1.2280)	0.9388

TABLE 3 | eQTL analysis to identify rs12537 associated genes.

p-value	Gene ID	Gene Symbol	Chr	Pos (hg19)	Z-score	FDR
4.91E-71	ENSG00000100325	ASCC2 ^a	22	30209434	−17.8204	0
1.08E-34	ENSG00000184117	NIPSNAP1 ^a	22	29964061	−12.2859	0
8.72E-30	ENSG00000100330	MTMR3 ^a	22	30352999	−11.3358	0
4.64E-23	ENSG00000100319	ZMAT5 ^a	22	30144972	9.8893	0
1.15E-08	ENSG00000100012	SEC14L3 ^a	22	30855991	−5.707	6.43E-05
6.63E-08	ENSG00000167065	DUSP18 ^a	22	31055957	5.401	0.000253
1.42E-06	ENSG00000099995	SF3A1 ^a	22	30740457	−4.8224	0.003981
9.10E-06	ENSG00000100296	THOC5 ^a	22	29926536	4.4376	0.023729
3.85E-08	ENSG00000189283	FHIT ^b	3	60486084	5.4974	0.000499
6.03E-08	ENSG00000126353	CCR7 ^b	17	38715872	5.418	0.000806
4.52E-07	ENSG00000120915	EPHX2 ^b	8	27375688	5.0458	0.004067
7.11E-07	ENSG00000138795	LEF1 ^b	4	109029406	4.9584	0.00614
1.28E-06	ENSG00000170915	PAQR8 ^b	6	52249397	4.8428	0.010193
8.11E-06	ENSG00000186854	TRABD2A ^b	2	85091453	4.4624	0.04903

^acis-eQTL effects genes.

^btrans-eQTL effects genes.

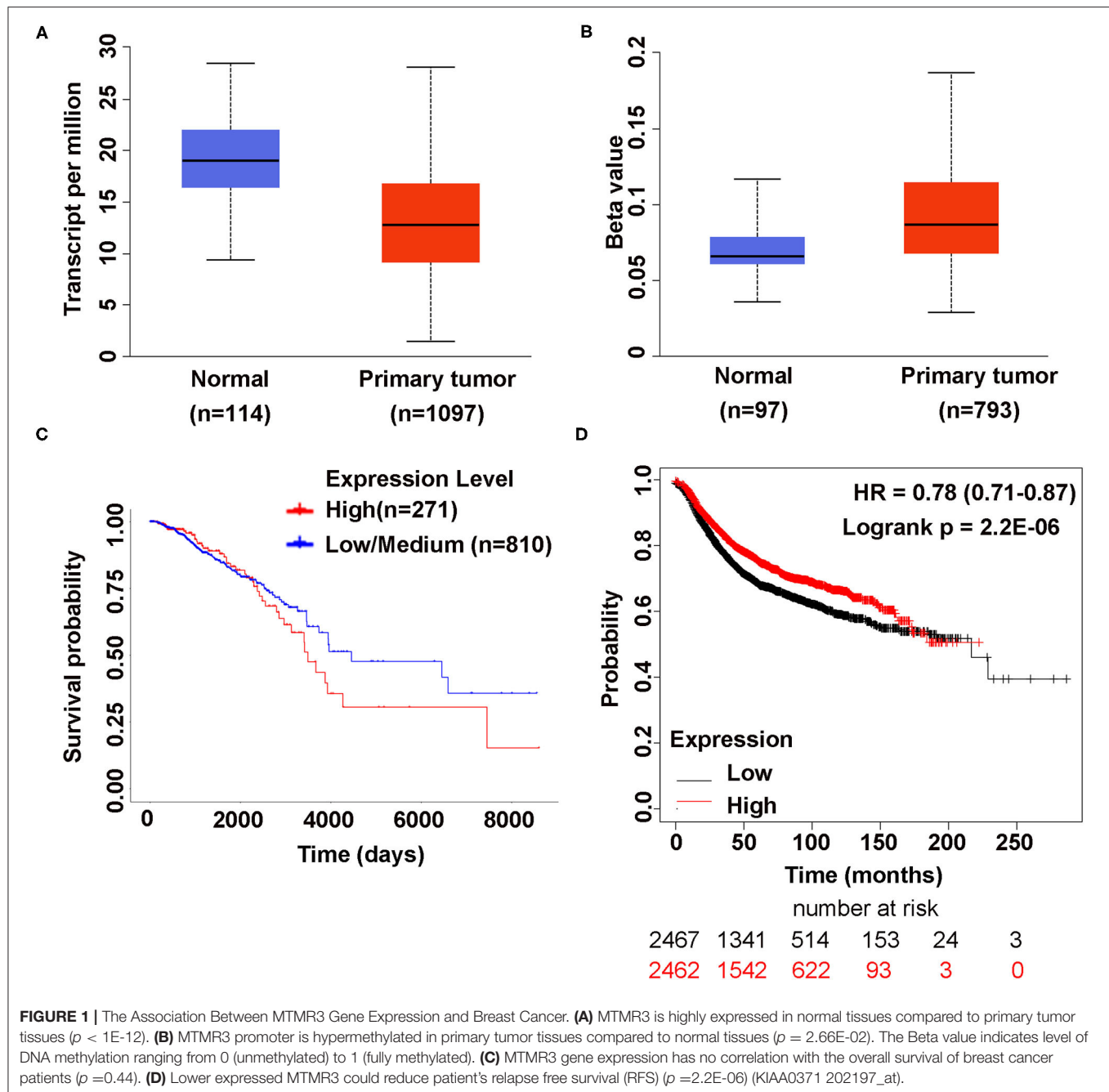
breast cancer. The average age of 6,308 female healthy controls was 47.4 ± 12.8 (Table 1).

Genotypic and Phenotype Analysis

To further explore the relationship between suggestive SNP rs12537 and breast cancer susceptibility, we combined our GWAS and replication data to perform a genotypic and phenotype analysis based on the clinical information we collected. The results show that the suggestive SNP rs12537 is not related to the familial history of cancer, menopausal status, HER-2 and the four molecular subtypes of breast cancer patients. And there is a statistical difference between PR positive patients and PR negative patients ($p = 0.01379$, OR = 0.8536, 95% CI: 0.7525–0.9638) in rs12537 variant (Table 2 and Supplementary Table 1). And we also performed genotypic and phenotype analysis on

the other three SNPs, rs9397435, rs11066150 and rs62112521 in Chinese Han women (12), but we found no correlation between these three SNPs and the clinical characteristics of breast cancer (Supplementary Table 2).

Surprisingly, ER positive patients and healthy controls also differ statistically ($p = 1.66E-8$, OR = 0.7744, 95% CI: 0.7085–0.8464) in rs12537 variant, and this difference has reached the level of the whole genome for $p < 5.00E-8$. Moreover, there is also a statistical difference between ER positive patients and ER negative patients ($p = 0.004006$, OR = 0.8309, 95% CI: 0.7323–0.9427) in rs12537 variant (Table 2). To exclude the influence of clinical features other than ER status on the results, we further analyzed and compared the clinical differences between ER positive, ER negative breast cancer patients and healthy controls (Supplementary Table 3), and we found that age, family history



and HER-2 expression were not correlated with ER. So rs12537 is a novel ER positive breast cancer associated SNP variant in Chinese Han women.

rs12537 Associated Gene MTMR3 and Breast Cancer

Expression quantitative trait locus (eQTL) has become a common tool to interpret the regulatory mechanisms of the variants associated with complex traits through genome-wide association studies (GWAS) (17, 18). To identify rs12537 associated genes, based on the eQTLGen database (<https://www.eqtlgen.org/>), we identified 8 cis-eQTL effects genes (ASCC2, NIPSNAP1, MTMR3, ZMAT5, SEC14L3, DUSP18, SF3A1 and THOC5) and 6 trans-eQTL effects genes (FHIT, CCR7, EPHX2, LEF1, PAQR8 and TRABD2A) (Table 3).

rs12537 associated gene, MTMR3, was reported to be associated with RA and SLE, gastric cancer and breast cancer (14, 15, 19). Therefore, we try to investigate the expression of MTMR3 in the cancer genome atlas (TCGA) database by UALCAN (20), discovering that compared to normal tissues MTMR3 was lower expressed in primary tumor tissues ($p < 1E-12$), but the promoter methylation level was higher ($p =$

2.66E-02). The Beta value indicates level of DNA methylation ranging from 0 (unmethylated) to 1 (fully methylated). MTMR3 gene expression has no correlation with the overall survival of breast cancer patients ($p = 0.44$). Lower expressed MTMR3 could reduce patient's relapse free survival (RFS) ($p = 2.2E-06$) (KIAA0371 202197_at).

2.66E-02). *MTMR3* expression was not associated with overall survival (OS) ($p = 0.44$) (Figures 1A–C). Moreover, based on Kaplan-Meier Plotter (www.kmplot.com) (21), we found highly expressed *MTMR3* could improve patients relapse free survival (RFS) ($p = 2.2E-06$) (Figure 1D), but there was no correlation between *MTMR3* expression and the OS, postoperative survival (PPS) as well as distant metastasis-free survival (DMFS) in breast cancer patients (Supplementary Figure 1).

DISCUSSION

Breast cancer is a complex multifactorial disease, with high incidence, strong invasiveness, metastasis and heterogeneity (1, 22, 23). A large number of sequencing studies have identified more than 200 susceptibility SNPs/genes (24). By combining sequencing analysis with the clinical characteristics of breast cancer patients, more SNPs/genes that have a stronger correlation with clinical characteristics were identified, which provides important theoretical support for precision treatment of breast cancer (8, 11, 25).

It is reported that more than 60% of breast cancers, including Luminal A and Luminal B breast cancers, were ER positive (8, 10). And ER positive breast cancer is a highly heterogeneous disease comprising different histological and mutational patterns, with varied clinical courses and responses to systemic treatment. GWASs have identified a lot of ER positive breast cancer associated SNPs, such as rs112545418, rs17132398 in 4p16, rs116638271, rs77274510 and rs117564384 in 11q13 and rs10941679 in 5p12 (26, 27). In our previous study, we designed a lncRNA array independently, and then performed the first genome-wide lncRNA association study on Han Chinese women, identifying a novel breast cancer-associated susceptibility SNP, rs11066150, a previously reported SNP, rs9397435 and two suggestive SNPs rs12537 and rs62112521 (12), but our study revealed that rs11066150, rs9397435 and rs6211252 had no relationship with the clinical characteristics of breast cancer (Supplementary Table 2). In the present research, we identified rs12537 as a novel susceptibility SNP in ER positive breast cancer in Han Chinese women. And this is the first time that rs12537 has been reported to be associated with ER positive breast cancer. However, only 4,263 ER patients (65.05% ER positive, 34.56% ER negative) and 6,308 healthy controls were included in this study, and a larger and better-matched population (including age, familial history, menopausal status, ER, PR and HER-2) may be needed for further verification.

The SNP rs12537 present in the miR-181a-binding site in the 3' UTR of the *MTMR3* gene (15) and T/C variant in *MTMR3* were reported to be associated with IgA nephropathy, RA, SLE and gastric cancer (13–16). As an autophagy-related gene involved in the negative regulation of autophagy initiation (24), rs12537 T/T carriers were associated with lower serum *MTMR3* expression and higher miR-181a expression than in other genotypes among SLE patients, and their interaction may lead to autophagy increasing (14). rs12537 CT genotype carriers in gastric cancer had low *MTMR3* mRNA expression than CC genotype carriers (15). Ectopic expression of miR-181a mimics or introduction of *MTMR3* small interfering RNA resulted in an increase in cell

proliferation, colony formation, migration, invasion, as well as suppression of apoptosis in gastric cancer (28).

DNA methylation plays a crucial role in the formation and process of cancers and it could be potential candidate biomarkers for cancers (29). Based on TCGA database, we found that *MTMR3* gene was lower expressed in breast cancer tissues than normal tissues and the promoter methylation level was higher. However, *MTMR3* expression had no correlation with overall survival. Here, we hypothesize that rs12537 variant in ER positive breast cancer patients could regulate the methylation of *MTMR3*, and further studies are required to fully understand the mechanism.

In conclusion, the results of our study show that rs12537 is a novel susceptibility SNP in ER positive breast cancer in Chinese Han population. Moreover, rs12537 associated gene *MTMR3* is lowly expressed but highly methylated in breast cancer. Considering that we have not found the correlation between *MTMR3* expression and overall survival based on TCGA database, multicentric studies involving a larger number of cases and genotypic data are needed to verify this result.

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 authors.

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

JX, YC, and BZ conceived of the idea. JX, GL, and MC conducted statistical analyses. WL and YW collected clinical data. JX and GL wrote the manuscript with inputs from all authors. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Talent Funding of the First Affiliated Hospital of Anhui Medical University (1519) and Natural Science Foundation of Higher Education of Anhui Province of China (KJ2017A181).

ACKNOWLEDGMENTS

The authors would like to thank all the participating patients and healthy controls, as well as all the doctors and nurses who have contributed to this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.708644/full#supplementary-material>

REFERENCES

- Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660
- Cao W, Chen HD, Yu YW, Li N, Chen WQ. Changing profiles of cancer burden worldwide and in China: a secondary analysis of the global cancer statistics 2020. *Chin Med J.* (2021) 134:783–91. doi: 10.1097/CM9.0000000000001474
- Ulitsky I, Bartel DP. lincRNAs: genomics, evolution, and mechanisms. *Cell.* (2013) 154:26–46. doi: 10.1016/j.cell.2013.06.020
- Milne RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet.* (2017) 49:1767–78. doi: 10.1038/ng.3785
- Stewart RL, Updike KL, Factor RE, Henry NL, Boucher KM, Bernard PS, et al. A multigene assay determines risk of recurrence in patients with triple-negative breast cancer. *Cancer Res.* (2019) 79:3466–78. doi: 10.1158/0008-5472.CAN-18-3014
- Zhang B, Chen MY, Shen YJ, Zhuo XB, Gao P, Zhou FS, et al. A large-scale, exome-wide association study of han chinese women identifies three novel loci predisposing to breast cancer. *Cancer Res.* (2018) 78:3087–97. doi: 10.1158/0008-5472.CAN-17-1721
- Kim J, Piao HL, Kim BJ, Yao F, Han Z, Wang Y, et al. Long noncoding RNA MALAT1 suppresses breast cancer metastasis. *Nat Genet.* (2018) 50:1705–15. doi: 10.1038/s41588-018-0252-3
- Xu Y, Chen M, Liu C, Zhang X, Li W, Cheng H, et al. Association study confirmed three breast cancer-specific molecular subtype-associated susceptibility loci in chinese han women. *Oncologist.* (2017) 22:890–4. doi: 10.1634/theoncologist.2016-0423
- Curtit E, Pivot X, Henriques J, Paget-Bailly S, Fumoleau P, Rios M, et al. Assessment of the prognostic role of a 94-single nucleotide polymorphisms risk score in early breast cancer in the SIGNAL/PHARE prospective cohort: no correlation with clinico-pathological characteristics and outcomes. *Breast Cancer Res.* (2017) 19:98. doi: 10.1093/annonc/mdw364.34
- Early Breast Cancer Trialists' Collaborative G, Davies C, Godwin J, Gray R, Clarke M, Cutter D, et al. Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *Lancet.* (2011) 378:771–84. doi: 10.1016/S0140-6736(11)60993-8
- Dorling L, Carvalho S, Allen J, González-Neira A, Luccarini C, Wahlström C, et al. Breast cancer risk genes - association analysis in More than 113,000 women. *N Engl J Med.* (2021) 384:428–39. doi: 10.1056/NEJMoa1913948
- Xu JK, Li GZ, Li Z, Li WJ, Chen RS, Zhang B, et al. Genome-wide long non-coding RNA association study on Han Chinese women identifies lncHSAT164 as a novel susceptibility gene for breast cancer. *Chin Med J (Engl).* (2021) 134:1138–45. doi: 10.1097/CM9.0000000000001429
- Yu XQ, Li M, Zhang H, Low HQ, Wei X, Wang JQ, et al. A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nat Genet.* (2011) 44:178–82. doi: 10.1038/ng.1047
- Senousy MA, Helmy HS, Fathy N, Shaker OG, Ayeldeen GM. Association of MTMR3 rs12537 at miR-181a binding site with rheumatoid arthritis and systemic lupus erythematosus risk in Egyptian patients. *Sci Rep.* (2019) 9:12299. doi: 10.1038/s41598-019-48770-5
- Lin Y, Nie Y, Zhao J, Chen X, Ye M, Li Y, et al. Genetic polymorphism at miR-181a binding site contributes to gastric cancer susceptibility. *Carcinogenesis.* (2012) 33:2377–83. doi: 10.1093/carcin/bgs292
- Cipollini M, Landi S, Gemignani F. MicroRNA binding site polymorphisms as biomarkers in cancer management and research. *Pharmgenomics Pers Med.* (2014) 7:173–91. doi: 10.2147/PGPM.S61693
- Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* (2016) 48:481–7. doi: 10.1038/ng.3538
- Karlsson Linner R, Biroli P, Kong E, Meddens SFW, Wedow R, Fontana MA, et al. Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat Genet.* (2019) 51:245–57. doi: 10.1038/s41588-018-0309-3
- Wang B, Mao JH, Wang BY, Wang LX, Wen HY, Xu LJ, et al. Exosomal miR-1910-3p promotes proliferation, metastasis, and autophagy of breast cancer cells by targeting MTMR3 and activating the NF-kappaB signaling pathway. *Cancer Lett.* (2020) 489:87–99. doi: 10.1016/j.canlet.2020.05.038
- Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi B, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia.* (2017) 19:649–58. doi: 10.1016/j.neo.2017.05.002
- Gyorffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. *Breast Cancer Res Treat.* (2010) 123:725–31. doi: 10.1007/s10549-009-0674-9
- Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet.* (2012) 8:e1002532. doi: 10.1371/journal.pgen.1002532
- Lindstrom S, Thompson DJ, Paterson AD, Li J, Gierach GL, Scott C, et al. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. *Nat Commun.* (2014) 5:5303. doi: 10.1038/ncomms6303
- Kapoor PM, Lindstrom S, Behrens S, Wang X, Michailidou K, Bolla MK, et al. Assessment of interactions between 205 breast cancer susceptibility loci and 13 established risk factors in relation to breast cancer risk, in the breast cancer association consortium. *Int J Epidemiol.* (2020) 49:216–32. doi: 10.1093/ije/dydz193
- Lin A, Li C, Xing Z, Hu Q, Liang K, Han L, et al. The LINK-A lncRNA activates normoxic HIF1alpha signalling in triple-negative breast cancer. *Nat Cell Biol.* (2016) 18:213–24. doi: 10.1038/ncb3295
- Ruiz-Narvaez EA, Sucheston-Campbell L, Bensen JT, Yao S, Haddad S, Haiman CA, et al. Admixture mapping of African-American women in the AMBER consortium identifies new loci for breast cancer and estrogen-receptor subtypes. *Front Genet.* (2016) 7:170. doi: 10.3389/fgene.2016.00170
- Ghoussaini M, French JD, Michailidou K, Nord S, Beesley J, Canisus S, et al. Evidence that the 5p12 variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through FGF10 and MRPS30 regulation. *Am J Hum Genet.* (2016) 99:903–11. doi: 10.1016/j.ajhg.2016.07.017
- Lin Y, Zhao J, Wang H, Cao J, Nie Y. miR-181a modulates proliferation, migration and autophagy in AGS gastric cancer cells and downregulates MTMR3. *Mol Med Rep.* (2017) 15:2451–6. doi: 10.3892/mmr.2017.6289
- Men C, Chai H, Song X, Li Y, Du H, Ren Q. Identification of DNA methylation associated gene signatures in endometrial cancer via integrated analysis of DNA methylation and gene expression systematically. *J Gynecol Oncol.* (2017) 28:e83. doi: 10.3802/jgo.2017.28.e83

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

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

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



Whole-Genome Sequencing Reveals Exonic Variation of *ASIC5* Gene Results in Recurrent Pregnancy Loss

Nourah H. Al Qahtani¹, Sayed AbdulAzeez², Noor B. Almandil³, Norah Fahad Alhur², Hind Saleh Alsuwat², Hatoon Ahmed Al Taifi¹, Ahlam A. Al-Ghamdi¹, B. Rabindran Jermy⁴, Mohamed Abouelhoda^{5,6}, Shazia Subhani^{5,6}, Lubna Al Asoom⁷ and J. Francis Borgio^{2,8*}

¹ Department of Obstetrics and Gynaecology, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, ² Department of Genetic Research, Institute for Research and Medical Consultations, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, ³ Department of Clinical Pharmacy Research, Institute for Research and Medical Consultations, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, ⁴ Department of Nanomedicine Research, Institute for Research and Medical Consultations, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, ⁵ Saudi Human Genome Project, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia, ⁶ Department of Genetics, King Faisal Specialist Hospital and Research Center, Riyadh, Saudi Arabia, ⁷ Department of Physiology, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, ⁸ Department of Epidemic Diseases Research, Institute for Research and Medical Consultations, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

OPEN ACCESS

Edited by:

D. Thirumal Kumar,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Jayalakshmi Mariakuttikan,
Madurai Kamaraj University, India
Huda Elfakharany,
National Heart Institute, Egypt

*Correspondence:

J. Francis Borgio
fbalexander@iau.edu.sa;
borgiomicro@gmail.com
orcid.org/0000-0001-7199-1540

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 23 April 2021

Accepted: 21 June 2021

Published: 30 July 2021

Citation:

Al Qahtani NH, AbdulAzeez S,
Almandil NB, Fahad Alhur N,
Alsuwat HS, Al Taifi HA,
Al-Ghamdi AA, Rabindran Jermy B,
Abouelhoda M, Subhani S, Al
Asoom L and Borgio JF (2021)
Whole-Genome Sequencing Reveals
Exonic Variation of *ASIC5* Gene
Results in Recurrent Pregnancy Loss.
Front. Med. 8:699672.
doi: 10.3389/fmed.2021.699672

Family trio next-generation sequencing-based variant analysis was done to identify the genomic reason on unexplained recurrent pregnancy loss (RPL). A family (dead fetus and parents) from Saudi Arabia with an earlier history of three unexplained RPLs at the ninth week of pregnancy was included in the study. Whole-genome sequencing (WGS) of a dead fetus and the parents was done to identify the pathogenic variation and confirmed through Sanger sequencing. WGS of dead fetus identifies a novel homozygous exonic variation (NM_017419.3:c.680G>T) in *ASIC5* (acid-sensing ion channel subunit family member 5) gene; the parents are heterozygous. Newly designed ARMS PCR followed by direct sequencing confirms the presence of heterozygous in one subject and absence of homozygous novel mutation among randomly selected healthy Saudis. The second family with heterozygous was confirmed with three unexplained RPLs. Pathogenicity analysis of R227I amino acid substitution in *ASIC5* protein through molecular docking and interaction analysis revealed that the mutations are highly pathogenic, decrease the stability of the protein, and prevent binding of amiloride, which is an activator to open the acid-sensing ion channel of *ASIC5*. The identified rare and novel autosomal recessive mutation, c.680G>T:p.R227I (*ASIC5*^{Saudi}), in two families confirm the *ASIC5* gene association with RPL and can be fatal to the fetus.

Keywords: exome, recurrent pregnancy loss, whole genome sequencing, *ASIC5*, Saudi Arabia, molecular docking, next generation sequencing, unknown spontaneous abortion

INTRODUCTION

Recurrent pregnancy loss (RPL), or recurrent miscarriage (RM) is described as three or more sequential unpremeditated abortions before 20 weeks of gestation (1), a condition termed “habitual abortion” or “repeated spontaneous abortions” (2). RPL affects couples at propagative age around the world. The etiologies of RPL in Saudis or Arabs and other populations tend to be multifactorial. Factors including genetic abnormalities (3–10), placental anomalies (11–13), psychological trauma

and stressful life events (14), and certain coagulation and immunoregulatory protein defects (15–18) were reported to be associated with RPL among women in the Gulf region. In some populations, other factors have been studied, such as anatomical, endocrine, hormonal problems, infection, smoking and alcohol consumption, and exposure to environmental factors, and these factors could increase the risk of RPL (19). Several studies have reported the relationship between various causes of recurrent miscarriage among Saudis and the rest of the population; however, 30–50% of RPLs were unexplained (5, 19). More studies on RPL only can reveal the cause. The objective of the study is to analyze the genetic basis of a family from Saudi Arabia with an earlier history of recurrent pregnancy loss at the ninth week of pregnancy using next-generation sequencing [whole-genome sequencing (WGS)] by complete analysis of whole genome of the fetus and parents followed by rigorous bioinformatics and confirmatory analyses (20–36). The study reports a novel homozygous exonic variation in the *ASIC5* gene in a dead fetus, while the parents are heterozygous.

MATERIALS AND METHODS

Ethics and Study Subjects

The study was approved by the Institutional Review Boards Committee of the Imam Abdulrahman Bin Faisal University (IRB-2017-13-137).

A family with a past history of three miscarriages has been included in the study with a written consent from the father and mother. During the fourth pregnancy, the mother experienced a similar type of miscarriage at the ninth week of pregnancy. Tissue (separated cautiously from maternal tissue to avoid contamination) samples and blood samples were collected from the fetus (proband) and parents, respectively. Miscarriage sample was collected in an RNAlater Cell Reagent (Qiagen, Hilden, Germany). The DNAs of the samples were isolated, and the most prevalent genetic disease, hemoglobinopathies, were screened using the Sanger sequencing. Genes (functional variants and deletions in *HBB*, *HBA1*, *HBA2*, *ATRX*, and *HBD*) related to the most prevalent mutations have been found to be normal. Hence, the WGS was done for the miscarriage tissue, mother, and father genomes.

Whole-Genome Sequencing and Trio Analysis

The trio analysis has been carried out using the best practice GATK pipeline (20). The program Fastx (http://hannonlab.cshl.edu/fastx_toolkit) was used to filter low-quality reads. Then the reads were aligned to the reference human genome (hg19) using the program BWA (21). The GATK haplotype caller was used to call the variants. The resulting variants were then annotated using in-house developed workflow including the following three sets of data sources:

1. *Public databases*: These were collected from the Annovar packages, and they include the basic positional information about genes and related proteins. They also include

information from the dbSNP database, the 1000 Genome database, ExAC, and gnomAD databases. Annovar also includes predictions of the functional effect of the variants from the tools PolyPhan, Sift, CADD, and MetaSVM. In addition to Annovar, we used the ClinVar and OMIM databases to annotate the variants and genes with up-to-date medical information.

2. *In-house databases*: We annotated the variants using the Saudi Human Genome Program variant DB to check for variant frequency in the Saudi population (22–24).
3. *Commercial databases*: We used the HGMD database to annotate the variant with clinical information.

After variant annotation, we ran filters according to the ACMG (American College of Medical Genetics and Genomics) guidelines. We excluded variants that are intergenic, synonymous, appearing more than 5% in population databases, or not damaging (as predicted by CADD, PolyPhan, SIFT, and MetaSVM). We also ran extra trio analysis to filter the variants according to the autosomal recessive, *de novo*, compound heterozygous, and x-linked. After applying these filters, the remaining variants were examined manually to match the annotated clinical information to the fetus phenotype.

Sanger Sequencing Validation

Whole-genome result was confirmed using Sanger sequencing. The presence of the homozygous NM_017419.3:c.680G>T in the proband and heterozygous in the parents were confirmed using Sanger sequencing. Highly specific primers (*ASIC5F*: 5'-CAGATAAAAAACATGTTTCCATACATCTTCAG-3' and *ASIC5R*: 5'-TTGTGGCATGAACATTCCCTGGA-3') were designed, and the selected region of the gene was amplified [PCR recipe: MOLEQULE-ON absolute master mix 12.5 µl, *ASIC5F* 1 µl (10 nM), *ASIC5R* 1 µl (10 nM), DNA Template 25 ng, and Dis H₂O to 25 µl; temperature profile: 95°C for 10 min; 35 cycles of 95°C/60 s, 60°C/60 s, 72°C/60 s; and 72°C for 5 min] and sequenced using BigDye Terminator Cycle Sequencing Kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA). Amplified PCR product (691 bp) of the *ASIC5* gene region was purified and sequenced using Genetic Analyzer 3500 (Thermo Fisher Scientific, Inc.) at the Department of Genetic Research, Institute for Research and Medical Consultations, Imam Abdulrahman Bin Faisal University (Dammam, Saudi Arabia). Sequences were analyzed using mutation surveyor software (Softgenetics, US) and DNA sequencing analysis software v.5.3 (Applied Biosystem; Thermo Fisher Scientific, Inc.).

Amplification Refractory Mutation System-Polymerase Chain Reaction-Based Variation Screening and Sanger Sequencing Validation

The amplification refractory mutation system-polymerase chain reaction (ARMS-PCR) was designed (primers will be available on request) to screen the presence of NM_017419.3:c.680G>T among healthy Saudis ($n = 200$). The subjects positive for the presence of

NM_017419.3:c.680G>T was confirmed through Sanger sequencing using primers (ASIC5F and ASIC5R). This is also to confirm the absence of the homozygous NM_017419.3:c.680G>T in the healthy Saudi subjects randomly selected.

Homology Protein Modeling and Functional Annotations

The homology modeling of wild (p.R227) and mutant (p.R227I) ASIC5 protein was performed using Swiss Model server (25), validated using PROCHECK (26). The structural functional annotations were completed using SAS-sequence server (27), ProFunc (28), and PDBsum (29). Mutant structures were generated using Swiss-PDB Viewer and PyMol (30). Energy minimization for the wild and mutants was estimated using GROMACS (31). Evolutionary conservation and functional aspect analysis of the R227 residue in the wild-type protein was performed using the ConSurf (32). PROVEAN and I-Mutant were used for analyzing the impact on the biological function of a protein due to an amino acid substitution R227I (33, 34). AutoDock Vina was used for molecular docking of the ligand with wild type and mutant ASIC5 protein (35), and the molecular visualization was done in PyMol and LigPlot (36).

RESULTS

Whole-Genome Sequencing and Trio Analysis

The family with a history of three unexplained miscarriages was included in the study. The couple is consanguineous but not first-degree relatives. There was no history of genetic and chronic diseases in the couple. The family was identified with a similar type of unknown spontaneous abortion at the ninth week of pregnancy. The mother was 30 years at the time of the fourth unexplained spontaneous miscarriage; the father was 34. The previous three unexplained miscarriages and the fourth were also of similar gestation. At this gestation, the gender of the proband cannot be determined even after miscarriage. The mother is devoid of uterine or cervical abnormalities. In order to identify the cause of the recurrent spontaneous abortion, WGS was done for the mother, father, and proband. The WGS of the trio (proband and parents) samples has revealed an inheritance of NM_017419.3:c.680G>T mutation in the ASIC5 gene from the parents (**Figure 1A** and **Supplementary Table 1**). Various heterozygous mutations observed in the proband are listed in the **Supplementary Material**, which were inherited either from the mother or father (**Supplementary Table 2**). The WGS result of NM_017419.3:c.680G>T variation in exon 4 of the ASIC5 gene has been confirmed through the Sanger sequencing (**Figure 1B**). The father and the mother were found to be carriers (heterozygous) of the c.680G>T:p.R227I at the ASIC5 gene, while the proband was homozygous to c.680G>T:p.R227I (GenBank: MN251164; ClinVar: SCV000930628; SNP ID: rs1248841709)

(**Figure 1**). The name of the novel variant was validated using Mutalyzer 2.0.32.

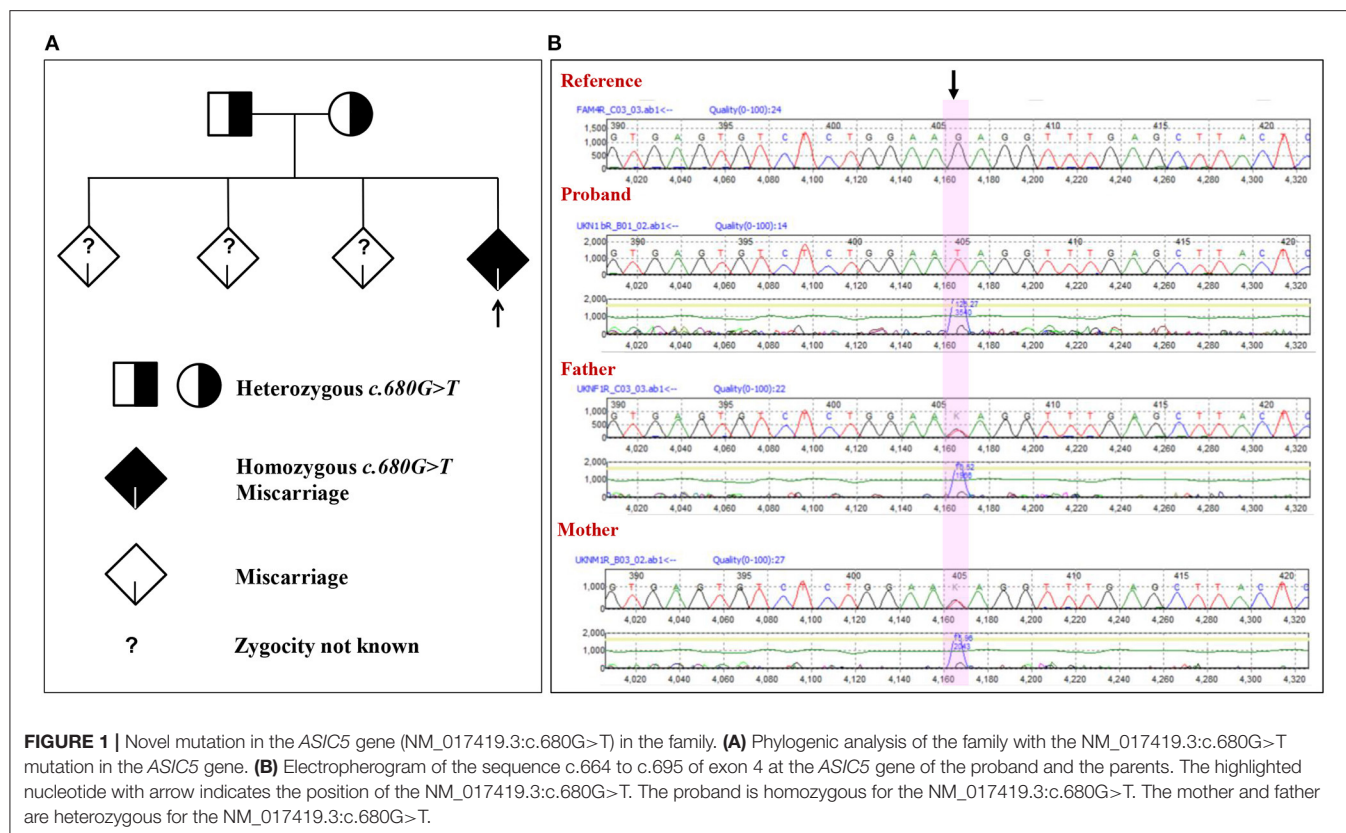
Amplification Refractory Mutation System-Polymerase Chain Reaction-Based Variation Screening and Sanger Sequencing Validation

In order to confirm the absence of the homozygous NM_017419.3:c.680G>T among the living population, a total of 200 healthy Saudis were selected randomly and checked for the mutation at the c.680 position in the ASIC5 gene using ARMS-PCR followed by Sanger sequencing. The results of the ARMS-PCR and direct sequencing of 200 healthy Saudis in the c.680 position in the ASIC5 gene revealed the absence of homozygous NM_017419.3:c.680G>T. Furthermore, this mutation is novel to the SHGP (Saudi Human Genome Program) database (about 9,500 cases). This suggests that the discovered mutation NM_017419.3:c.680G>T is rare, and their absence of a homozygous state in the healthy Saudis is validated. Furthermore, a female subject was observed with a heterozygous NM_017419.3:c.680G>T in the ASIC5 gene. The female subject with heterozygous mutations is a single daughter, and her mother experienced the unexplained RPL similar with the earlier family in the ninth week of pregnancy consecutively three times.

Molecular Docking and Interaction Analysis

The predicted structure of the wild ASIC5 on the Ramachandran plot showed ϕ/ψ angles of 83.1% residues in the most favored regions, 15.4% in the additional allowed regions, 1.1% in the generously allowed regions, and 0.3% in the disallowed regions (**Figure 2B**). The total residue span of the secondary structure consist of 23.0% residues involved in the formation of the strands, 23.0% residues in alpha helices, 2.6% residues in 3–10 helices, and 51.5% residues in other structural moieties. Analysis of secondary structure in ProFunc showed the presence of 3 β -sheets, 4 β -hairpins, 1 psi loop, 3 β -bulges, 14 strands, 14 helices, 5 helix–helix interactions, 34 β -turns, and 9 γ -turns. Homology modeling of the mutant structure (R227I) of the ASIC5 showed deviations from the wild type; the mutant structure on the Ramachandran plot showed ϕ/ψ angles of 84.3% residues in the most favored regions, 14.3% in the additional allowed regions, 1.1% in the generously allowed regions, and 0.3% in the disallowed regions. The total residue span of the secondary mutant structure consisting of 22.7% residues involving the formation of the strands, 23.7% residues in alpha helices, 1.8% residues in 3–10 helices, and 51.8% residues in other structural moieties. Analysis of the secondary structure of the mutant in ProFunc showed the presence of 3 β -sheets, 4 β -hairpins, 1 psi loop, 2 β -bulges, 14 strands, 13 helices, 5 helix–helix interactions, 42 β -turns, and 8 γ -turns.

ConSurf analysis revealed that R227I is a functional residue, which is highly conserved and exposed. A total of 97 HMMER hits were considered for this analysis, while 91 of them were unique, including the query. PROVEAN analysis showed that



R227I is a deleterious amino acid substitution as evident from PROVEAN score -3.830 . I-Mutant analysis predicted that the free energy change value (DDG) between wild type and mutant type was less than zero ($DDG < 0$), which declares the decrease in protein stability. Wild (RMSD = 0.045 \AA) and mutant (RMSD = 0.088 \AA) proteins were superimposed, quantitative measure of similarity analysis revealed an increase of 95.56% root-mean-square deviation of atomic positions in the mutant (**Figure 2C**).

Molecular docking studies of wild (p.R227) and mutant (p.R227I) ASIC5 protein with amiloride, a potent inhibitor of acid-sensing ion channel proteins, were performed, and it was observed that the binding behavior of amiloride with the mutant model compared with the wild-type model was completely different (**Figure 2**). R227 residue is not directly involved in binding with the ligand, but it assists atomic interactions through binding of the ligand with protein molecules at specific sites (**Figure 2F**). In particular, a halogen bonding occurs between the chlorine atom (colored green) of amiloride with the amino group (NH_2) of Gln305 (colored blue). The oxygen atom of the carbonyl group (colored red) of amiloride interacts with the hydrogen of the amino group (NH_2) of Gln265 through $\text{N-H} \cdots \text{O}$ hydrogen bonding. In a similar fashion, the hydrogen of amiloride interacts with the oxygen group of Glu203. However, the R227I prevents the binding of ligand with the ASIC5 molecule at a specific site (**Figure 2G**). In this mutant model, an alteration in the protein coordination site occurs (Gly126 and Asn243) and, therefore, fails to coordinate with amiloride functional groups.

DISCUSSION

Studies on tissues of miscarriage specimens from women with RPL observed the chromosomal aberrations from 29 to 46% of miscarriage tissues, while majority of the RPL may be due to alternative mechanisms or other than chromosomal aberrations (37–39). The present observation suggests that coding variants in *ASIC5* gene can be one among the alternative mechanisms for RPL. The role of the acid-sensing ion channel subunit family member 5 (*ASIC5*) or *ACCN5* or bile acid-sensitive ion channel (*BASIC*) gene in humans, in general, and the development of the fetus, in particular, is scanty (40–42). Very limited studies are available on the gene *ASIC5* and related expression. This gene, *ASIC5*, was reported to be expressed in the amniotic fluid (43), fetal gut, brain, liver, heart, ovary, and testis (44). *ASIC5* is overexpressed in the fetal gut (41.0) and plasma (27.5). *ASIC5* was observed to a key player in the physiology of unipolar brush cells of the vestibulocerebellum (42, 45, 46). The complete functions of the *ASIC5* gene and its product are yet to be identified (40–42). Animal studies on the autosomal recessive mouse mutant of the gene encoding the L-type calcium channel revealed that the homozygous mutant animals die at birth; however, the heterozygous for the mutant is not distinguishable from that of wild animals (47). The study resembles the present observation of the heterozygous mutant of the healthy parents, while death of the fetus with a homozygous mutant in the gene belongs to the amiloride-sensitive Na^+ channel. The R227I

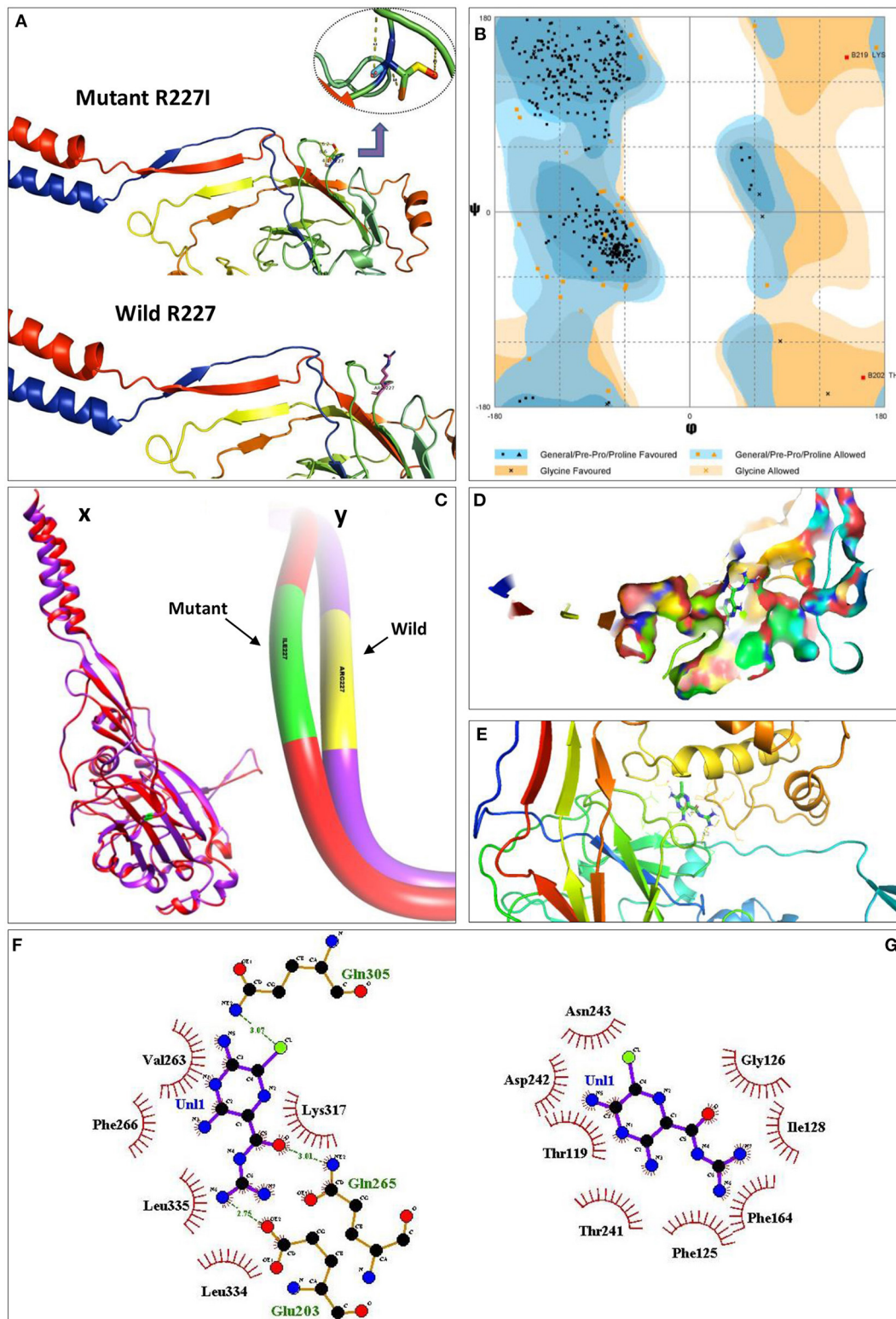


FIGURE 2 | Pathogenicity analysis of R227I mutation in the ASIC5 protein through molecular docking and interaction analysis. **(A)** Structural models of the wild (R227) and mutated (R227I) ASIC5 proteins. **(B)** Ramachandran plot for the predicted structure of the ASIC5 protein. Eighty-three portorage residues of the ASIC5 protein are in the most favored regions. Cx, superimposed structures of the wild (R227) and mutated (R227I) ASIC5 proteins; Cy, deviated region of R227I from R227 on superimposed wild and mutant ASIC5. **(D)** Amiloride with ASIC5 at the active binding site. **(E)** 3D amiloride with surrounding amino acids of ASIC5 protein. **(F,G)** Protein–ligand interaction. **(F)** Wild ASIC5 (R227) protein with ligand, amiloride. **(G)** Mutant ASIC5 (R227I) protein with ligand, amiloride.

prevents the binding of amiloride with ASIC5 protein. However, more confirmatory studies are mandatory to prove the failure in amiloride-R227I (ASIC5) binding in wet lab, which is mandatory for an activator to open its own channel (41, 48). Acid-sensing ion channel subunit channels play an important role in the fetal developmental pathology due to acidosis; furthermore, prolonged acidosis is significantly associated with mortality of the fetus (49, 50). Increased apoptosis was observed in the retina due to the mutant *ASIC2* gene compared with the wild type (51). Mammalian degenerin (*MDEG*) or *ASIC2* (acid-sensing ion channel subunit 2) gene mutant study on the development of *Xenopus* reported that the *Xenopus* oocytes with *ASIC2* mutation start to mature and die (52). This indicates the pathophysiology of the mutation in the acid-sensing ion channel subunit genes.

Earlier reports reported that in 39% of the Saudi females who had RPL, the origin of the patient in the study was unexplained or had no identifiable cause (5). Various reasons including genetic factors were stated for recurrent pregnancy loss among Saudi women (3, 4, 6, 14, 18). Consanguineous marriages are also considerably ($p = 0.046$) impacting (3). Genome-wide association study (GAWS) revealed the association of *ASIC5* ($p = 0.0029$; **Supplementary Table 3**) and level of manganese (53, 54). Furthermore, the level of manganese in the placental tissue of Saudi women with recurrent pregnancy loss was significantly ($p < 0.0001$) decreased (11). This suggests that the identified mutation in *ASIC5* might have played a role in the level of manganese in the present women. A recent study on the prognosis markers of glioblastoma revealed the expression of *ASIC5* as associated prognosis markers (55). *ASIC5* was found to be activated in the ethanol-(100 mM)-exposed neonatal rat cardiomyocytes along with other six molecules (*CYP2A6*, *PRL*, *CHRNA4*, *CNR1*, *CRH*, and *SLC40A1*) (56). Low (in 50%) *ASIC5* protein expression in melanoma were observed with <4% mutation rates (57).

Preparing the mutated animal model to study the impact of the mutant on the fetal development is not available in our laboratory, which is a limitation of the study. Hence, the region with the mutation, c.680G>T in the *ASIC5* gene, was screened using ARMS-PCR followed by sequencing using designed primers to identify the presence of c.680G>T in randomly selected Saudis in the study region, which confirms the absence of the homozygous NM_017419.3:c.680G>T and the presence of heterozygous NM_017419.3:c.680G>T in a female subject and her mother with RPL. The study confirms the influence of the association of the novel exonic mutation with RPL. However, nationwide studies are mandatory to identify the prevalence of this rare mutation and mutations in this gene among unexplained miscarriages cases and their impact on the recurrent pregnancy loss and fetal development. This can reveal the role of *ASIC5*. The protein–protein interaction analysis of *ASIC5* protein, with the protein observed with the mutation in the proband using STRING, revealed lack of interaction (**Supplementary Table 2**). However, the analysis using STRING cannot reveal any specific impact of mutated protein–protein interactions due to specific amino acid changes (58).

Based on the earlier reports on the member of the DEG/ENaC (degenerin/epithelial sodium channel) protein family and the

current observations, it may be concluded that the R227I amino acid substitution in the *ASIC5* is highly deleterious; the mutant *ASIC5* showed decreased stability and of the protein and prevents the binding of amiloride, a potent inhibitor of acid-sensing ion channel proteins (59).

The observed novel *ASIC5* gene-coding variant (*ASIC5*^{Saudi}) in two families confirm the *ASIC5* association with the results of RPL. Hence, this mutation is pathogenic, which may cause serious illness to the fetus and cause fetal mortality. The molecular mechanism behind the death of the fetus in relation to the homozygous NM_017419.3:c.680G>T at exon 4 (*ASIC5*^{Saudi}) in the *ASIC5* gene should be studied in detail. Early prenatal diagnosis of pathogenic variation like *ASIC5*^{Saudi} can provide a choice for the parent to decide pregnancy termination within the allowed time among high-consanguinity population (60).

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**.

ETHICS STATEMENT

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Institutional Review Board (IRB) at Imam Abdulrahman Bin Faisal University. IRB approval number: IRB-2017-13-137 dated 07 June 2017 and extended on 14 Dec 2020. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NHA, SA, NBA, HAA, AA-G, BR, and JB conceived and designed the research and analyzed the experiments. SA, NBA, NF, HSA, BR, and JB performed and analyzed the experiments. SA, MA, SS, and JB performed the whole-genome analysis. NHA, HAA, LA, and AA-G performed the clinical analysis. SA, HSA, and JB wrote the paper with the contributions of NHA, NBA, NF, HAA, AA-G, BR, MA, LA, and SS. All authors reviewed and approved the final manuscript.

FUNDING

This study was partially supported by The Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University (to JB, Grant No: 2017-100-IRMC). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors thank the Dean, Institute for Research and Medical Consultations (IRMC), Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, for her continuous support

and encouragement. We thank the Saudi Human Genome Program, King Abdulaziz City for Science and Technology team, for their support. The authors thank Mr. Ranilo M. Tumbaga, Mr. Horace T. Pacifico, and Mrs. Jee Entusiasamo Aquino for their assistance.

REFERENCES

- Baek KH, Lee EJ, Kim YS. Recurrent pregnancy loss: the key potential mechanisms. *Trends Mol Med.* (2007) 13:310–7. doi: 10.1016/j.molmed.2007.05.005
- Turki RF, Banni HA, Assidi M, Al-Qahtani MH, Abduljabbar HS, Jamel HS, et al. Analysis of chromosomal and genetic disorders in patients with recurrent miscarriages in Saudi Arabia. *BMC Genom.* (2014) 15:P73. doi: 10.1186/1471-2164-15-S2-P73
- McNamee K, Dawood F, Farquharson R. Recurrent miscarriage and thrombophilia: an update. *Curr Opin Obstetr Gynecol.* (2012) 24:229–34. doi: 10.1097/GCO.0b013e32835585dc
- Turki RF, Assidi M, Banni HA, Zahed HA, Karim S, Schulten HJ, et al. Associations of recurrent miscarriages with chromosomal abnormalities, thrombophilia allelic polymorphisms and/or consanguinity in Saudi Arabia. *BMC Med Gen.* (2016) 17:69. doi: 10.1186/s12881-016-0331-1
- Al-Ghamdi AA, Makhshen SF. Etiology of recurrent pregnancy loss in Saudi females. *Saudi J Med Med Sci.* (2016) 4:187. doi: 10.4103/1658-631X.188258
- Awartani KA, Al Shabibi MS. Description of cytogenetic abnormalities and the pregnancy outcomes of couples with recurrent pregnancy loss in a tertiary-care center in Saudi Arabia. *Saudi Med J.* (2018) 39:239. doi: 10.15537/smj.2018.3.21592
- Tan Y, Yin X, Zhang S, Jiang H, Tan K, Li J. Clinical outcome of preimplantation genetic diagnosis and screening using next generation sequencing. *Gigascience.* (2014) 3:30. doi: 10.1186/2047-217X-3-30
- Gaboon NE, Mohamed AR, Elsayed SM, Zaki OK, Elsayed MA. Structural chromosomal abnormalities in couples with recurrent abortion in Egypt. *Turkish J Med Sci.* (2015) 45:208–13. doi: 10.3906/sag-1310-5
- Alkhuriji AF, Alhimaidi AR, Babay ZA, Wary AS. The relationship between cytokine gene polymorphism and unexplained recurrent spontaneous abortion in Saudi females. *Saudi Med J.* (2013) 34:484–9.
- Al-Hassan S, Hellani A, Al-Shahrani A, Al-Deery M, Jaroudi K, Coskun S. Sperm chromosomal abnormalities in patients with unexplained recurrent abortions. *Arch Androl.* (2005) 51:69–76. doi: 10.1080/014850190518062
- Ghneim HK and Alshehly MM. Biochemical markers of oxidative stress in Saudi women with recurrent miscarriage. *J Korean Med Sci.* (2016) 31:98–105. doi: 10.3346/jkms.2016.31.1.98
- Ghneim HK, Al-Sheikh YA, Alshehly MM, Aboul-Soud MA. Superoxide dismutase activity and gene expression levels in Saudi women with recurrent miscarriage. *Mol Med Rep.* (2016) 13:2606–12. doi: 10.3892/mmr.2016.4807
- Atia TA. Placental apoptosis in recurrent miscarriage. *Kaohsiung J Med Sci.* (2017) 33:449–52. doi: 10.1016/j.kjms.2017.06.012
- Rouzi AA, Alamoudi R, Turkistani J, Almansouri N, Alkafy S, Alsenani N, et al. Miscarriage knowledge among Saudi women. *Fertility Sterility.* (2017) 108:e383. doi: 10.1016/j.fertnstert.2017.07.1111
- Gader AG, Al-Mishari AA, Al-Jabbari AW, Awadalla SA, Al-Momen AM. Thrombophilia in Saudi women with recurrent fetal loss. *J Appl Hematol.* (2010) 1:97.
- Al Omar SY, Mansour L, Alkhuriji AF, Alwasel S, Al-Qahtani S. Genetic association between the HLA-G 14-bp insertion/deletion polymorphism and the recurrent spontaneous abortions in Saudi Arabian women. *Genet Mol Res.* (2015) 14:286–93. doi: 10.4238/2015.January.23.2
- Gowri V, Udayakumar AM, Bisio W, Al Farsi Y, Rao K. Recurrent early pregnancy loss and consanguinity in Omani couples. *Acta obstetrica et gynecologica Scandinavica.* (2011) 90:1167–9. doi: 10.1111/j.1600-0412.2011.01200.x
- Almasry SM, Elmansy RA, Elfayomy AK, Algaidi SA. Ultrastructure alteration of decidual natural killer cells in women with unexplained recurrent miscarriage: a possible association with impaired decidual vascular remodelling. *J Mol Histol.* (2015) 46:67–78. doi: 10.1007/s10735-014-9598-8
- Ford HB, Schust DJ. Recurrent pregnancy loss: etiology, diagnosis, and therapy. *Rev Obstetr Gynecol.* (2009) 2:76.
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* (2010) 20:1297–303. doi: 10.1101/gr.107524.110
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* (2010) 26:589–95. doi: 10.1093/bioinformatics/btp698
- Abouelhoda M, Faquih T, El-Kalioby M, Alkuraya FS. Revisiting the morbid genome of Mendelian disorders. *Genome Biol.* (2016) 17:235. doi: 10.1186/s13059-016-1102-1
- Abouelhoda M, Sobahy T, El-Kalioby M, Patel N, Shamseldin H, Monies D, et al. Clinical genomics can facilitate countrywide estimation of autosomal recessive disease burden. *Genet Med.* (2016) 18:1244. doi: 10.1038/gim.2016.37
- The Saudi Mendliome Group. Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biol.* (2015) 16:134.
- Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* (2003) 31:3381–5. doi: 10.1093/nar/gkg520
- Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. *Proteins Struct Funct Bioinform.* (1992) 12:345–64. doi: 10.1002/prot.340120407
- Milburn D, Laskowski RA, Thornton JM. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.* (1998) 11:855–9. doi: 10.1093/protein/11.10.855
- Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res.* (2005) 33:W89–93. doi: 10.1093/nar/gki414
- Laskowski RA. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.* (2001) 29:221–2. doi: 10.1093/nar/29.1.221
- Guex N, Peitsch MC. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling. *Electrophoresis.* (1997) 18:2714–23. doi: 10.1002/elps.1150181505
- Lindahl E, Azuara C, Koehl P, Delarue M. NOMAD-Ref: visualization, deformation and refinement of macromolecular structures based on all-atom normal mode analysis. *Nucleic Acids Res.* (2006) 34:52–6. doi: 10.1093/nar/gkl082
- Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, et al. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* (2005) 33:W299–302. doi: 10.1093/nar/gki370
- Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* (2005) 33:W306–10. doi: 10.1093/nar/gki375
- Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics.* (2015) 31:2745–7. doi: 10.1093/bioinformatics/btv195
- Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem.* (2010) 31:455–61. doi: 10.1002/jcc.21334
- Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng Design Select.* (1995) 8:127–34. doi: 10.1093/protein/8.2.127
- Carp H, Toder V, Aviram A, Daniely M, Mashiach S, Barkai G. Karyotype of the abortus in recurrent miscarriage.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.699672/full#supplementary-material>

- Fertility Sterility*. (2001) 75:678–82. doi: 10.1016/S0015-0282(00)01801-X
38. Stephenson MD, Awartani KA, Robinson WP. Cytogenetic analysis of miscarriages from couples with recurrent miscarriage: a case-control study. *Hum Reprod*. (2002) 17:446–51. doi: 10.1093/humrep/17.2.446
 39. Li TC, Makris M, Tomsu M, Tuckerman E, Laird S. Recurrent miscarriage: aetiology, management and prognosis. *Hum Reprod Update*. (2002) 8:463–81. doi: 10.1093/humupd/8.5.463
 40. Xu S, Liu C, Ma Y, Ji HL, Li X. Potential roles of amiloride-sensitive sodium channels in cancer development. *BioMed Res Int*. (2016) 2016:2190216. doi: 10.1155/2016/2190216
 41. Hanukoglu I. ASIC and ENaC type sodium channels: conformational states and the structures of the ion selectivity filters. *FEBS J*. (2017) 284:525–45. doi: 10.1111/febs.13840
 42. Boiko NY, Kreko-Pierce T, Pugh J, Stockand JD. Ataxia in the ASIC5 knockout mouse associated with decreased excitability of vestibulocerebellum unipolar brush cells. *FASEB J*. (2019) 33:824–13. doi: 10.1096/fasebj.2019.33.1_supplement.824.13
 43. McCall MN, Uppal K, Jaffee HA, Zilliox MJ, Irizarry RA. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Res*. (2010) 39:D1011–5. doi: 10.1093/nar/gkq1259
 44. Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, et al. Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*. (2003) 31:142–6. doi: 10.1093/nar/gkg050
 45. Boiko N, Kucher V, Wang B, Stockand JD. Restrictive expression of acid-sensing ion channel 5 (asic5) in unipolar brush cells of the vestibulocerebellum. *PLoS ONE*. (2014) 9:e91326. doi: 10.1371/journal.pone.0091326
 46. Boiko N, Kucher V, Stockand JD. Dysfunction of acid-sensing ion channel 5 (ASIC5) in unipolar brush cells of the vestibulocerebellum causes ataxia. *FASEB J*. (2018) 32:750–5. doi: 10.1096/fasebj.2018.32.1_supplement.750.5
 47. Lehmann-Horn F, Jurkat-Rott K. Voltage-gated ion channels and hereditary disease. *Physiol Rev*. (1999) 79:1317–72. doi: 10.1152/physrev.1999.79.4.1317
 48. Schaefer L, Sakai H, Mattei MG, Lazdunski M, Lingueglia E. Molecular cloning, functional expression and chromosomal localization of an amiloride-sensitive Na⁺ channel from human small intestine. *FEBS Lett*. (2000) 471:205–10. doi: 10.1016/S0014-5793(00)01403-4
 49. Xiong ZG, Pignataro G, Li M, Chang SY, Simon RP. Acid-sensing ion channels (ASICs) as pharmacological targets for neurodegenerative diseases. *Curr Opin Pharmacol*. (2008) 8:25–32. doi: 10.1016/j.coph.2007.09.001
 50. Bobrow CS, Soothill PW. Causes and consequences of fetal acidosis. *Arch Dis Childhood Fetal Neonatal Ed*. (1999) 80:F246–9. doi: 10.1136/fn.80.3.F246
 51. Zha XM. Acid-sensing ion channels: trafficking and synaptic function. *Mol Brain*. (2013) 6:1. doi: 10.1186/1756-6606-6-1
 52. Waldmann R, Champigny G, Voilley N, Lauritzen I, Lazdunski M. The mammalian degenerin MDEG, an amiloride-sensitive cation channel activated by mutations causing neurodegeneration in *Caenorhabditis elegans*. *J Biol Chem*. (1996) 271:10433–6. doi: 10.1074/jbc.271.18.10433
 53. Ng E, Lind PM, Lindgren C, Ingelsson E, Mahajan A, Morris A, et al. Genome-wide association study of toxic metals and trace elements reveals novel associations. *Hum Mol Genet*. (2015) 24:4739–45. doi: 10.1093/hmg/ddv190
 54. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*. (2016) 44:W90–7. doi: 10.1093/nar/gkw377
 55. Gong Z, Hong F, Wang H, Zhang X, Chen J. An eight-mRNA signature outperforms the lncRNA-based signature in predicting prognosis of patients with glioblastoma. *BMC Med Genet*. (2020) 21:1–13. doi: 10.1186/s12881-020-0992-7
 56. Mashimo K, Ohno Y. Effect of ethanol on gene expression of beating neonatal rat cardiomyocytes-further research with ingenuity pathway analysis software. *J Nippon Med School*. (2020) 88:209–19. doi: 10.1272/jnms.JNMS.2021_88-501
 57. Böhme I, Schönherr R, Eberle J, Bosserhoff AK. Membrane transporters and channels in melanoma. In: *Reviews of Physiology, Biochemistry and Pharmacology*. Berlin, Heidelberg: Springer (2000). p. 1–106.
 58. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. (2018) 47:D607–13. doi: 10.1093/nar/gky1131
 59. Lingueglia E, Lazdunski M. Pharmacology of ASIC channels. *Wiley Interdiscip Rev Membrane Transport Signal*. (2013) 2:155–71. doi: 10.1002/wmts.88
 60. AbdulAzeez S, Al Qahtani NH, Almandil NB, Al-Amodi AM, Aldakeel SA, Ghanem NZ, et al. Genetic disorder prenatal diagnosis and pregnancy termination practices among high consanguinity population, Saudi Arabia. *Sci Rep*. (2019) 9:17248. doi: 10.1038/s41598-019-53655-8

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

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

Copyright © 2021 Al Qahtani, AbdulAzeez, Almandil, Fahad Alhur, Alsuwat, Al Taifi, Al-Ghamdi, Rabindran Jermy, Abouelhoda, Subhani, Al Asoom and Borgio. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Identification of Key Genes and Pathways in Persistent Hyperplastic Primary Vitreous of the Eye Using Bioinformatic Analysis

Derin M. Thomas*, Chitra Kannabiran and D. Balasubramanian

Kallam Anji Reddy Molecular Genetics Laboratory, Prof. Brien Holden Eye Research Center, LV Prasad Eye Institute, Hyderabad, India

OPEN ACCESS

Edited by:

Balu Kamaraj,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia

Reviewed by:

Sabeena Mustafa,
King Abdullah International Medical
Research Center (KAIMRC),
Saudi Arabia
Umashankar Vetrivel,
Indian Council of Medical Research
(ICMR), India

*Correspondence:

Derin M. Thomas
derinmarythomas@gmail.com

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 03 April 2021

Accepted: 07 June 2021

Published: 13 August 2021

Citation:

Thomas DM, Kannabiran C and
Balasubramanian D (2021)
Identification of Key Genes and
Pathways in Persistent Hyperplastic
Primary Vitreous of the Eye Using
Bioinformatic Analysis.
Front. Med. 8:690594.
doi: 10.3389/fmed.2021.690594

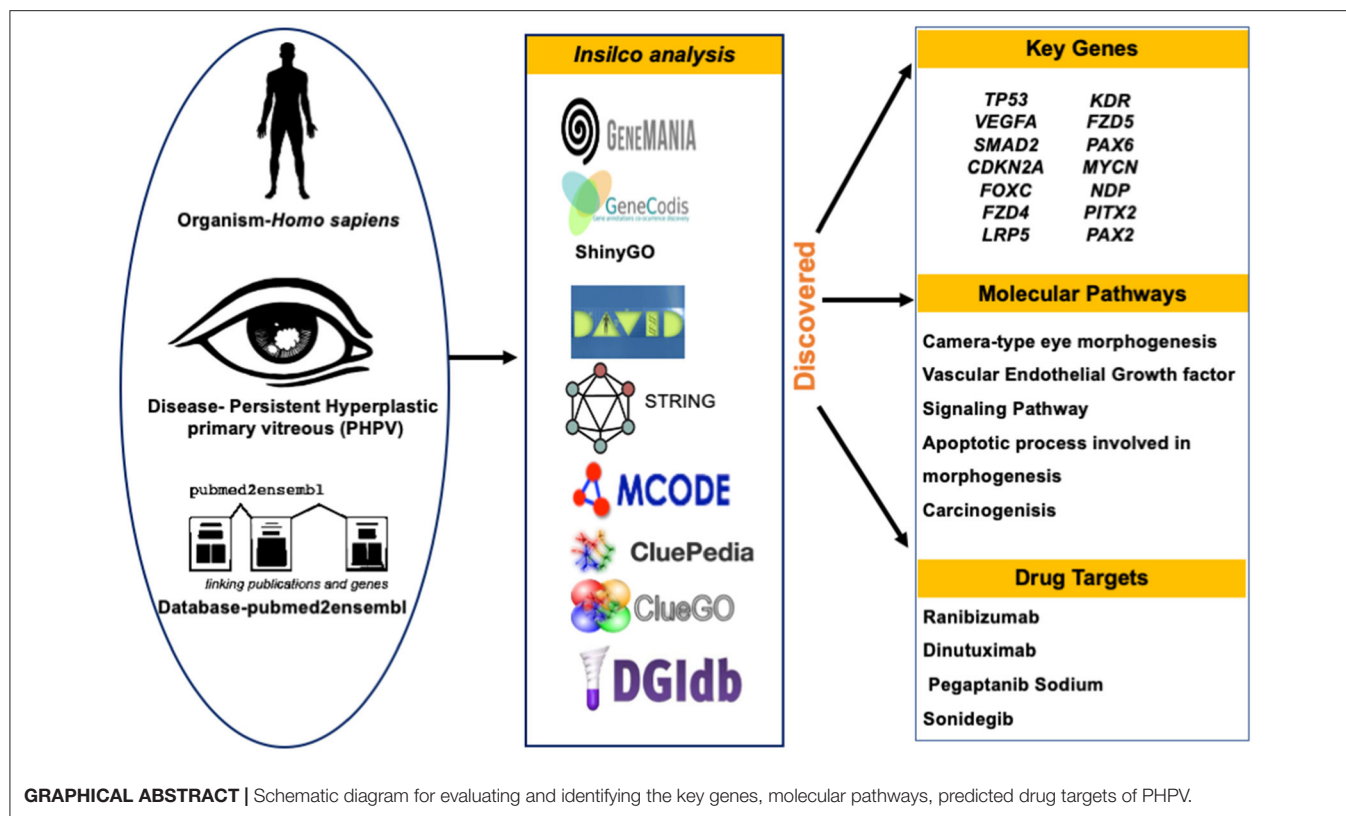
Background: The failure of the embryonic hyaloid vascular system to regress naturally causes persistent hyperplastic primary vitreous (PHPV), a congenital eye disease. PHPVs molecular pathway, candidate genes, and drug targets are unknown. The current paper describes a comprehensive analysis using bioinformatics to identify the key genes and molecular pathways associated with PHPV, and to evaluate potential therapeutic agents for disease management.

Methods: The genes associated with PHPV were identified using the pubmed2ensembl text mining platform. GeneCodis was employed to evaluate the Gene Ontology (GO) biological process terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Search Tool for the Retrieval of Interacting Genes (STRING) constructed a protein-protein interaction (PPI) network from the text mining genes (TMGs) in Cytoscape. The significant modules were clustered using Molecular Complex Detection (MCODE), and the GO and KEGG analysis for the hub genes were analyzed with the Database of Annotation, Visualization and Integrated Discovery (DAVID) tool. ClueGO, CluePedia, and ShinyGo were used to illustrate the functions and pathways of the clustered hub genes in a significant module. The Drug-Gene Interaction database (DGIdb) was used to evaluate drug-gene interactions of the hub genes to identify potential PHPV drug candidates.

Results: A total of 50 genes associated with PHPV were identified. Overall, 35 enriched GO terms and 15 KEGG pathways were discovered by the gene functional enrichment analysis. Two gene modules were obtained from the PPI network constructed with 31 nodes with 42 edges using MCODE. We selected 14 hub genes as core candidate genes: *TP53*, *VEGFA*, *SMAD2*, *CDKN2A*, *FOXC*, *FZD4*, *LRP5*, *KDR*, *FZD5*, *PAX6*, *MYCN*, *NDP*, *PITX2*, and *PAX2*, primarily associated with camera-type eye morphogenesis, pancreatic cancer, the apoptotic process involved in morphogenesis, and the VEGF receptor signaling pathway. We discovered that 26 Food and Drug Administration (FDA)-approved drugs could target 7 of the 14 hub genes.

Conclusions: In conclusion, the results revealed a total of 14 potential genes, 4 major pathways, 7 drug gene targets, and 26 candidate drugs that could provide the basis of novel targeted therapies for targeted treatment and management of PHPV.

Keywords: persistent hyperplastic primary vitreous, gene ontology, bioinformatic analysis, hub genes, molecular pathway



INTRODUCTION

Persistent hyperplastic primary vitreous (PHPV) is a rare vitreoretinal disorder which accounts for up to 5% of blindness (1). PHPV's pathophysiology occurs during the embryonic stage, with vessel development occurring in the third week of pregnancy. The hyaloid artery system expands and extends to the anterior part of the eye forming the iridohyaloid or capsulopupillary artery during this period. At this point of development, the posterior tunica vasculosa lentis, which is an anastomosis of vessels at the back of the lens begins to develop and nourishes the lens. Secondary vitreous begins to develop in place of primary vitreous during the second trimester of pregnancy. The pathological persistence of fetal intraocular vessels including the hyaloid artery in embryonic vitreous causes this congenital eye disease (2, 3). Apoptosis or macrophage activation causes hyaloidal artery regression which is accompanied by vasa hyaloidal propria, iridohyaloid, and tunica vasculosa lentis (4). White retrolental tissue, an anteriorly swollen lens, centrally dragging ciliary structures, and varying degrees of lenticular opacification are the most prominent clinical symptoms (5). These vascular remnants can hinder the normal retinal development leading to retinal detachment and optic nerve or macula anomalies, and it can also appear in anterior, lateral, or combined forms in various patients (5, 6). PHPV is usually detected in infants within the first 3 months of life due to leukocoria, microphthalmos, and strabismus (7). **Figure 1** represents the typical morphology of a PHPV subject. PHPV is

also known as persistent fetal vasculature (PFV) (5). Bilateral PHPV is rare and sporadic compared to unilateral PHPV; however, it is an autosomal dominant or recessive trait that may be inherited (8, 9).

A large number of genes are involved in the development and regression of the hyaloid artery. PHPV traits have been observed in human and animal models. In humans, PHPV incidence was found to be an autosomal dominant inheritance pattern in an Egyptian family (10). Mutations in the *NDP* gene and the *COX15* gene on chromosome 10 have been found in cases of bilateral PHPV (11–14). The *ZNF408* gene, which had previously been found in retinitis pigmentosa and autosomal dominant familial exudative vitreoretinopathy (ADFEVR) was also identified in PHPV cases of microcornea, posterior megalolenticonus, and coloboma syndrome (MPPC syndrome) (15). *FZD4* (frizzled-type receptor 4) was reported to be associated in certain PHPV cases as well as a gene linked to familial exudative vitreoretinopathy (FEVR) (16). In animal models, various signaling pathways have been implicated in the pathogenesis of PHPV including protooncogene *ski*, *p53*, tumor suppressor gene *Arf*, ephrin-B2, β A3/A1-crystallin, *LRP5*, *ang-2*, *Bax* and *Bak*, *FZD4*, and *ephrin-A5*. FEVR, incontinentia pigmenti, retinoblastoma, and retinopathy of prematurity (RoP) are some of the conditions that mimic PHPV-like symptoms (17–23). However, the regulatory mechanisms responsible and genes involved in the process of fetal vascular regression continue to be unclear, as does the underlying cause of failure of regression.

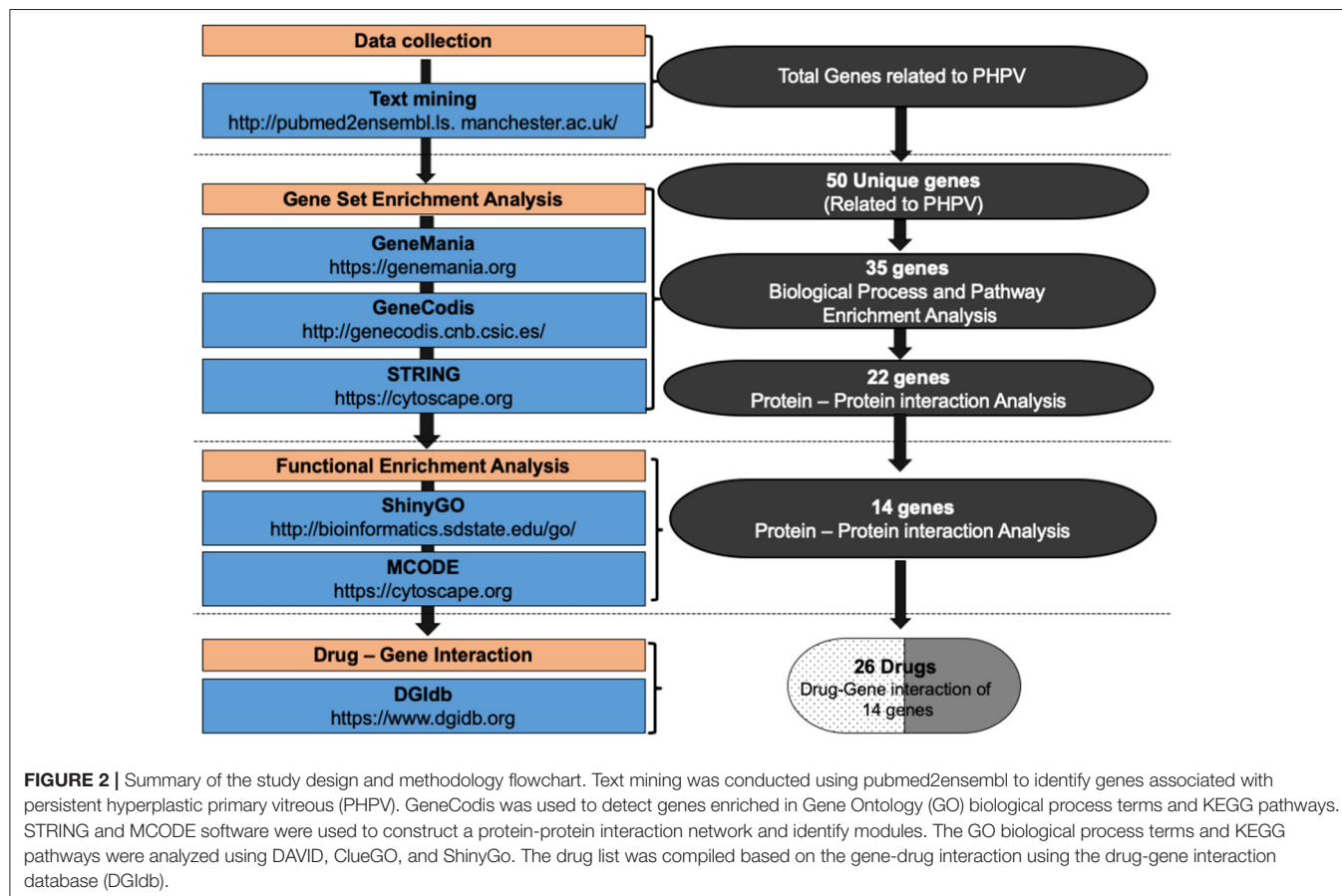
The current surgical management of PHPV is primarily based on the pathological presence of individual cases. Depending on the ocular pathology of PHPV, the limbal and pars plicata incisions are the two most frequent surgical incision methods (24). The most common criteria for surgical intervention



FIGURE 1 | Morphology of persistent hyperplastic primary vitreous in a 1-year-old boy with white vascularized retrolental tissue. Image obtained with prior informed consent from the parents of the patient.

are severe media opacities due to cataract or retrolental membranes, progressive anterior chamber shallowing due to cataract, uncontrolled glaucoma or secondary ocular hypotony related to ciliary process dragging, vitreous hemorrhage, and retinal detachment following vitreoretinal traction (5, 25). In cases with advanced pathology, such as acute optic nerve hypoplasia, severe retinal detachment, or microphthalmia, surgery is not a preferred choice since post-operative vision is often low (24). Non-surgical management is currently used in non-progressive conditions and patients with non-central opacity that does not cause any visual impairment. If a non-surgical alternative is used, diligent follow-up should be carried out to detect any potential risks, such as cataract progression or glaucoma (26). The disease's heterogeneity continues to render PHPV diagnosis and treatment challenging.

Since PHPV is a rare disease, understanding the mechanisms that constitute a group of phenotypes is often restricted by small sampling sizes. Therefore, comprehending the molecular mechanisms underlying the expression of the mutated gene, which leads to improper vascular remodeling and the formation of PHPV is often individual-specific and critical for diagnosis, prevention, and therapeutic management. The assessment and analysis of molecular pathways and genetic variant analysis using conventional variant detection



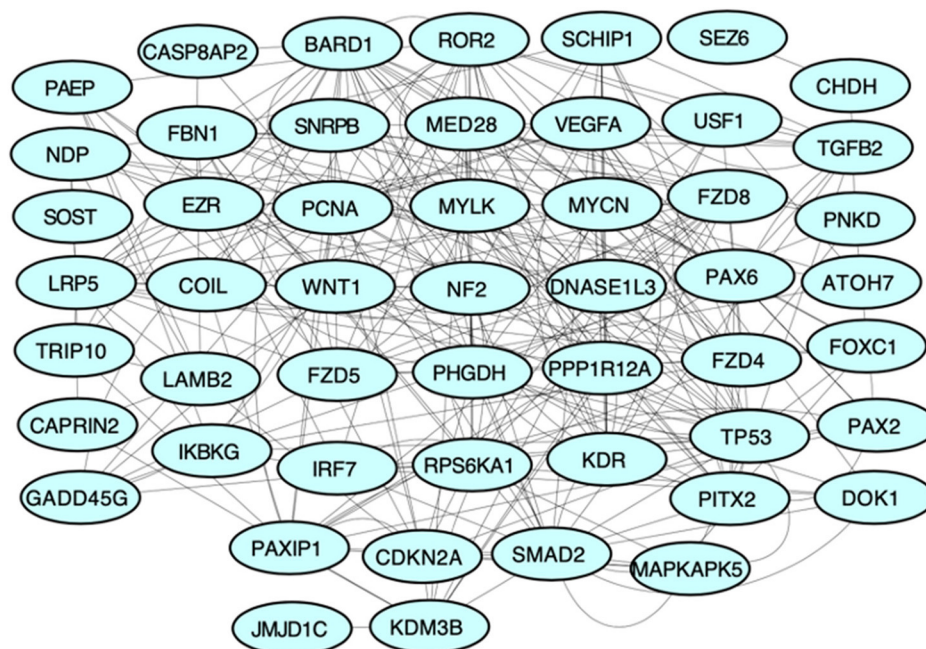


FIGURE 3 | Protein-protein interaction network of all TMGs related to PHPV. The network, genetic interactions, co-expression analysis, and pathways evaluated using Genemania, a Cytoscape plugin, are represented in the figure.

approaches such as Sanger sequencing, next generation sequencing, FISH, aCGH, and GTG banding can be time-consuming, expensive, and results in complicated data analysis for unspecified variants (27–31). Text mining is an effective tool for generating a hypothesis since it can reveal novel correlations between genes and the disease pathologies (32). Integration of text mining with biological knowledge and a bioinformatic approach provides new insights into the potential to reconfigure existing drugs (33). By integrating biological databases and *in silico* tools, the present paper aims to explore possible molecular mechanisms (if any) and classify the causative genes responsible for the heterogenic disease PHPV, thus discovering new drug targets for the treatment of the disease.

METHODS

Selection of Key Genes Using Text Mining Analysis

To identify genes related to PHPV, text mining analysis was performed using pubmed2ensembl (<http://pubmed2ensembl.ls.manchester.ac.uk>) which revealed associations between genes and the literature for data extraction. It is a freely accessible database that connects over 2,000,000 articles in PubMed publications to 150,000 Ensembl genes from 50 species (34, 35). To create a list of key genes, we used search terminology “Persistent Hyperplastic Primary Vitreous” and “Persistent Fetal Vasculature” from 100,000 relevant document IDs. The search terms used were confined to avoid overlapping genes related with

other ocular disorders. The species dataset was set to “*Homo sapiens* (GRCh37)” and the query result was constrained using “filter on MEDLINE: PubMed ID”. The unduplicated genes were extracted and the TMGs were recovered as the intersection of gene hits from the two sets. **Figure 2** represents the methodology flowchart and summary of the study design.

Pathway Enrichment and Biological Process Analysis

The TMGs obtained from text mining were analyzed for biological process annotations. The tool GeneCodis, a web-based server, was used (<http://genecodis.cnb.csic.es/>) to execute an enrichment analysis of the TMGs. GeneCodis assesses functional analysis of gene lists that integrates different sources of information which includes Gene Ontology (GO) [a collection of terminology that describe gene products in terms of Biological Process (BP), Molecular Function (MF), and Cellular Component (CC)], KEGG pathways (offers evidence on biological metabolic pathways that are well-known), and InterPro motifs (36). The organism chosen for the analysis was set as *Homo sapiens*. The TMGs were used as the input set, and genes with significantly enriched biological processes relevant to eye development and vasculogenesis were chosen using an adjusted *P*-value and analyzed using the GO and BP categories. Using the same method, the genes from the selected annotations were used for KEGG pathway analysis and the genes obtained by the KEGG pathway analysis were further analyzed (28). GeneMania (version 3.5.2), a Cytoscape plugin (version 3.8.2), was used to construct a gene-gene functional interaction network from the TMGs. The

TABLE 1 | Top 15 enriched Gene Ontology (GO) biological process terms assigned to the text mining genes.

Biological process	Genes in query set	Total genes in the genome	Corrected P-value	Genes
Extracellular matrix-cell signaling	3	4	2.33E-08	<i>FZD4, NDP, LRP5</i>
Multicellular organism development	9	1,103	2.51E-08	<i>FZD4, FZD5, VEGFA, PAX2, PAEP, LRP5, KDR, ATOH7, FOXC1</i>
Regulation of transcription by RNA polymerase ii	9	1,102	3.12E-08	<i>VEGFA, TP53, PAX2, MYCN, SMAD2, JMJD1C, ATOH7, FOXC1, KDM3B</i>
Norrin signaling pathway	3	3	3.50E-08	<i>FZD4, NDP, LRP5</i>
Positive regulation of transcription, DNA-templated	8	701	3.77E-08	<i>FZD4, TP53, PAX2, NDP, MYCN, SMAD2, LRP5, FOXC1</i>
Positive regulation of transcription by RNA polymerase ii	9	1,068	4.73E-08	<i>IKBKG, FZD5, VEGFA, TP53, PAX2, MYCN, SMAD2, LRP5, FOXC1</i>
Negative regulation of gene expression	6	291	1.27E-07	<i>VEGFA, TP53, TGFB2, MYCN, SMAD2, KDR</i>
Retina vasculature morphogenesis in camera-type eye	3	7	1.53E-07	<i>FZD4, NDP, LRP5</i>
Heart development	5	234	2.04E-06	<i>TP53, TGFB2, PCNA, SMAD2, FOXC1</i>
Vascular endothelial growth factor signaling pathway	3	16	2.17E-06	<i>VEGFA, KDR, FOXC1</i>
Regulation of transcription, DNA-templated	7	998	4.75E-06	<i>TP53, PAX2, MYCN, SMAD2, JMJD1C, ATOH7, FOXC1</i>
Negative regulation of cell population proliferation	5	438	3.78E-05	<i>FZD5, TP53, TGFB2, NF2, SMAD2</i>
Positive regulation of epithelial to mesenchymal transition	3	48	4.27E-05	<i>TGFB2, SMAD2, FOXC1</i>
Post-embryonic camera-type eye development	2	4	4.52E-05	<i>FZD5, VEGFA</i>
Heart morphogenesis	3	53	4.76E-05	<i>VEGFA, TGFB2, FOXC1</i>
<i>In utero</i> embryonic development	4	202	4.76E-05	<i>VEGFA, TP53, SMAD2, FOXC1</i>
Positive regulation of endothelial cell chemotaxis by VEGF-activated vascular endothelial growth factor receptor signaling pathway	2	5	4.90E-05	<i>VEGFA, KDR</i>
Vascular endothelial growth factor receptor-2 signaling pathway	2	5	4.90E-05	<i>VEGFA, KDR</i>
Positive regulation of cell population proliferation	5	519	4.96E-05	<i>VEGFA, TGFB2, PAX2, LRP5, KDR</i>
Cell differentiation	6	943	5.02E-05	<i>VEGFA, PAX2, SMAD2, KDR, ATOH7, FOXC1</i>
Positive regulation of gene expression	5	486	5.03E-05	<i>VEGFA, TP53, MYCN, SMAD2, FOXC1</i>
Vasculogenesis	3	61	5.39E-05	<i>FZD4, VEGFA, KDR</i>
Wnt signaling pathway	4	226	5.41E-05	<i>FZD4, FZD5, NDP, LRP5</i>
Retinal blood vessel morphogenesis	2	6	5.88E-05	<i>FZD4, LRP5</i>
Positive regulation of transcription from RNA polymerase ii promoter in response to hypoxia	2	6	5.88E-05	<i>VEGFA, TP53</i>

advanced statistical options used were max resultant genes = 20, max resultant attributes = 10, and the automatically selected network weighting function. The resulting network comprised functional annotations from GO as well as genes most closely related to the original list.

Construction of Protein-Protein Interaction Network and Module Analysis

STRING (version 1.6.0) was used to construct the PPI network of 35 enriched genes based on GO. STRING is a web-based database comprising nearly 24.6 million proteins and over 3.1 billion interactions from 5,090 distinct species [https://string-db.org/cgi/input.pl; (37)]. The fundamental metrics of nodes in network theory are connectivity degree (k), Betweenness Centrality (BC),

Closeness Centrality (CC), Eigenvector Centrality (EC), and eccentricity. However, the main advantage of PPI network analysis is to accommodate a wide range of biological processes including inputs pathway information, providing confidence scores based on evidence from conserved genomic neighborhoods, gene-fusion events, co-occurrence events, co-expression data, experimental data, database information, text mining, and homology. In the PPI network, nodes with a high degree known, as hub proteins, are critical proteins because they may correlate to disease-causing genes while nodes with a high BC, known as bottlenecks, prefer to signify important genes because they can be compared to highly used intersections on major highways or bridges. The confidence score of 0.900 was specified as the minimum criterion. The molecular interaction network was then visualized and hub genes were identified using

TABLE 2 | Top 10 enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways assigned to the text mining genes.

PHPV-KEGG pathway	Genes in the query set	Total genes in the genome	Corrected P-value	Genes
Pathways in cancer	9	368	1.24E-12	<i>IKBKG, FZD4, FZD5, VEGFA, TP53, TGFB2, SMAD2, LRP5, LAMB2</i>
Proteoglycans in cancer	7	141	5.54E-12	<i>FZD4, FZD5, VEGFA, TP53, TGFB2, SMAD2, KDR</i>
Hepatocellular carcinoma	6	90	3.16E-11	<i>FZD4, FZD5, TP53, TGFB2, SMAD2, LRP5</i>
Gastric cancer	6	89	3.94E-11	<i>FZD4, FZD5, TP53, TGFB2, SMAD2, LRP5</i>
Pancreatic cancer	5	62	8.60E-10	<i>IKBKG, VEGFA, TP53, TGFB2, SMAD2</i>
Human papillomavirus infection	6	179	1.39E-09	<i>IKBKG, FZD4, FZD5, VEGFA, TP53, LAMB2</i>
Hippo signaling pathway	5	81	2.42E-09	<i>FZD4, FZD5, TGFB2, NF2, SMAD2</i>
MAPK signaling pathway	5	192	1.66E-07	<i>IKBKG, VEGFA, TP53, TGFB2, KDR</i>
Wnt signaling pathway	4	69	1.96E-07	<i>FZD4, FZD5, TP53, LRP5</i>
Cell cycle	4	74	2.35E-07	<i>TP53, TGFB2, PCNA, SMAD2</i>
PI3K-Akt signaling pathway	5	222	2.49E-07	<i>IKBKG, VEGFA, TP53, LAMB2, KDR</i>
Breast cancer	4	87	3.77E-07	<i>FZD4, FZD5, TP53, LRP5</i>
Fluid shear stress and atherosclerosis	4	95	4.97E-07	<i>IKBKG, VEGFA, TP53, KDR</i>
Hepatitis B	4	127	1.49E-06	<i>IKBKG, TP53, TGFB2, PCNA</i>
Basal cell carcinoma	3	31	1.82E-06	<i>FZD4, FZD5, TP53</i>

the Cytoscape software which visually presents the integration of gene expression, biological network, and genotype (38). In this study, the hub nodes were classified by a high score based on the network's scale-free property and was used for centrality analysis by analyzing the network topology (39) and considered the sub-network of these key proteins as the backbone which was worth exploring further in the signaling pathways involved in eye development. Further, a built in Cytoscape plugin Molecular Complex Detection (MCODE, version 2.0.0) was used to distinguish the significant gene modules (clusters) and hub genes from the PPI network (40). The cutoff parameters were “degree cutoff = 2,” “node score cutoff = 0.2,” “k-core = 2,” and “max depth = 100” (41).

Drug-Gene Interactions

The Drug-Gene Interaction Database (DGIdb) (www.dgiddb.org) is an online resource that consolidates data from various sources to illustrate drug-gene interactions and gene druggability (42). We investigated drug-gene interactions used in significant module genes as the potential targets for existing drugs or compounds using DGIdb (Version 3.0). The PubChem database was used to obtain the chemical structure of the identified drugs (<https://pubchem.ncbi.nlm.nih.gov>). It has over 25 million specific chemical structures and 90 million bioactivity outcomes linked to thousands of macromolecular targets.

RESULTS

Identification of Candidate Genes

We obtained 50 unique genes in *Homo sapiens* associated with PHPV using the TMG approach. **Figure 3** depicts the network, genetic interactions, co-expression analysis, and pathways of the 50 TMGs assessed by GeneMania. From these, 35 genes were

selected as candidate genes for enrichment analysis based on their GO and molecular pathways.

Enrichment Analysis of TMGs

The most enriched terminology directly linked to the pathology of vasculature morphogenesis of the eye contributing to PHPV was identified using GeneCodis (with $P = 1.00E-07$), GO, biological process (BP), and KEGG. The GO and BP annotations analysis identified 35 significantly enriched genes. The 10 most enriched functions were “extracellular matrix cell signaling” ($P = 2.33E-08$), “multicellular organism development” ($P = 2.51E-08$), “regulation of transcription by RNA polymerase ii” ($P = 3.12 E-08$), “Norris signaling pathway” ($P = 3.50E-08$), “regulation of transcription, DNA templated” ($P = 3.77E-08$), “negative regulation of gene expression” ($P = 1.27E-07$), “retina vasculature morphogenesis in camera-type eye” ($P = 1.53E-07$), “heart development” ($P = 2.04E-06$), “vascular endothelial growth factor (VEGF) signaling pathway” ($P = 2.17E-06$), and “post-embryonic camera type eye development” ($P = 4.52E-05$). Overall, 15 major pathways involving 12 TMGs were discovered by KEGG enrichment analysis. The five most significantly enriched pathways were “pathways in cancer” ($P = 11.24E-12$), “proteoglycans in cancer” ($P = 5.54E-12$), “hepatocellular carcinoma” ($P = 3.16E-11$), “gastric cancer” ($P = 3.94E-11$), and “pancreatic cancer” ($P = 8.69E-10$), involving 9, 7, 6, 6, and 5 text mining genes, respectively. **Table 1** displays 15 enriched GO terms and **Table 2** exhibits the KEGG analysis of 10 enriched molecular pathways of the TMGs.

PPI Network Construction, Modular Analysis, and Key Genes Identification

STRING was used to construct a PPI network for the 35 target genes with a high confidence score >0.900. There

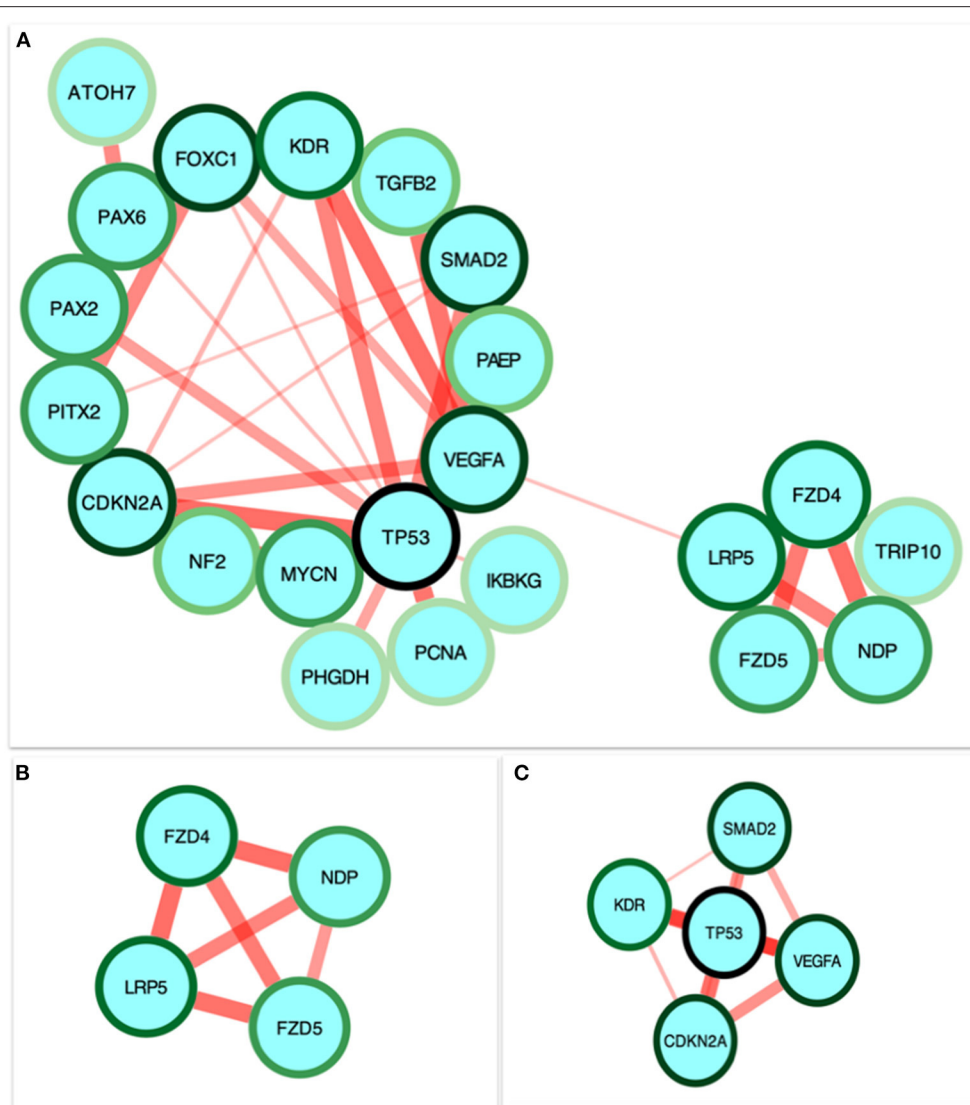


FIGURE 4 | Identification and enrichment analysis of the text mining genes (TMGs). **(A)** The protein-protein interaction (PPI) network of the 35 target TMGs was visualized using Cytoscape. The gradient of the genes node border color reflects its role in the eye development process (a dark green node border represents the strongest degree of association and a light green node border represents the weakest degree of association). **(B,C)** The two modules were obtained from the PPI network using MCODE. **(B)** Module 1, the most significant module with four nodes. **(C)** Module 2 with five nodes.

were 31 nodes and 42 edges in the network (**Figure 4A**). Using a cluster analysis of filtering nodes, 14 hub node genes were identified among 31 nodes (**Table 3**). The hub genes identified were *TP53*, *VEGFA*, *SMAD2*, *CDKN2A*, *FOXC1*, *FZD4*, *LRP5*, *KDR*, *FZD5*, *PAX6*, *MYCN*, *NDP*, *PITX2*, and *PAX2*. The REVIGO analysis of the hub genes revealed five clusters based on GO similarity which were primarily related to eye development, Wnt signaling pathway, cell proliferation, regulation of cell migration, and regulation of angiogenesis (**Figure 5**). The modular analysis performed using MCODE yielded two modules. The PPI network relies on a total of 9 genes, as module 1 (*FZD4*, *FZD5*, *LRP5*, and *NDP*) contained 4 genes with 10 edges and module 2 (*TP53*, *KDR*,

VEGFA, *CDKN2A*, and *SMAD2*) contained 5 genes with 6 edges (**Figures 4B,C**). According to the pathway enrichment analysis using KEGG and the ShinyGo platform, the genes in module 1 were associated with VEGF signaling pathway, regulation of execution process of apoptosis, and cell migration involved in sprouting angiogenesis. The module 2 genes were significantly associated with eye development, retinal vasculature development, and Wnt signaling pathway (**Figure 6**). Overall, the enrichment analysis revealed that these genes were substantially enriched in cell proliferation, anatomical structure morphogenesis, and regulation of developmental process which play a crucial role in vasculature formation of the lens causing PHPV.

TABLE 3 | Hub node genes in the protein-protein interaction network identified with a filtering node degree ≥ 2 .

Genes	Degree	MCODE cluster	MCODE node status	MCODE SCORE
<i>TP53</i>	12	Clustered	Module 1	4
<i>VEGFA</i>	9	Clustered	Module 1	4
<i>SMAD2</i>	7	Seed	Module 1	4
<i>CDKN2A</i>	6	Clustered	Module 1	4
<i>FOXC1</i>	5	Unclustered	–	1
<i>FZD4</i>	4	Clustered	Module 2	3
<i>LRP5</i>	4	Clustered	Module 2	3
<i>KDR</i>	4	Clustered	Module 1	4
<i>FZD5</i>	3	Clustered	Module 2	3
<i>PAX6</i>	3	Unclustered	–	2
<i>MYCN</i>	3	Unclustered	–	3
<i>NDP</i>	3	Seed	Module 2	3
<i>PITX2</i>	3	Unclustered	–	2
<i>PAX2</i>	3	Unclustered	–	2

Drug-Gene Interaction Analysis of Core Genes

In the drug-gene interaction study, we selected 14 hub genes as potential drug targets (Table 4). Overall, 7 of the 14 are potential gene targets and 26 FDA-approved drugs are expected to have drug-gene interactions. *FOXC1*, *FZD4*, *LRP5*, *FZD5*, *PAX6*, *NDP*, and *PITX2* were the exceptions. The major interactions among drugs, genes, and pathways are depicted in Table 4. Table 5 represents the chemical structure and formula of the identified drugs.

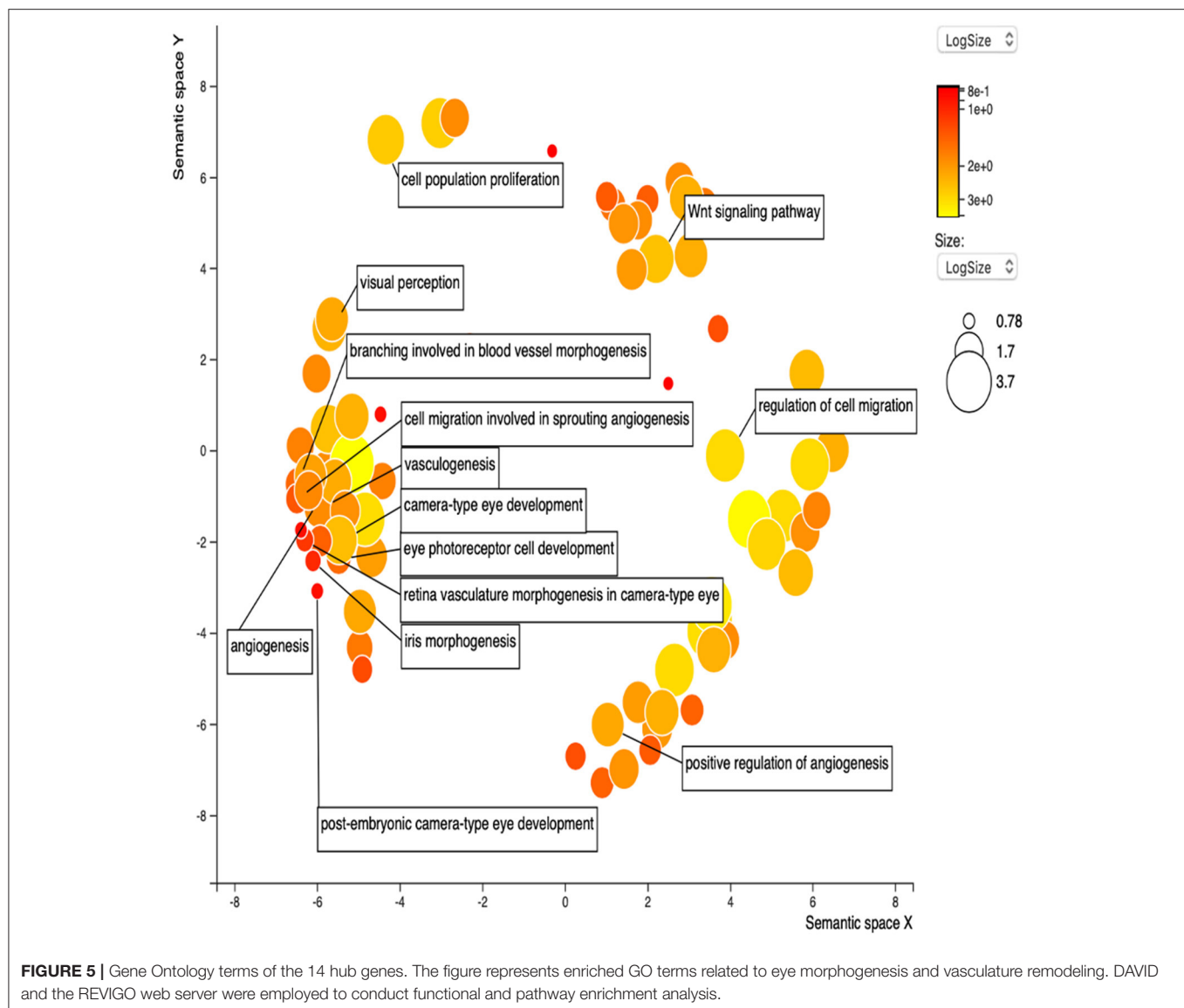
DISCUSSION

PHPV is a disease that leads to blindness or severe vision loss, although there are currently few therapeutic choices (24, 25). On the other hand, PHPV patients are more likely to develop cataracts and closed-angle glaucoma early on in life. Terminal glaucoma, uveitis, retinal detachment, and intra-ocular hemorrhage can be inevitable for these patients (35). As a consequence, the molecular mechanisms that contribute to PHPV must be established. Our analysis discloses that the molecular mechanism of PHPV overlaps with various other signaling pathways contributing to a broader range of therapeutic targets and prognostic biomarkers. The present paper reports 35 genes that might be involved in the development of the eye’s vasculature process in the PHPV condition. The enriched GO and BP terms assigned to these genes were associated mainly with extracellular matrix-cell signaling, multicellular organism development, regulation of transcription by RNA polymerase ii, Norrin signaling pathway, and retina vasculature morphogenesis in camera-type eyes. The PPI network and enrichment analysis identified 14 hub genes, *TP53*, *VEGFA*, *SMAD2*, *CDKN2A*, *FOXC*, *FZD4*, *LRP5*, *KDR*, *FZD5*, *PAX6*,

MYCN, *NDP*, *PITX2*, and *PAX2* that were involved in camera-type eye morphogenesis, pancreatic cancer, the apoptotic process involved in morphogenesis, and the VEGF receptor signaling pathway (Figure 7). The functional analysis and pathways of the key genes in module 1 and module 2 illustrated using ClueGO are displayed in Figure 7A. Figure 7B displays the distribution of functions and pathways among core genes, while Figure 7C reveals KEGG pathways and enriched GO terms, with colors allocated to each pathway.

Based on the evaluations, four genes such as *FZD4*, *LRP5*, *FZD5*, *NDP* were involved in the process of eye development, retina vasculature development, retinal blood vessel morphogenesis, and the Wnt signaling pathway [Figure 6B; (43, 44)]. The architecture of the retinal vasculature is dependent on highly organized signaling between various cell types of retina, combining internal metabolic conditions with external influences such as oxygen and nutrient supply. In various organs, including the eye, the Wnt signaling pathway is essential for vascular morphogenesis. During eye development, and in vascular eye disorders, Wnt ligands and receptors are key regulators of ocular angiogenesis and also control the development of structured layers of vasculature in retinas as well as the regression of hyaloid vessels (45). FEVR (an inherited disease in which the peripheral retina is hypovascularized to varying degrees) has been attributed to mutations in Wnt pathway components *FZD4*, *LRP5*, and the secreted cysteine-knot protein Norrin (46, 47). Norrin is a non-Wnt ligand with a high affinity *FZD4* receptor located in the retina and activates the Wnt/ β -catenin pathway. Norrie disease, retinopathy of prematurity, and Coats disease are vascular retinopathies caused by defects in the Norrin gene (48). In humans, mutations in *NDP* and *FZD4* have been identified in a limited number of unilateral and bilateral PHPV patients (14, 49–51). *ATOH7* mutation (N46H-homozygous) in a family of autosomal recessive PHPV disease traits linked to 10q21 has been identified (52). These variations include deletions, insertions, and missense and nonsense mutations. However, individuals with X-linked FEVR, autosomal dominant FEVR, retinopathy of prematurity, and Norrie disease have also been reported to have mutations in *NDP* and *FZD4* genes (53). According to the GO analysis, five genes *TP53*, *VEGFA*, *SMAD2*, *CDKN2A*, and *KDR* (Figure 6A) are involved in the process of regulation of cell migration by the VEGF signaling pathway, angiogenesis, regulation of muscle cell apoptotic process, and embryonic organ development process. Apoptosis is another crucial process in eye development involving extensive programmed cell death associated with morphogenesis (54).

Previous research on transgenic mice models supports our *in silico* analysis of PHPV to validate the function of these hub genes in hyaloid vasculature regression such as knockouts of the *Arf* tumor suppressor gene (23, 55, 56), *p53* (21, 57), and *Frizzled-5* (57) which were associated with PHPV-like phenotypes in mouse models. During mouse eye development, the *arf* tumor suppressor gene promoted hyaloid vasculature regression and its deficiency may cause a retrolental membrane with persistent hyaloid vessels (9, 23). In *Atoh7* knockout mice, hyaloid vessels persist in the vitreous and proliferate to



supply the retina which lacks intrinsic vasculature (58, 59). The ephrin-A5 family of receptor tyrosine has been demonstrated to be significant in the regression of the primary vitreous in mouse models (60). Furthermore, in mice lacking *LRP5*, a Wnt receptor displayed hyaloid vasculature that lasted throughout their lives (61, 62). Given the correlation between transgenic mouse PHPV phenotypes and the hub genes in human congenital defects affecting the eye morphogenesis or retinal vasculature and molecular signaling pathways in module 1 and module 2, it suggests that the pathogenesis of PHPV is regulated by genes in modules 1 and 2.

Twenty-six drugs identified by the drug-gene interaction analysis were classified as anti-neoplastic agents, ocular vascular disorder agents, kinase inhibitors, immune system functioning agents, or corticosteroids. Among them, four potential drugs such as Ranibizumab, Dinutuximab, Pegaptanib Sodium, and

Sonidegib were identified based on their high drug-gene interaction score (Table 4). Ranibizumab is a recombinant humanized monoclonal antibody fragment that binds to human vascular endothelial growth factor A (VEGF-A) and thereby prevents it from binding to its receptor and blocking the development of new blood vessels (63). Pegaptanib Sodium is an anti-angiogenic drug used to treat neovascular diseases. It specifically binds to the 165 isoform of VEGF, a protein that is involved in angiogenesis and increased blood vessel leakage (64). Ranibizumab and Pegaptanib Sodium are typically used to treat wet age-related macular degeneration, a type of eye disease (61, 62, 65, 66). They are also used to treat macular edema after retinal vein occlusion, diabetic macular edema, and diabetic retinopathy. Dinutuximab is a GD2-binding human/mouse chimeric monoclonal antibody. It has been proven that the action of pro and anti-angiogenic factors regulates

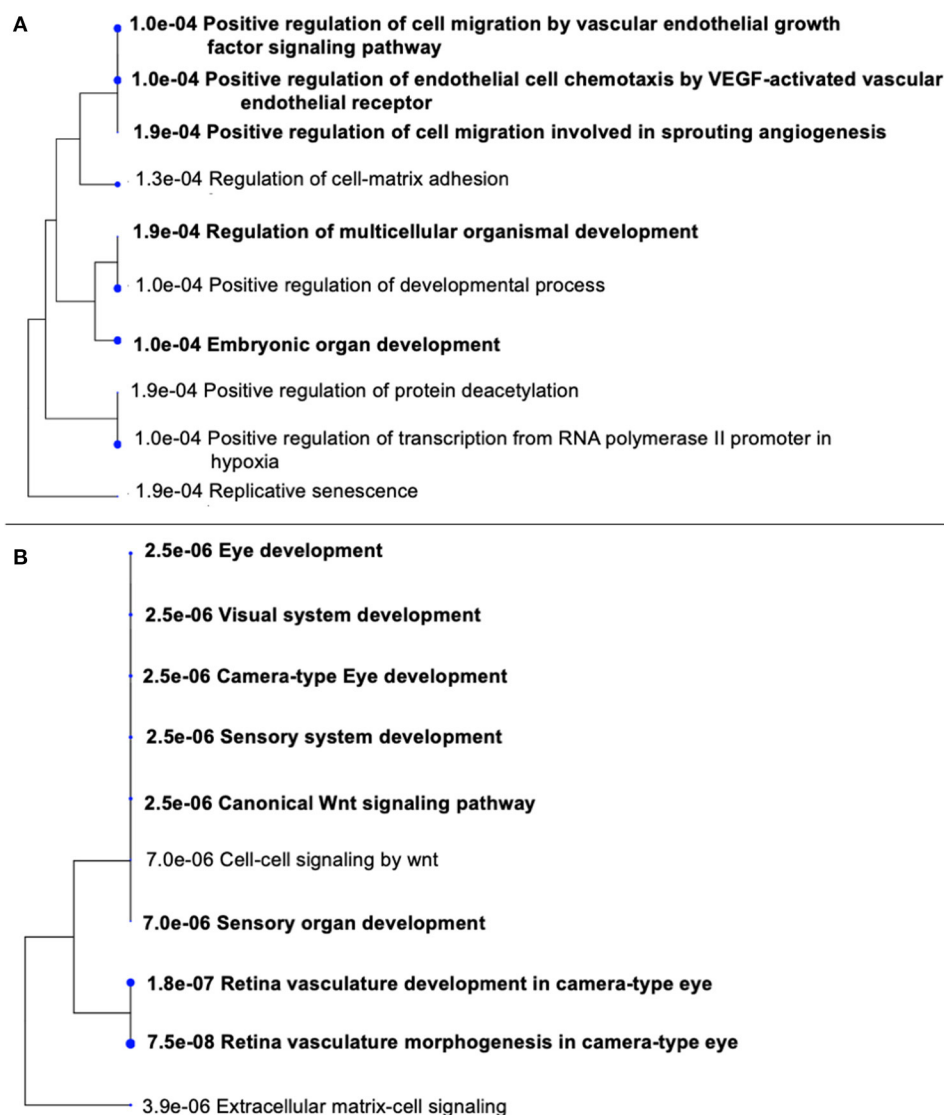


FIGURE 6 | Gene Ontology (GO) terms in the three modules. **(A)** Significantly enriched GO terms in module 1. **(B)** Significantly enriched GO terms in module 2. The functional and pathway enrichment analyses in PHPV were related with high score ($P > 0.005$) based on a tree illustration created using the ShinyGo web server.

angiogenesis in the development of new capillaries from a pre-existing capillary network (67). Dinutuximab binds to GD2 on the cell surface and induces GD2 expressing cells to lyse by antibody-dependent cell-mediated cytotoxicity and complement-dependent cytotoxicity (68, 69). Sonidegib is an anticancer drug that inhibits the hedgehog (Hh) pathway which is involved in cell differentiation, tissue polarity, and stem cell maintenance during embryonic growth. Hh is essential for the development of the hyaloid loop on the lens's ventral surface by promoting VEGF-mediated angiogenesis. In a zebrafish model, the loss of Hh signaling induced excess sprouting of blood vessels in the dorsal eye and impaired the growth of blood vessels in the ventral eye (70). Regulation of the Hh signaling pathway has been associated

with the growth and progression of cancers such as basal cell carcinoma, medulloblastoma (71), and periocular basal cell carcinoma (72).

In PHPV, the ocular fetal vasculature does not go through normal developmental regression. The reasons could be due to presumed loss of apoptosis in PHPV; these natural apoptotic pathways could be pathologically disrupted (22, 73). Apoptosis is a process of cell death that is regulated by a number of gene families. In mice models, the macrophage has been established as a key mediator in studies examining the mechanisms of regression (74, 75). *In silico* drug-gene analysis using the hub genes of PHPV revealed high interaction with anticancer compounds in the present study. It is understood that vascular quiescence can be regulated by a combination of pro and

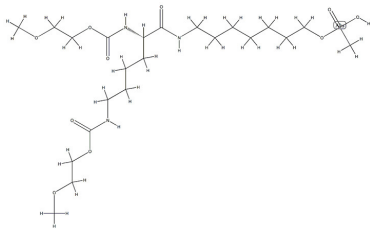
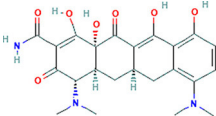
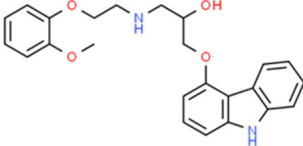
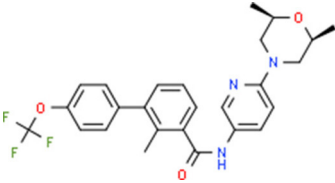
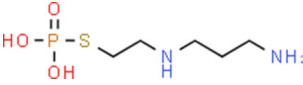
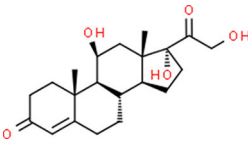
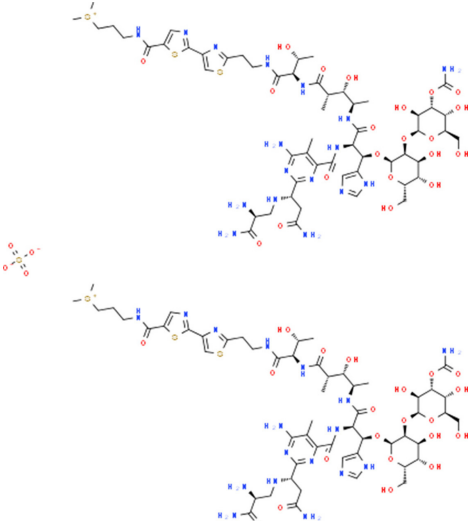
TABLE 4 | Details of the 26 Food and Drug Administration (FDA)-approved drugs that potentially target 7 of the 14 hub genes.

Drug	Gene	Interaction	Interaction score	Drug class	Approved	Reference (PubMed ID)
Ranibizumab	<i>VEGFA</i>	Inhibitor	6.51	Antineoplastic agents, ocular vascular disorder agents	Yes	18046235, 18054637
Pegaptanib Sodium	<i>VEGFA</i>	Antagonist	2.31	Antagonist agent vascular endothelial growth factor (VEGF)	Yes	23953100
Aflibercept	<i>VEGFA</i>	Antibody, binder, inhibitor	1.89	Antagonist agent vascular endothelial growth factor-A (VEGF-A) and placental growth factor (PLGF)	Yes	22813448, 20124951
Minocycline	<i>VEGFA</i>	Inhibitor	0.28	Antibiotic agent	Yes	11875741, 16224178
Carvedilol	<i>VEGFA</i>	Other/unknown	0.18	Beta blocker agent	Yes	15071347, 15942707, 15732037
Dinutuximab	<i>MYCN</i>	Other/unknown	6.31	Antineoplastic agents	Yes	–
Sonidegib	<i>MYCN</i>	Other/unknown	2.52	Antineoplastic agents	Yes	24651015
Amifostine	<i>SMAD2</i>	Unknown	2.24	Chemo protectant, antineoplastic adjunct, or cytoprotective agent	Yes	–
Hydrocortisone	<i>SMAD2</i>	Unknown	0.3	Corticosteroids	Yes	26343583, 27488531, 22983396
Bleomycin	<i>SMAD2</i>	Unknown	0.57	Antitumor antibiotic	Yes	–
Tretinoin	<i>SMAD2</i>	Unknown	0.26	Anticancer drugs	Yes	–
Progesterone	<i>PAX2</i>	Unknown	1.48	Antimineralocorticoid; neurosteroid	Yes	11850818
Vandetanib	<i>KDR</i>	Inhibitor	0.0	Tyrosine kinase inhibitors	Yes	26578684, 20124951
Ramucirumab	<i>KDR</i>	Inhibitor, antagonist, antibody	0.63	Binds to the vascular endothelial growth factor receptor-2	Yes	20048182, 2182794
Lenvatinib	<i>KDR</i>	Inhibitor		Kinase inhibitors, antineoplastic agent	Yes	17943726
Sunitinib	<i>KDR</i>	Inhibitor	0.25	Kinase inhibitors, antineoplastic agent	Yes	27149458, 20142593, 25639617
Sorafenib	<i>KDR</i>	Antagonist, inhibitor		Kinase inhibitors, antineoplastic agent	Yes	16824050, 16418310, 26344591
Axinib	<i>KDR</i>	Unknown	0.32	Kinase inhibitors, antineoplastic agent	Yes	–
Regorafenib	<i>KDR</i>	Inhibitor	0.14	Kinase inhibitors, antineoplastic agent	Yes	27004155
Sorafenib	<i>KDR</i>	Antagonist, inhibitor	0.12	Kinase inhibitors, antineoplastic agent	Yes	16824050, 16418310, 26344591
Abemaciclib	<i>CDKN2A</i>	Unknown	0.92	Antineoplastic agent	Yes	27217383, 26183925
Palbociclib	<i>CDKN2A</i>	Unknown	0.88	Antineoplastic agent	Yes	26715889, 21278246, 22711607
Alectinib	<i>CDKN2A</i>	Unknown	0.65	Antineoplastic agent	Yes	–
Zinc Chloride	<i>TP53</i>	Chaperone	0	Immune system functioning agent	Yes	17327663, 29843463
Trifluridine	<i>TP53</i>	Unknown	0.07	Ophthalmological antiviral agent	Yes	25700705
Bortezomib	<i>TP53</i>	Inhibitor	0	Anticancer drug	Yes	28679691

anti-angiogenic factors. Previous reports have demonstrated that the equilibrium of angiogenic factors such as VEGF and placental growth factor is crucial in vascular regression and

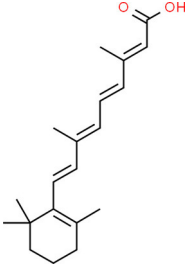
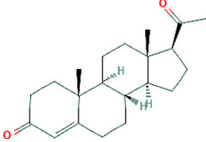
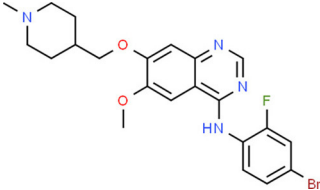
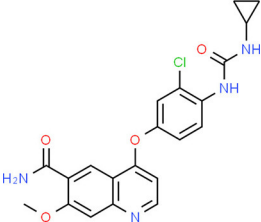
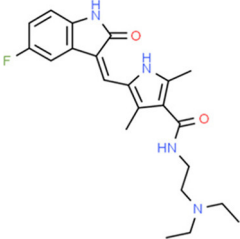
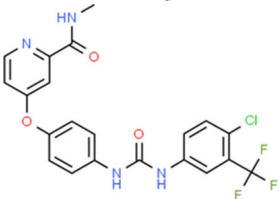
mice lacking angiopoietin 2 which regulates angiogenesis by binding to the Tie2 receptor, maintaining fetal vessels in the eyes (18, 76). In addition, the identified drugs can be used

TABLE 5 | Chemical structure of the potential drugs that target the seven candidate genes.

Drug	Gene target	Chemical structure	Chemical formula/protein formula
Pegaptanib sodium	VEGFA		C ₂₂ H ₄₄ N ₃ O ₁₀ P
Minocycline	VEGFA		C ₂₃ H ₂₇ N ₃ O ₇
Carvedilol	VEGFA		C ₂₄ H ₂₆ N ₂ O ₄
Sonidegib	MYCN		C ₂₆ H ₂₆ F ₃ N ₃ O ₃
Amifostine	SMAD2		C ₅ H ₁₅ N ₂ O ₃ PS
Hydrocortisone	SMAD2		C ₂₁ H ₃₀ O ₅
Bleomycin	SMAD2		C ₁₁₀ H ₁₆₈ N ₃₄ O ₄₆ S ₇

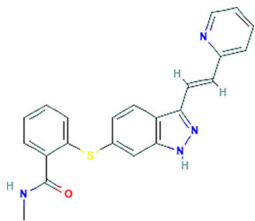
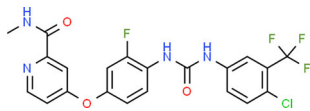
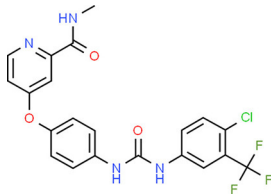
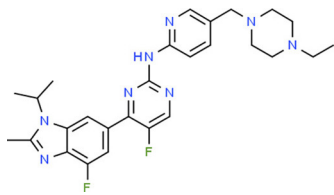
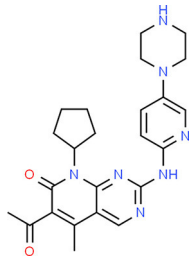
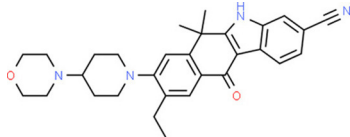
(Continued)

TABLE 5 | Continued

Drug	Gene target	Chemical structure	Chemical formula/protein formula
Tretinoin	SMAD2		C ₂₀ H ₂₈ O ₂
Progesterone	PAX2		C ₂₁ H ₃₀ O ₂
Vandetanib	KDR		C ₂₂ H ₂₄ BrFN ₄ O ₂
Lenvatinib	KDR		₂₁ H ₁₉ ClN ₄ O ₄
Sunitinib	KDR		C ₂₂ H ₂₇ FN ₄ O ₂
Sorafenib	KDR		C ₂₁ H ₁₆ ClF ₃ N ₄ O ₃


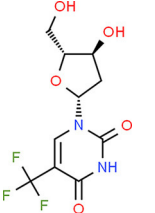
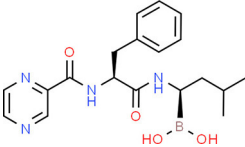
(Continued)

TABLE 5 | Continued

Drug	Gene target	Chemical structure	Chemical formula/protein formula
Axinib	KDR		C ₂₂ H ₁₈ N ₄ OS
Regorafenib	KDR		C ₂₁ H ₁₅ ClF ₄ NO ₃
Sorafenib	KDR		C ₂₁ H ₁₆ ClF ₃ N ₄ O ₃
Abemaciclib	CDKN2A		C ₂₇ H ₃₂ F ₂ N ₈
Palbociclib	CDKN2A		C ₂₄ H ₂₉ N ₇ O ₂
Alectinib	CDKN2A		C ₃₀ H ₃₄ N ₄ O ₂

(Continued)

TABLE 5 | Continued

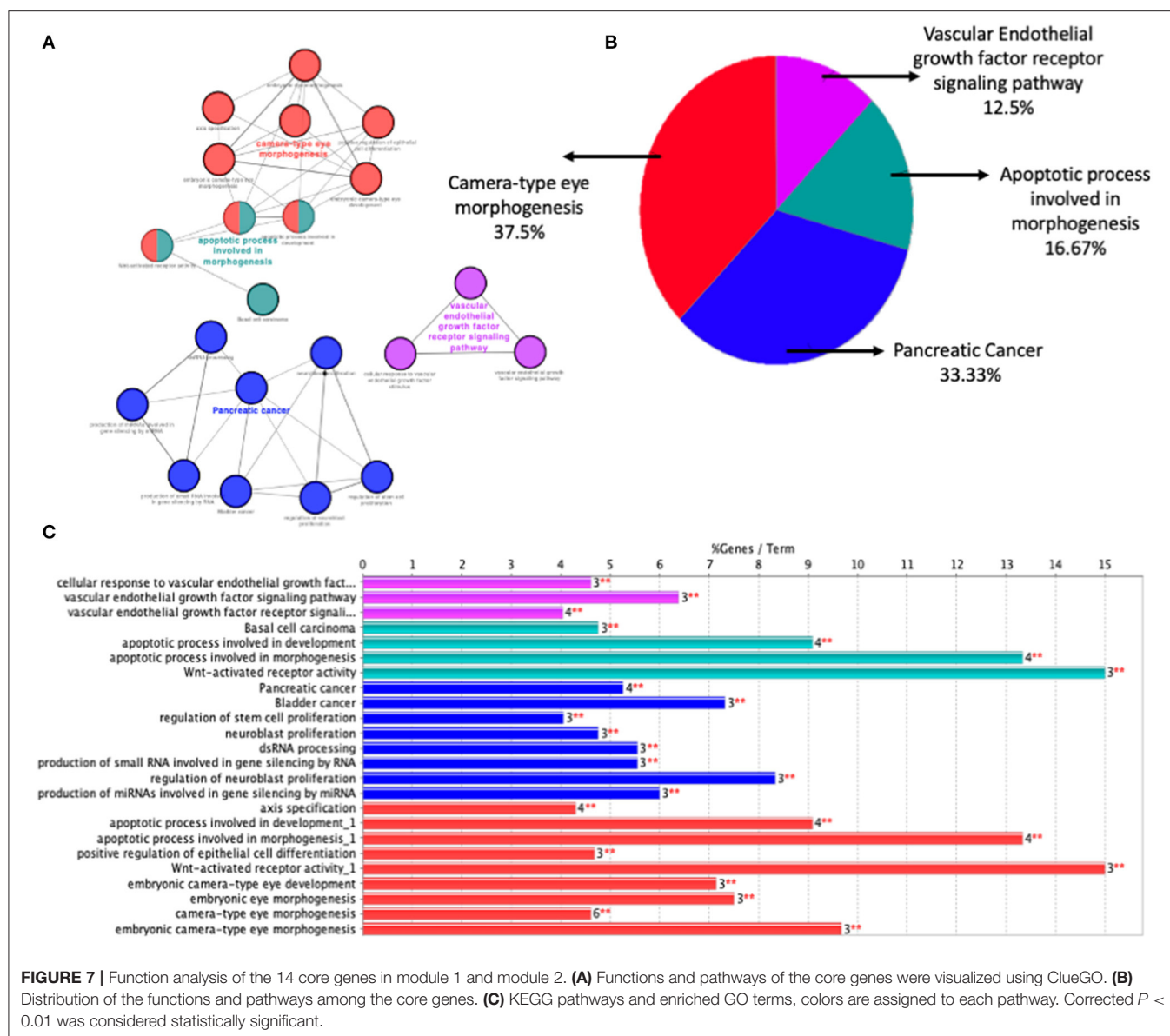
Drug	Gene target	Chemical structure	Chemical formula/protein formula
Zinc chloride	TP53		Cl ₂ Zn
Trifluridine	TP53		C ₁₀ H ₁₁ F ₃ N ₂ O ₅
Bortezomib	TP53		C ₁₉ H ₂₅ BN ₄ O ₄
Ranibizumab	VEGFA	Monoclonal antibody	C ₂₁₅₈ H ₃₂₈₂ N ₅₆₂ O ₆₈₁ S ₁₂
Aflibercept	VEGFA	Monoclonal antibody	C ₄₃₁₈ H ₆₇₈₈ N ₁₁₆₄ S ₃₂
Dinutuximab	MYCN	Monoclonal antibody	C ₆₄₂₂ H ₉₉₈₂ N ₁₇₂₂ O ₂₀₀₈ S ₄₈
Ramucirumab	KDR	Monoclonal antibody	C ₆₃₇₄ H ₉₈₆₄ N ₁₆₉₂ O ₁₉₉₆ S ₄₆

for pharmacological screening in mice and zebrafish models to identify compounds affecting vasculature development that could be of therapeutic importance. The results of the study could lead to a better understanding of the potential molecular pathway and possible hyaloid vasculature mechanism, as well as the development of novel therapeutics to prevent or cure this blinding disease, PHPV. Since the current paper focuses on the appropriate path for understanding molecular pathways and therapeutic options for PHPV through *in silico* analysis, further experimental analysis using animal models is highly recommended to confirm the significance of the candidate genes and pro- and antiangiogenic factors in hyaloid vasculature development and physiology. This continues to be a limitation of the study.

CONCLUSION AND FUTURE PERSPECTIVES

To conclude, no specific candidate genes, molecular pathways, or drug targets have been associated with PHPV until now. *TP53*, *VEGFA*, *SMAD2*, *CDKN2A*, *FOXC*, *FZD4*, *LRP5*, *KDR*, *FZD5*, *PAX6*, *MYCN*, *NDP*, *PITX2*, and *PAX2* were identified for the first time as hub genes using *in silico* tools that may be involved in the development of retinal vasculature and dysfunction of these genes, leading to PHPV. These genes appear to be predominantly associated with functions

related to eye morphogenesis, cancer, apoptosis, and VEGF receptor signaling pathways. Previous reports in knockout *TP53*, *VEGFA*, *FZD4*, and *NDP* transgenic mouse models confirmed the failure of regression of hyaloid vessels and abnormalities in the retinal vasculature. In the future, these *in silico* analyses will be validated by mutation screening of the hub genes in PHPV patients in order to identify pathogenic variants and gene product expressivity. Clearly, more research is warranted on animals and in human patients as the phenotypic differences will differ from individual to individual based on the expressivity of the gene product. In addition, we identified four genes that may be potential drug targets. Precision medicine for a fetal ocular condition like PHPV presents new challenges but with a possibility. Since PHPV is a rare and often autosomal recessive condition, the present paper is useful when there is little pathological knowledge about the disease or where there is substantial pathway heterogeneity, underlying the clinical phenotype. As a result, a combination of therapeutic methods such as surgical intervention and candidate gene identification may be used not only to analyze biological pathways unique to specific cases, but also to propose potential drug combinations based on gene products annotated to the disease associated with PHPV. This research sheds light on the potential of personalized intervention in the treatment of PHPV indicating a substantial advancement in management strategy.



DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

Written informed consent was obtained from the individual(s), and minor(s)' legal guardian/next of kin, for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

DT, CK, and DB conceived, designed the study, and wrote the manuscript. All authors contributed and wrote parts of the manuscript and approved the submitted version.

ACKNOWLEDGMENTS

The authors thank the Department of Biotechnology for providing the financial support (project no. BT/PR28214/MED/97/392/2018) and two reviewers for their helpful comments.

REFERENCES

1. Khaliq S, Hameed A, Ismail M, Anwar K, Leroy B, Payne AM, et al. Locus for autosomal recessive nonsyndromic persistent hyperplastic primary vitreous. *Investigat Ophthalmol Vis Sci.* (2001) 42:2225–8.
2. Reese AB. Persistent hyperplastic primary vitreous: the Jackson Memorial Lecture. *Am J Ophthalmol.* (1955) 40:317–31. doi: 10.1016/0002-9394(55)91866-3
3. Maqsood H, Younus S, Fatima M, Saim M, Qazi S. Bilateral persistent hyperplastic primary vitreous: a case report and review of the literature. *Cureus.* (2021) 13:e13105. doi: 10.7759/cureus.13105
4. Ito M, Yoshioka M. Regression of the hyaloid vessels and pupillary membrane of the mouse. *Anatomy Embryol.* (1999) 200:403–11. doi: 10.1007/s004290050289
5. Goldberg MF. Persistent fetal vasculature (PFV): an integrated interpretation of signs and symptoms associated with persistent hyperplastic primary vitreous (PHPV) LIV Edward Jackson Memorial Lecture. *Am J Ophthalmol.* (1997) 124:587–626. doi: 10.1016/S0002-9394(14)70899-2
6. Müllner-Eidenböck A, Amon M, Moser E, Klebermass N. Persistent fetal vasculature and minimal fetal vascular remnants: a frequent cause of unilateral congenital cataracts. *Ophthalmology.* (2004) 111:906–13. doi: 10.1016/j.ophtha.2003.07.019
7. Antoszyk JH, Antoszyk A. Persistent fetal vasculature 743.51. In: Roy FH, Fraunfelder FW, Fraunfelder FT, editors. *Roy and Fraunfelder's Current Ocular Therapy*. Elsevier Health Sciences (2008). p. 660–2.
8. Jain TP. Bilateral persistent hyperplastic primary vitreous. *Indian J Ophthalmol.* (2009) 57:53. doi: 10.4103/0301-4738.44487
9. Shastry BS. Persistent hyperplastic primary vitreous: congenital malformation of the eye. *Clin Exp Ophthalmol.* (2009) 37:884–90. doi: 10.1111/j.1442-9071.2009.02150.x
10. Galal AH, Kotoury AI, Azzab AA. Bilateral persistent hyperplastic primary vitreous: an Egyptian family supporting a rare autosomal dominant inheritance. *Genet Counsel.* (2006) 17:441–7.
11. Hasbrook M, Yonekawa Y, van Laere L, Shah AR, Capone A. Bilateral persistent fetal vasculature and a chromosome 10 mutation including COX15. *Can J Ophthalmol.* (2017) 52:e203–5. doi: 10.1016/j.jcjo.2017.04.019
12. Payabvash S, Anderson JS, Nascene DR. Bilateral persistent fetal vasculature due to a mutation in the Norrie disease protein gene. *Neuroradiol J.* (2015) 28:623–7. doi: 10.1177/1971400915609350
13. Dhingra S, Shears DJ, Blake V, Stewart H, Patel CK. Advanced bilateral persistent fetal vasculature associated with a novel mutation in the Norrie gene. *Br J Ophthalmol.* (2006) 90:1324–5. doi: 10.1136/bjo.2005.088625
14. Pendergast SD, Trese MT, Liu X, Shastry BS. Study of the Norrie disease gene in 2 patients with bilateral persistent hyperplastic primary vitreous. *Archiv Ophthalmol.* (1998) 116:381–2.
15. Weiner GA, Nudleman E. Microcornea, posterior megalolenticonus, persistent fetal vasculature, and coloboma syndrome associated with a new mutation in ZNF408. *Ophthalm Surg Lasers Imaging Retina.* (2019) 50:253–6. doi: 10.3928/23258160-20190401-10
16. Robitaille JM, Wallace K, Zheng B, Beis MJ, Samuels M, Hoskin-Mott A, et al. Phenotypic overlap of familial exudative vitreoretinopathy (FEVR) with persistent fetal vasculature (PFV) caused by FZD4 mutations in two distinct pedigrees. *Ophthalm Genet.* (2009) 30:23–30. doi: 10.1080/13816810802464312
17. McGannon P, Miyazaki Y, Gupta PC, Traboulsi EI, Colmenares C. Ocular abnormalities in mice lacking the Ski proto-oncogene. *Investig Ophthalmol Visual Sci.* (2006) 47:4231–7. doi: 10.1167/jovs.05-1543
18. Hackett SE, Wiegand S, Yancopoulos G, Campochiaro PA. Angiopoietin-2 plays an important role in retinal angiogenesis. *J Cell Physiol.* (2002) 192:182–7. doi: 10.1002/jcp.10128
19. Freeman-Anderson NE, Zheng Y, McCalla-Martin AC, Treanor LM, Zhao YD, Garfin PM, et al. Expression of the Arf tumor suppressor gene is controlled by Tgfb2 during development. *Development.* (2009) 136:2081–9. doi: 10.1242/dev.033548
20. Salvucci O, Ohnuki H, Maric D, Hou X, Li X, Yoon SO, et al. EphrinB2 controls vessel pruning through STAT1-JNK3 signalling. *Nat Commun.* (2015) 6:6576. doi: 10.1038/ncomms7576
21. Reichel MB, Ali RR, D'Esposito F, Clarke AR, Luthert PJ, Bhattacharya SS, et al. High frequency of persistent hyperplastic primary vitreous and cataracts in p53-deficient mice. *Cell Death Differ.* (1998) 5:156–62. doi: 10.1038/sj.cdd.4400326
22. Hahn P, Lindsten T, Tolentino M, Thompson CB, Bennett J, Dunaief JL. Persistent fetal ocular vasculature in mice deficient in bax and bak. *Archiv Ophthalmol.* (2005) 123:797–802. doi: 10.1001/archophth.123.6.797
23. McKeller RN, Fowler JL, Cunningham JJ, Warner N, Smeyne RJ, Zindy F, et al. The Arf tumor suppressor gene promotes hyaloid vascular regression during mouse eye development. *Proc Natl. Acad Sci USA.* (2002) 99:3848–53. doi: 10.1073/pnas.052484199
24. Dass AB, Trese MT. Surgical results of persistent hyperplastic primary vitreous. *Ophthalmology.* (1999) 106:280–4. doi: 10.1016/S0161-6420(99)90066-0
25. Smith RE. Persistent hyperplastic primary vitreous: results of surgery. *Trans Am Acad Ophthalmol Otolaryngol.* (1974) 78:911–25.
26. Prakhunhungsit S, Berrocal AM. Diagnostic and management strategies in patients with persistent fetal vasculature: current insights. *Clin Ophthalmol.* (2020) 14:4325–35. doi: 10.2147/OPHTH.S236117
27. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet.* (2006) 7:85–97. doi: 10.1038/nrg1767
28. Conrad DF, Hurler ME. The population genetics of structural variation. *Nat Genet.* (2007) 39:S30–6. doi: 10.1038/ng2042
29. Barkholt L, Flory E, Jekerle V, Lucas-Samuel S, Ahnert P, Bisset L, et al. Risk of tumorigenicity in mesenchymal stromal cell-based therapies—bridging scientific observations and regulatory viewpoints. *Cytotherapy.* (2013) 15:753–9. doi: 10.1016/j.jcyt.2013.03.005
30. Bishop R. Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Bioscience Horizons.* (2010) 3:85–95. doi: 10.1093/biohorizons/hzq009
31. Bejjani BA, Shaffer LG. Application of array-based comparative genomic hybridization to clinical diagnostics. *J Mol Diagn.* (2006) 8:528–33. doi: 10.2353/jmoldx.2006.060029
32. Mosca E, Bertoli G, Piscitelli E, Vilaro L, Reinbold RA, Zucchi I, et al. Identification of functionally related genes using data mining and data integration: a breast cancer case study. *BMC Bioinformatics.* (2009) 10(Suppl. 12):S8. doi: 10.1186/1471-2105-10-S12-S8
33. Butcher EC, Berg EL, Kunkel EJ. Systems biology in drug discovery. *Nat Biotechnol.* (2004) 22:1253–9. doi: 10.1038/nbt1017
34. Yu S, Tranchevent L-C, De Moor B, Moreau Y. Gene prioritization and clustering by multi-view text mining. *BMC Bioinformatics.* (2010) 11:28. doi: 10.1186/1471-2105-11-28
35. Hu D, Jiang J, Lin Z, Zhang C, Moonasar N, Qian S. Identification of key genes and pathways in scleral extracellular matrix remodeling in glaucoma: potential therapeutic agents discovered using bioinformatics analysis. *Int J Med Sci.* (2021) 18:1554–65. doi: 10.7150/ijms.52846
36. Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, et al. GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.* (2009) 37:W317–22. doi: 10.1093/nar/gkp416
37. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* (2003) 31:258–61. doi: 10.1093/nar/gkg034
38. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics.* (2011) 27:431–2. doi: 10.1093/bioinformatics/btq675
39. Tang Y, Li M, Wang J, Pan Y, Wu F-X. CytoNCA: a cytoscape plugin for centrality analysis and evaluation of protein interaction networks. *Biosystems.* (2015) 127:67–72. doi: 10.1016/j.biosystems.2014.11.005
40. Wu W, Fang C, Zhang C, Hu N, Wang L. Identification of 10 important genes with poor prognosis in non-small cell lung cancer through bioinformatic analysis. *J Biomed Res Rev.* (2020) 3:41–50. doi: 10.21203/rs.3.rs-21044/v1
41. Bandettini WP, Kellman P, Mancini C, Booker OJ, Vasu S, Leung SW, et al. MultiContrast Delayed Enhancement (MCODE) improves detection of subendocardial myocardial infarction by late gadolinium enhancement cardiovascular magnetic resonance: a clinical validation study. *J Cardiovasc Magnet Reson.* (2012) 14:1–10. doi: 10.1186/1532-429X-14-83

42. Cotto KC, Wagner AH, Feng Y-Y, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res.* (2018) 46:D1068–73. doi: 10.1093/nar/gkx1143
43. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex network: a systematic exploration of the human interactome. *Cell.* (2015) 162:425–40. doi: 10.1016/j.cell.2015.06.043
44. Chen T-C, Lin K-T, Chen C-H, Lee S-A, Lee P-Y, Liu Y-W, et al. Using an in situ proximity ligation assay to systematically profile endogenous protein-protein interactions in a pathway network. *J Proteome Res.* (2014) 13:5339–46. doi: 10.1021/pr5002737
45. Wang Z, Liu C-H, Huang S, Chen J. Wnt signaling in vascular eye diseases. *Progr Retinal Eye Res.* (2019) 70:110–33. doi: 10.1016/j.preteyeres.2018.11.008
46. Gong Y, Slee RB, Fukai N, Rawadi G, Roman-Roman S, Reginato AM, et al. LDL receptor-related protein 5 (LRP5) affects bone accrual and eye development. *Cell.* (2001) 107:513–23. doi: 10.1016/S0092-8674(01)00571-2
47. Xu Q, Wang Y, Dabdoub A, Smallwood PM, Williams J, Woods C, et al. Vascular development in the retina and inner ear: control by Norrin and Frizzled-4, a high-affinity ligand-receptor pair. *Cell.* (2004) 116:883–95. doi: 10.1016/s0092-8674(04)00216-8
48. Fuhrmann S. Wnt signaling in eye organogenesis. *Organogenesis.* (2008) 4:60–7. doi: 10.4161/org.4.2.5850
49. Wu W-C, Drenser K, Trese M, Capone A, Dailey W. Retinal phenotype-genotype correlation of pediatric patients expressing mutations in the Norrie disease gene. *Archiv Ophthalmol.* (2007) 125:225–30. doi: 10.1001/archophth.125.2.225
50. Dhinra S, Shears DJ, Blake V, Stewart H, Patel CK. Advanced bilateral persistent fetal vasculature. *Am J Ophthalmol.* (1981) 91:312–7.
51. Aponte EP, Pulido JS, Ellison JW, Quiram PA, Mohny BG. A novel NDP mutation in an infant with unilateral persistent fetal vasculature and retinal vasculopathy. *Ophthalmic Genet.* (2009) 30:99–102. doi: 10.1080/13816810802705755
52. Prasov L, Masud T, Khaliq S, Mehdi SQ, Abid A, Oliver ER, et al. ATOH7 mutations cause autosomal recessive persistent hyperplasia of the primary vitreous. *Hum Mol Genet.* (2012) 21:3681–94. doi: 10.1093/hmg/dd5197
53. Shastri BS, Hiraoka M. Molecular genetics of familial exudative vitreoretinopathy and Norrie disease. *Curr Genomics.* (2000) 1:259–69. doi: 10.2174/1389202003351445
54. Lang RA. Apoptosis in mammalian eye development: lens morphogenesis, vascular regression and immune privilege. *Cell Death Different.* (1997) 4:12–20. doi: 10.1038/sj.cdd.44.00211
55. Martin AC, Thornton JD, Liu J, Wang X, Zuo J, Jablonski MM, et al. Pathogenesis of persistent hyperplastic primary vitreous in mice lacking the arf tumor suppressor gene. *Investig Ophthalmol Visual Sci.* (2004) 45:3387–96. doi: 10.1167/iops.04-0349
56. Thornton JD, Swanson DJ, Mary MN, Pei D, Martin AC, Pounds S, et al. Persistent hyperplastic primary vitreous due to somatic mosaic deletion of the arf tumor suppressor. *Investig Ophthalmol Visual Sci.* (2007) 48:491–9. doi: 10.1167/iops.06-0765
57. Zhang J, Fuhrmann S, Vetter ML. A nonautonomous role for retinal frizzled-5 in regulating hyaloid vitreous vasculature development. *Investig Ophthalmol Visual Sci.* (2008) 49:5561–7. doi: 10.1167/iops.08-2226
58. Brzezinski JA, Schulz SM, Crawford S, Wroblewski E, Brown NL, Glaser T. Math5 null mice have abnormal retinal and persistent hyaloid vasculatures. In: *Developmental Biology*. San Diego, CA: Academic Press Inc.; Elsevier Science (2003). p. 553.
59. Edwards MM, McLeod DS, Li R, Grebe R, Bhutto I, Mu X, et al. The deletion of Math5 disrupts retinal blood vessel and glial development in mice. *Exp Eye Res.* (2012) 96:147–56. doi: 10.1016/j.exer.2011.12.005
60. Son AI, Sheleg M, Cooper MA, Sun Y, Kleiman NJ, Zhou R. Formation of persistent hyperplastic primary vitreous in ephrin-A5^{-/-} mice. *Investig Ophthalmol Visual Sci.* (2014) 55:1594–606. doi: 10.1167/iops.13-12706
61. Gaudreault J, Fei D, Beyer JC, Ryan A, Rangell L, Shiu V, et al. Pharmacokinetics and retinal distribution of ranibizumab, a humanized antibody fragment directed against VEGF-A, following intravitreal administration in rabbits. *Retina.* (2007) 27:1260–6. doi: 10.1097/IAE.0b013e318134eecd
62. Bakri SJ, Snyder MR, Reid JM, Pulido JS, Ezzat MK, Singh RJ. Pharmacokinetics of intravitreal ranibizumab (Lucentis). *Ophthalmology.* (2007) 114:2179–82. doi: 10.1016/j.ophtha.2007.09.012
63. Klettner A, Roider J. Comparison of bevacizumab, ranibizumab, and pegaptanib *in vitro*: efficiency and possible additional pathways. *Investig Ophthalmol Visual Sci.* (2008) 49:4523–7. doi: 10.1167/iops.08-2055
64. Shukla D, Namperumalsamy P, Goldbaum M, Cunningham ET Jr. Pegaptanib sodium for ocular vascular disease. *Indian J Ophthalmol.* (2007) 55:427. doi: 10.4103/0301-4738.36476
65. Chong V. Ranibizumab for the treatment of wet AMD: a summary of real-world studies. *Eye.* (2016) 30:270–86. doi: 10.1038/eye.2015.217
66. Miyake M, Yamashiro K, Akagi-Kurashige Y, Kumagai K, Nakata I, Nakanishi H, et al. Vascular endothelial growth factor gene and the response to anti-vascular endothelial growth factor treatment for choroidal neovascularization in high myopia. *Ophthalmology.* (2014) 121:225–33. doi: 10.1016/j.ophtha.2013.06.043
67. Hamano Y, Kalluri R. Tumorstatin, the NC1 domain of $\alpha 3$ chain of type IV collagen, is an endogenous inhibitor of pathological angiogenesis and suppresses tumor growth. *Biochem Biophys Res Commun.* (2005) 333:292–8. doi: 10.1016/j.bbrc.2005.05.130
68. Mujoo K, Cheresh DA, Yang HM, Reisfeld RA. Disialoganglioside GD2 on human neuroblastoma cells: target antigen for monoclonal antibody-mediated cytotoxicity and suppression of tumor growth. *Cancer Res.* (1987) 47:1098–104.
69. Barker E, Reisfeld RA. A mechanism for neutrophil-mediated lysis of human neuroblastoma cells. *Cancer Res.* (1993) 53:362–7.
70. Weiss O, Kaufman R, Mishani E, Inbal A. Ocular vessel patterning in zebrafish is indirectly regulated by Hedgehog signaling. *Int J Develop Biol.* (2017) 61:277–84. doi: 10.1387/ijdb.160273ai
71. Kool M, Jones DTW, Jäger N, Northcott PA, Pugh TJ, Hovestadt V, et al. Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. *Cancer Cell.* (2014) 25:393–405.
72. Jain S, Song R, Xie J. Sonidegib: mechanism of action, pharmacology, and clinical utility for advanced basal cell carcinomas. *Onco Targets Ther.* (2017) 10:1645. doi: 10.2147/OTT.S130910
73. Papermaster DS, Tilly JL. Apoptosis of tissues of the eye during development and disease. In: *When Cells Die: A Comprehensive Evaluation of Apoptosis and Programmed Cell Death*. New York, NY: Wiley-Liss (1998). p. 321–46.
74. Lang RA, Bishop JM. Macrophages are required for cell death and tissue remodeling in the developing mouse eye. *Cell.* (1993) 74:453–62. doi: 10.1016/0092-8674(93)80047-1
75. Balazs EA, Toth LZ, Ozanics V. Cytological studies on the developing vitreous as related to the hyaloid vessel system. *Albrecht von Graefes Archiv für klinische und experimentelle Ophthalmologie.* (1980) 213:71–85. doi: 10.1007/BF00413534
76. Feeney SA, Simpson DAC, Gardiner TA, Boyle C, Jamison P, Stitt AW. Role of vascular endothelial growth factor and placental growth factors during retinal vascular development and hyaloid regression. *Investig Ophthalmol Visual Sci.* (2003) 44:839–47. doi: 10.1167/iops.02-0040

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

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

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



Multi-Trait Genomic Risk Stratification for Type 2 Diabetes

Palle Duun Rohde^{1,2*}, Mette Nyegaard^{2,3}, Mads Kjolby^{3,4,5,6} and Peter Sørensen⁷

¹ Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark, ² Department of Health Science and Technology, Aalborg University, Aalborg, Denmark, ³ Department of Biomedicine, Aarhus University, Aarhus, Denmark, ⁴ Department of Population Health and Genomics, University of Dundee, Dundee, United Kingdom, ⁵ Department of Clinical Pharmacology, Aarhus University Hospital, Aarhus, Denmark, ⁶ Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark, ⁷ Centre for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark

OPEN ACCESS

Edited by:

Balu Kamaraj,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia

Reviewed by:

J. Francis Borgio,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia
Udhaya Kumar S,
Vellore Institute of Technology, India

*Correspondence:

Palle Duun Rohde
palledr@bio.aau.dk

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 18 May 2021

Accepted: 05 August 2021

Published: 08 September 2021

Citation:

Rohde PD, Nyegaard M, Kjolby M and
Sørensen P (2021) Multi-Trait
Genomic Risk Stratification for Type 2
Diabetes. *Front. Med.* 8:711208.
doi: 10.3389/fmed.2021.711208

Type 2 diabetes mellitus (T2DM) is continuously rising with more disease cases every year. T2DM is a chronic disease with many severe comorbidities and therefore remains a burden for the patient and the society. Disease prevention, early diagnosis, and stratified treatment are important elements in slowing down the increase in diabetes prevalence. T2DM has a substantial genetic component with an estimated heritability of 40–70%, and more than 500 genetic loci have been associated with T2DM. Because of the intrinsic genetic basis of T2DM, one tool for risk assessment is genome-wide genetic risk scores (GRS). Current GRS only account for a small proportion of the T2DM risk; thus, better methods are warranted for more accurate risk assessment. T2DM is correlated with several other diseases and complex traits, and incorporating this information by adjusting effect size of the included markers could improve risk prediction. The aim of this study was to develop multi-trait (MT)-GRS leveraging correlated information. We used phenotype and genotype information from the UK Biobank, and summary statistics from two independent T2DM studies. Marker effects for T2DM and seven correlated traits, namely, height, body mass index, pulse rate, diastolic and systolic blood pressure, smoking status, and information on current medication use, were estimated (i.e., by logistic and linear regression) within the UK Biobank. These summary statistics, together with the two independent training summary statistics, were incorporated into the MT-GRS prediction in different combinations. The prediction accuracy of the MT-GRS was improved by 12.5% compared to the single-trait GRS. Testing the MT-GRS strategy in two independent T2DM studies resulted in an elevated accuracy by 50–94%. Finally, combining the seven information traits with the two independent T2DM studies further increased the prediction accuracy by 34%. Across comparisons, body mass index and current medication use were the two traits that displayed the largest weights in construction of the MT-GRS. These results explicitly demonstrate the added benefit of leveraging correlated information when constructing genetic scores. In conclusion, constructing GRS not only based on the disease itself but incorporating genomic information from other correlated traits as well is strongly advisable for obtaining improved individual risk stratification.

Keywords: UK Biobank, genetic risk scores, GRS, multi-trait analysis, precision medicine

INTRODUCTION

Type 2 diabetes mellitus (T2DM) is a chronic disease with severe comorbidities, such as myocardial infarction, loss of kidney function, blindness, and risk of amputations (1). Globally, the prevalence of T2DM is expected to increase exponentially in developing countries (2, 3), and it is a disease that places a severe economic burden on health systems. Accurate disease risk assessment is important for early disease diagnosis for initiating lifestyle changes early in the disease progression or prompt the clinician to treat high-risk patients more aggressively, which is expected to slow down disease progression, reduce disease symptoms, and prevent severe morbidity and mortality. Thus, methods for accurate disease risk assessment are absolutely critical for reducing morbidity and mortality.

Studies have unambiguously shown that T2DM is a complex, multifactorial disease, where an individual's risk of developing the disease is influenced by a combination of genetic variation at multiple sites across the genome acting in concert with environmental factors (4–6). The heritability of T2DM has been estimated to be 40–70% (7, 8), and more than 500 distinct genetic loci have been implicated with T2DM risk (6, 9–12). As T2DM is greatly impacted by genetics, genomic information has the potential to not only aid with early disease diagnosis but importantly also to stratify patients across disease subtypes (13) to initiate treatment intervention and lifestyle changes early in the disease progression.

During the last decade, an enormous effort has been in method development and construction of disease risk scores based on genomic information (14–17). However, until recently, these genome-wide genetic risk scores (GRS) have mainly been constructed using a single-trait approach. Because much of the variation within the human genome contributes to a large number of different complex traits and diseases (18), the accuracy of risk stratification can be improved by developing multi-trait (MT)-GRS accounting for the genetic correlation *among* traits. Using correlated information to construct GRS has theoretically—and to a minor extend empirical—been shown to increase the accuracy of disease risk prediction (6–8). T2DM is strongly correlated with a range of complex diseases and traits, such as overweight (19), cardiovascular disease (1, 19–21), hypertension (19, 22), and chronic kidney disease (19, 23); hence, T2DM is an excellent case for developing accurate GRS by leveraging correlated information.

The objective of the current study was to investigate the predictive performance of a MT-GRS model that combines marker effects from genome-wide association studies (GWAS) of T2DM and a number of correlated traits. The types of information included in this study were body mass index (BMI), height, smoking status, pulse rate, diastolic and systolic blood pressure, and a quantity of current medication use, as the total count of different prescription and over-the-counter medications is a proxy for general health and disease status. The aim of the present study was to investigate whether a MT-GRS model based on loci for multiple correlated traits had increased predictive discriminative power compared with a traditional single-trait (ST)-GRS model. This strategy was first applied within the UK

Biobank (UKB) (24), and then extended to include information on two UKB-independent GWAS summary statistics and, finally, a combined model incorporating information from the UKB and the two independent T2DM GWAS data sets.

MATERIALS AND METHODS

Phenotype and Genotype Data

Only unrelated British Caucasian individuals from the UKB (24) ($n = 335,652$ subjects) were used in the current study (excluding individuals with more than 5,000 missing genotype values or if having chromosomal aneuploidy). T2DM status was determined based on in-hospital records (by ICD-10 E.11, UKB data field 41270, which contains both main and secondary diagnoses) and self-reported disease state (UKB data field 20002) counting a total of 18,809 individuals. Seven additional phenotypes were also included: standing height, BMI, diastolic and systolic blood pressure, pulse rate, smoking status, and current medication use (measured as the number of different prescription and over-the-counter medications taken). These phenotypes were all adjusted for sex, age, UKB assessment center, and the first 10 genetic principal components (to account for any cryptic relatedness that were not accounted for by restricting to unrelated Caucasian British individuals), following inverse rank normalization to approximate normality.

Genotyped variants with minor allele frequency <0.01 , genotype missingness $>5\%$, or variants within the major histocompatibility complex were excluded from the analyses, resulting in a total of 599,297 genetic variants.

Prediction of Diabetes Risk

T2DM risk was determined using GRS based on either summary statistics obtained within the UKB cohort and other T2DM-related GWAS studies (Table 1). The overall workflow is depicted in Figure 1 and is described in detail below.

UKB Summary Statistics

The White-British UKB cohort of unrelated individuals (335,652 subjects) was split into 10 folds with no overlap of samples within each fold, and for each fold, the marker effects for T2DM, standing height, BMI, diastolic and systolic blood pressure, pulse rate, smoking status, and current medication use, were estimated using logistic or linear regression as implemented in PLINK2 (26). In all analyses, the same set of covariates were included as those used during phenotypic adjustment as this has been shown to increase statistical power (27).

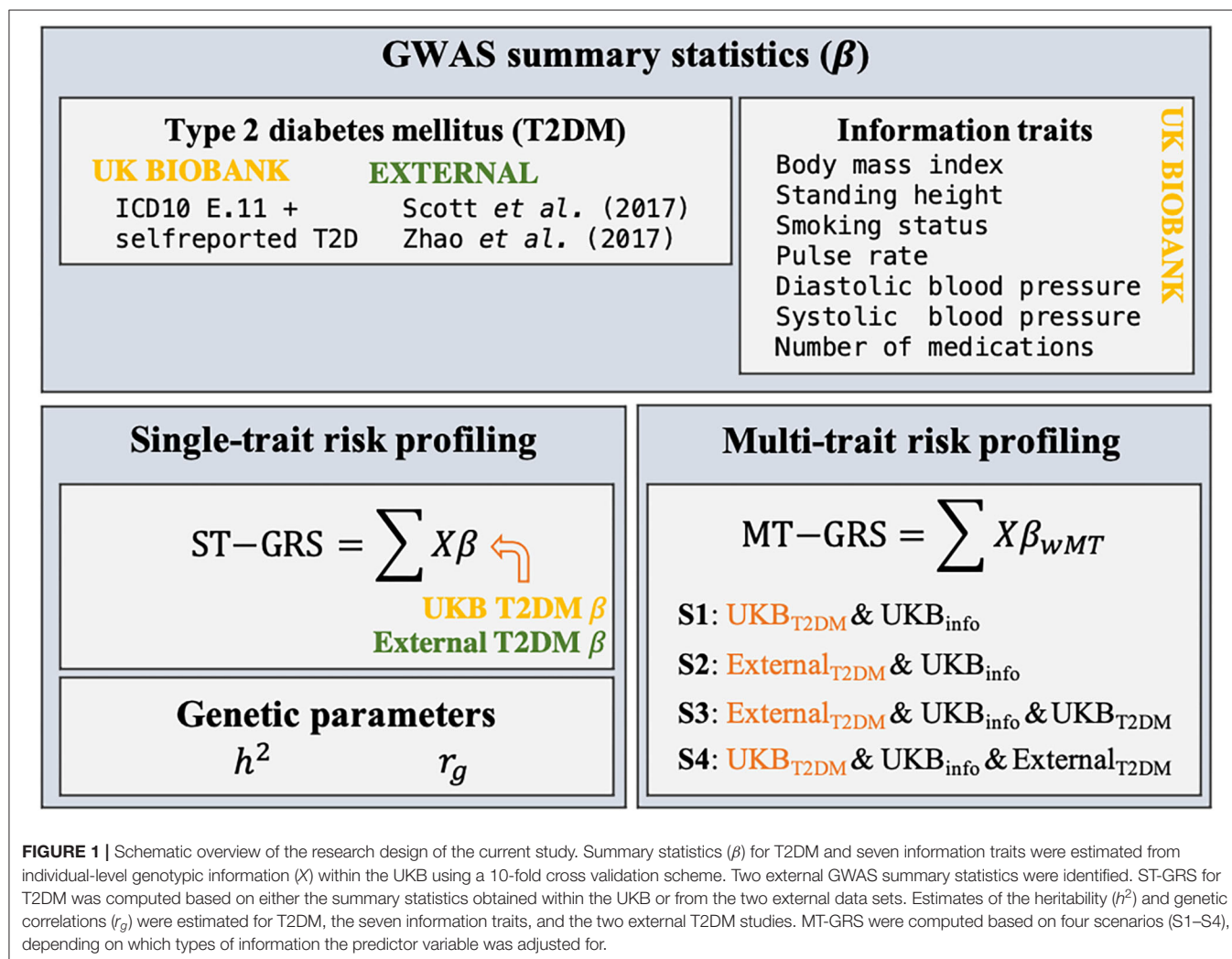
Publicly Available Type 2 Diabetes Summary Statistics

Two recently published GWAS for T2DM were identified (Table 1). Common for the studies were that they did not include UKB data, and therefore provide an independent training set. The regression coefficients were flipped such that the marker effect of the effect allele matched the effect allele within the UKB data.

TABLE 1 | Type 2 diabetes studies with available GWAS summary statistics independent of UKB.

Study	References	n_{total}	n_{case}	m_{total}	m_{UKB}
Scott et al. (2017)	(10)	159,208	26,676	12,056,346	595,528
Zhao et al. (2017)	(25)	265,678	73,337	8,796,184	558,105

n_{total} is the sample size of the listed study. n_{case} is the number of T2DM cases within the listed study. m_{total} is the number of genetic variants used in the GWAS of the listed study. m_{UKB} is the number of variants in the listed study that were among the 599,297 quality-controlled genotyped variants in the UKB.



Estimation of Genetic Parameters

Linkage disequilibrium (LD) between the genotyped variants was estimated as the squared Pearson's correlation coefficient (r^2) between two genetic variants adjusted for sample size (N) as the standard estimator of the Pearson's correlation coefficient has an upward bias (28). The adjusted squared Pearson's correlation coefficient (\tilde{r}^2) is obtained as (28):

$$\tilde{r}^2 = r^2 - \frac{1 - r^2}{N - 2}, \quad (1)$$

which was computed with the R package qgg (29). LD scores (l) for all variants within a window size of 5,000

markers (2,500 markers around the i -th variant) were computed as

$$l_i = \sum_{k=1}^{m=5000} \tilde{r}_{i,k}^2. \quad (2)$$

The MT-GRS model relies on selection index theory to obtain marker weights that require estimates of genetic parameters (30). The heritability (h^2) and the genetic correlation (r_g) between traits can be computed based on GWAS summary statistics using LD score regression (28). The heritability was estimated as the regression of the summary statistics on the LD score:

TABLE 2 | UKB cohort description ($n = 335,652$) of T2DM cases and controls (count (%) or mean \pm standard deviation).

Characteristics	Controls	T2DM cases	Information trait
N	316,935	18,809	
Age (years)	56.4 \pm 8.0	60.5 \pm 6.7	
Sex, male	144,070 (45.5%)	11,693 (62.2%)	
BMI (kg/m ²)	27.1 \pm 4.5	31.9 \pm 5.8	X
Height (cm)	168.8 \pm 9.2	170.0 \pm 9.3	X
Pulse rate (BPM)	69.1 \pm 11.1	73.6 \pm 13.1	X
Systolic blood pressure (mmHg)	138.0 \pm 18.6	142.6 \pm 18.0	X
Diastolic blood pressure (mmHg)	82.3 \pm 10.1	82.3 \pm 10.3	X
Smoking status			X
Never	175,002 (55.4%)	7,687 (41.1%)	
Former	109,007 (34.5%)	8,663 (46.3%)	
Current	31,867 (10.1%)	2,345 (12.6%)	
Number of medications	2.3 \pm 2.4	5.7 \pm 3.7	X

$$\hat{h}^2 = (\mathbf{Z}'\mathbf{Z})^{-1} (\mathbf{Z}'\mathbf{y}), \quad (3)$$

where $\mathbf{Z} = \mathbf{n}_{eff} \times l/m$, with l being the LD score (see Equation 2), m is the number of genetic variants, and \mathbf{n}_{eff} is the effective number of individuals and is $\mathbf{n}_{eff} = \text{median}\left(1/2 \times af \times (1 - af) \times SE(\hat{\mathbf{b}})^2\right)$, where af is the allele frequency, and $SE(\hat{\mathbf{b}})$ is the estimated standard error of the marker regression estimate. The response variable is $\mathbf{y} = \left(\frac{\hat{\mathbf{b}}}{SE(\hat{\mathbf{b}})}\right)^2$, where $\hat{\mathbf{b}}$ is the estimated regression coefficient for the genetic variants [for binary traits, the odds ratios (ORs) were converted to $\hat{\mathbf{b}} = \log(OR)$, and $SE(\hat{\mathbf{b}}) = \left|\hat{\mathbf{b}}/P(X < (1 - p)/2)\right|$, where $P(X < (1 - p)/2)$ is the normal cumulative distribution given the marker P -value, p (31)]. Similarly, the genetic correlation between traits 1 and 2 can be estimated as:

$$\hat{r}_g = \frac{(\mathbf{Z}'\mathbf{Z})^{-1} (\mathbf{Z}'\mathbf{y})}{\sqrt{\hat{h}_1^2} \sqrt{\hat{h}_2^2}}, \quad (4)$$

where $\mathbf{Z} = \sqrt{n_1} \sqrt{n_2} \times \frac{l}{m}$, and $\mathbf{y} = \frac{\hat{\mathbf{b}}_1}{SE(\hat{\mathbf{b}}_1)} \times \frac{\hat{\mathbf{b}}_2}{SE(\hat{\mathbf{b}}_2)}$. LD score regression was implemented in the R package qgg (29) and was computed for each of the 10-folds of random data subdivisions for T2DM and the seven information traits (Table 2), and among the information traits and the publicly available T2DM summary statistics (Table 1).

ST-GRS

The ST-GRS was computed as,

$$\text{ST-GRS} = \sum_{i=1}^m \mathbf{x}_i \hat{\mathbf{b}}_i, \quad (5)$$

where \mathbf{x}_i is the i -th column of the genotype matrix containing allelic counts, $\hat{\mathbf{b}}_i$ is the estimated marker effect for the i -th marker, and m is the number of variants left after LD pruning ($r^2 < 0.1$, <0.5 , or <0.9) and P -value thresholding ($P < 0.001$, 0.01 , 0.05 , 0.1 , 0.2 , 0.3 , 0.5 , 0.7 , 0.9 , and 0.99). The genetic scoring was performed with the R package qgg (29).

MT-GRS

The accuracy of GRS can be improved by leveraging information from correlated traits by adjusting the marker effects ($\hat{\mathbf{b}}$) (30). The adjustment of the marker effects for the focal trait (f , i.e., T2DM) is obtained by computing index weights for each marker (\mathbf{w}_i')

$$\hat{\mathbf{b}}_{wMT_i} = \mathbf{w}_i' \hat{\mathbf{b}}_i. \quad (6)$$

From quantitative genetic theory, selection indices have been developed for MT selection, in which many ST individual genetic effects (i.e., breeding values) are combined with an index weight allowing selection of the individuals with the best MT phenotype (32, 33). The optimal weights can be derived as $\mathbf{w} = \mathbf{V}^{-1} \mathbf{C}$, where \mathbf{C} is a $k \times 1$ column vector of covariances between the $\hat{\mathbf{b}}$ values of the k traits and the true marker effects of the focal trait (\mathbf{b}_f), and \mathbf{V} is a $k \times k$ variance-covariance matrix of the $\hat{\mathbf{b}}$ values:

$$\mathbf{w} = \begin{bmatrix} \text{var}(\hat{\mathbf{b}}_1) & \dots & \text{cov}(\hat{\mathbf{b}}_1, \hat{\mathbf{b}}_k) \\ \vdots & \ddots & \vdots \\ \text{cov}(\hat{\mathbf{b}}_k, \hat{\mathbf{b}}_1) & \dots & \text{var}(\hat{\mathbf{b}}_k) \end{bmatrix}^{-1} \begin{bmatrix} \text{cov}(\mathbf{b}_f, \hat{\mathbf{b}}_1) \\ \dots \\ \text{cov}(\mathbf{b}_f, \hat{\mathbf{b}}_k) \end{bmatrix}. \quad (7)$$

The diagonal elements of variance-covariance matrix, \mathbf{V} , are

$$\text{var}(\hat{\mathbf{b}}_k) = \frac{h_k^2}{M} + \frac{1}{N_k}, \quad (8)$$

where M is the effective number of chromosomal segments [here $M = 60,000$ (30, 34)] and N_k is the number of observations for trait k . The off-diagonal elements of \mathbf{V} for trait k and l are

$$\text{cov}(\hat{\mathbf{b}}_k, \hat{\mathbf{b}}_l) = \frac{r_g h_k h_l}{M}, \quad (9)$$

which is the same for the elements of \mathbf{C} . Combining Equations (8) and (9), Equation (7) becomes

$$\mathbf{w} = \begin{bmatrix} \frac{h_1^2}{M} + \frac{1}{N_1} & \dots & \frac{r_g h_1 h_k}{M} \\ \vdots & \ddots & \vdots \\ \frac{r_g h_k h_1}{M} & \dots & \frac{h_k^2}{M} + \frac{1}{N_k} \end{bmatrix}^{-1} \begin{bmatrix} \frac{h_1^2}{M} \\ \dots \\ \frac{r_g h_1 h_k}{M} \end{bmatrix}. \quad (10)$$

The MT-GRS is then obtained as the sum of adjusted marker effects,

$$\text{MT-GRS} = \sum_{i=1}^m \mathbf{x}_i \hat{\mathbf{b}}_{wMT_i}. \quad (11)$$

MT-GRS was computed by applying LD pruning ($r^2 < 0.1$, <0.5 , or <0.9) and P -value thresholding ($P < 0.001$, 0.01 , 0.05 , 0.1 ,

0.2, 0.5, 0.75, and 0.99) based on UKB genotypes and T2DM summary statistics; thus, the same LD pruning and P -value thresholding were applied across traits.

Four MT scenarios were applied, resulting in four different predictors (**Figure 1**): (1) UKB T2DM summary statistics combined with the seven UKB information traits; (2) external T2DM summary statistics [i.e., results from Scott et al. (10) and Zhao et al. (25)] combined with the seven UKB information traits; (3) external T2DM summary statistics combined with the seven UKB information traits and UKB T2DM summary statistics; and (4) UKB T2DM summary statistics combined with the seven UKB information traits and the two external T2DM summary statistics.

GRS Accuracy

The accuracy of ST-GRS and MT-GRS was determined using Nagelkerke's variance explained (R^2),

$$R^2 = \frac{1 - e^{-LR/n}}{1 - e^{-(2L_0)/n}} \quad (12)$$

where LR is the likelihood ratio comparing two nested logistic regression models, L_0 is the log-likelihood of a model neglecting the GRS, and n is the number of observations. The full model included sex, age, UKB assessment center, the first 10 genetic principal components, and the GRS, whereas the reduced model did not contain the GRS effect. For visualization, the GRS were divided into percentiles, and the disease prevalence within each bin was computed; the OR for each percentile was computed adjusting for sex, age, UKB assessment center, and the first 10 genetic principal components, and the OR was expressed relative to the 50-th percentile.

RESULTS

ST Prediction and Genetic Parameters

The analysis of T2DM was performed using 335,662 unrelated individuals from UKB with more than 18,000 T2DM cases (**Table 2**). A larger proportion of T2DM cases were males and smokers; on average, T2DM cases were older than individuals without T2DM, had higher BMI, and on average used more medications than non-diabetic individuals (**Table 2**).

The UKB cohort was split into 10 training and validation sets, and within-cohort marginal marker effects of common genotyped variants were estimated for each training set. After LD pruning and P -value thresholding, ST-GRS were computed for individuals within the validation sets. The maximum prediction accuracy for ST-GRS was $R^2 = 0.032$ when using variants with LD $r^2 < 0.9$ and $P < 0.05$ (**Figure 2**; **Supplementary Table 2**).

Across the 10 training sets, the average heritability for T2DM on the observed scale was 0.07 (0.31 on the liability scale). Seven information traits were included and used in the MT genetic risk scoring (**Table 2**). All seven traits showed non-zero heritability estimates (**Figure 3A**), and the strongest genetic correlation was observed between diastolic and systolic blood pressure (**Figure 3B**). Current medication use was the trait that showed the highest genetic correlation to most of the other traits,

and only standing height showed negative genetic correlation to the other traits (**Figure 3B**).

Leveraging Correlated Information for MT Prediction

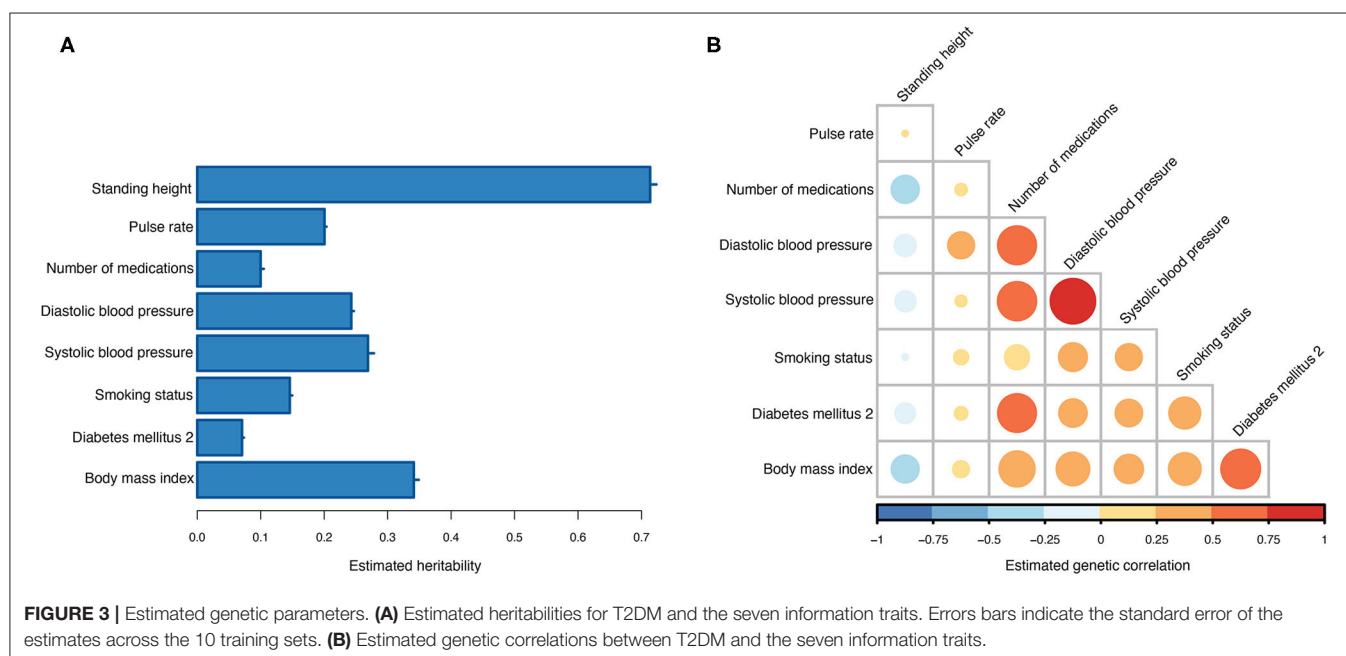
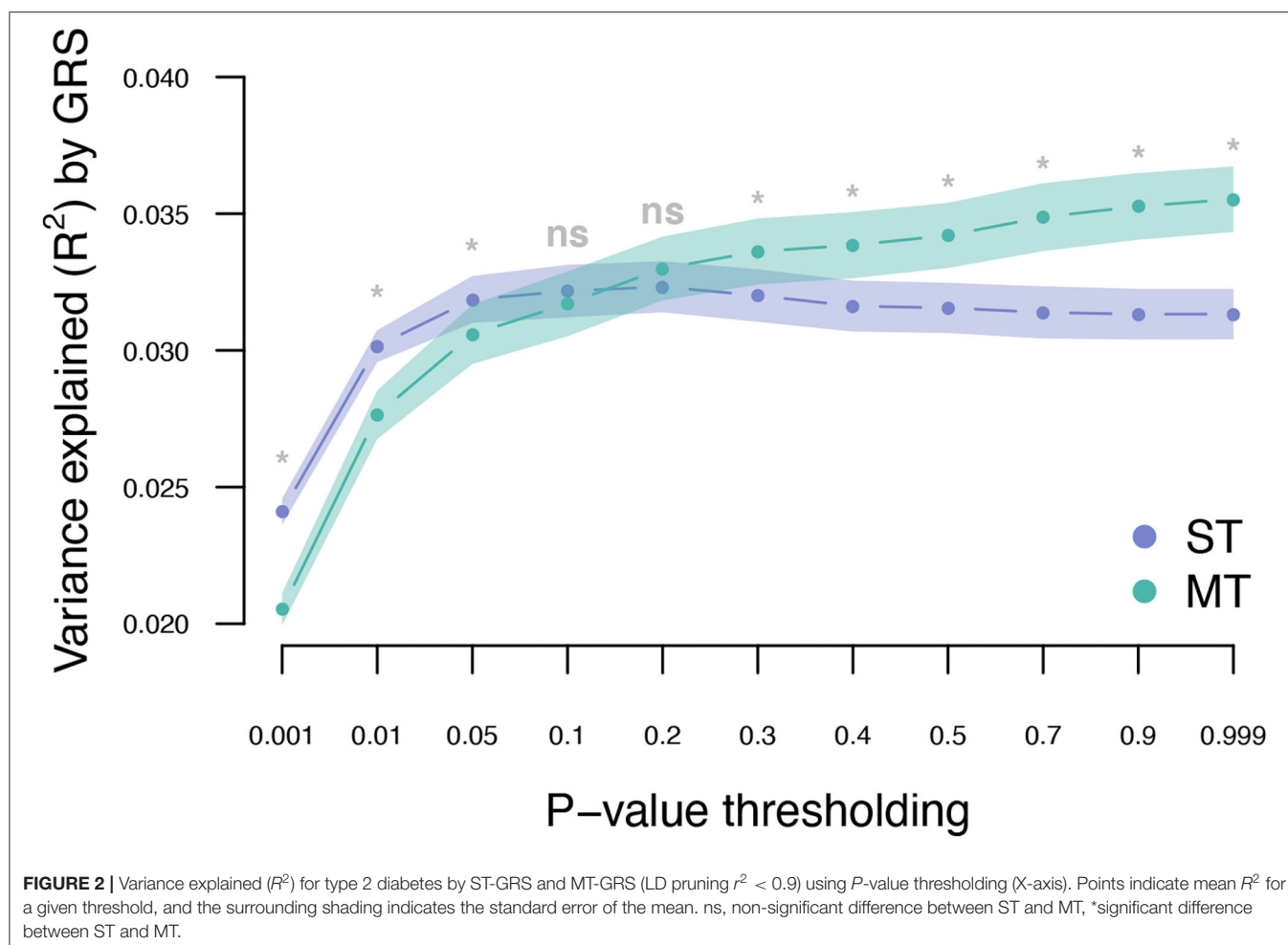
The T2DM marginal effects were adjusted using the estimated genetic parameters to compute MT-GRS (Scenario 1; **Figure 1**). Across the three levels of LD pruning, the predictive ability was generally improved when the marginal SNP effects were adjusted by the seven information traits (**Supplementary Figure 1**; **Supplementary Table 2**). The highest prediction accuracy ($R^2 = 0.036$) was obtained at LD $r^2 < 0.9$ and $P < 0.999$ (**Figure 2**; **Supplementary Table 2**), which corresponds to an improved prediction accuracy by 12.5%.

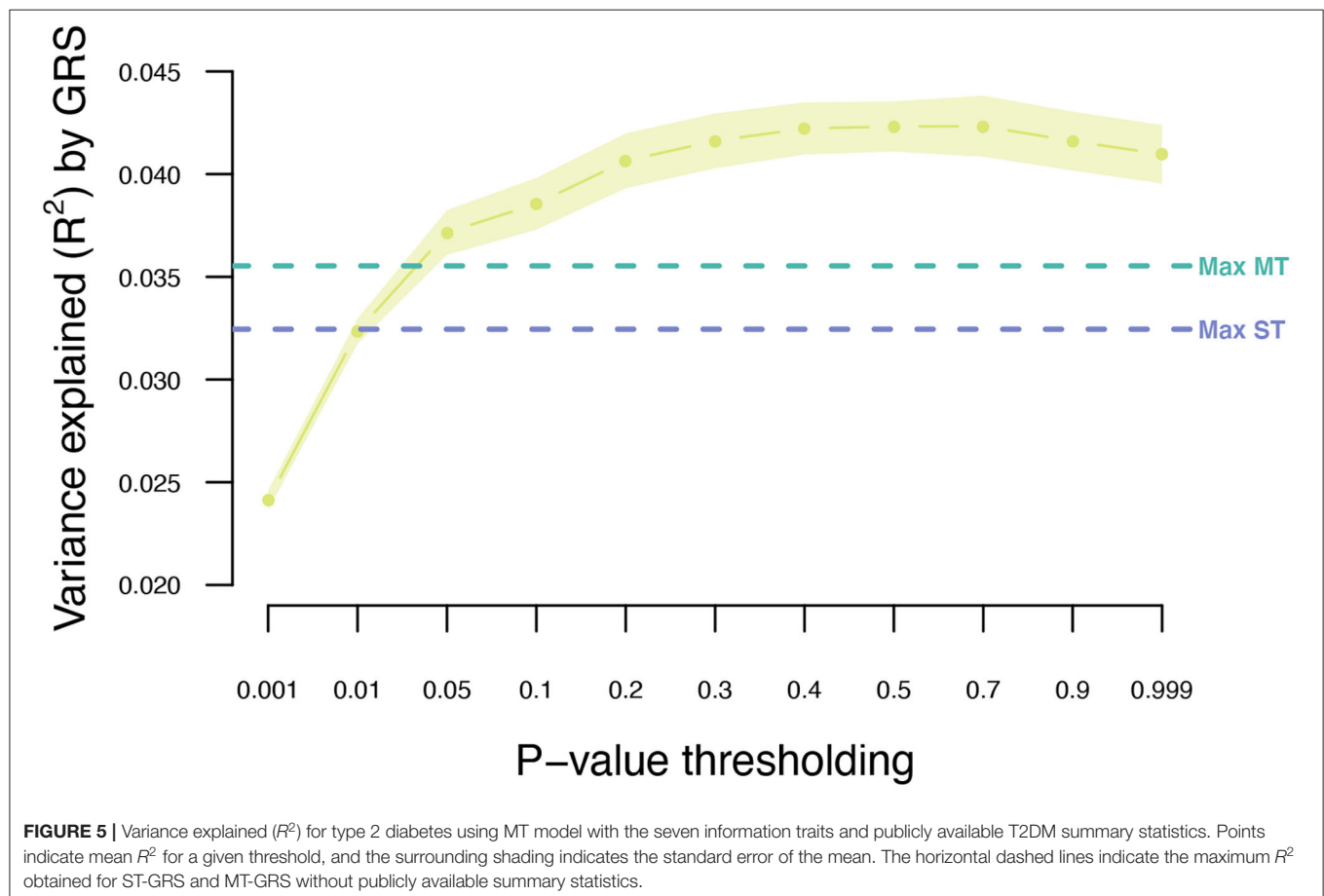
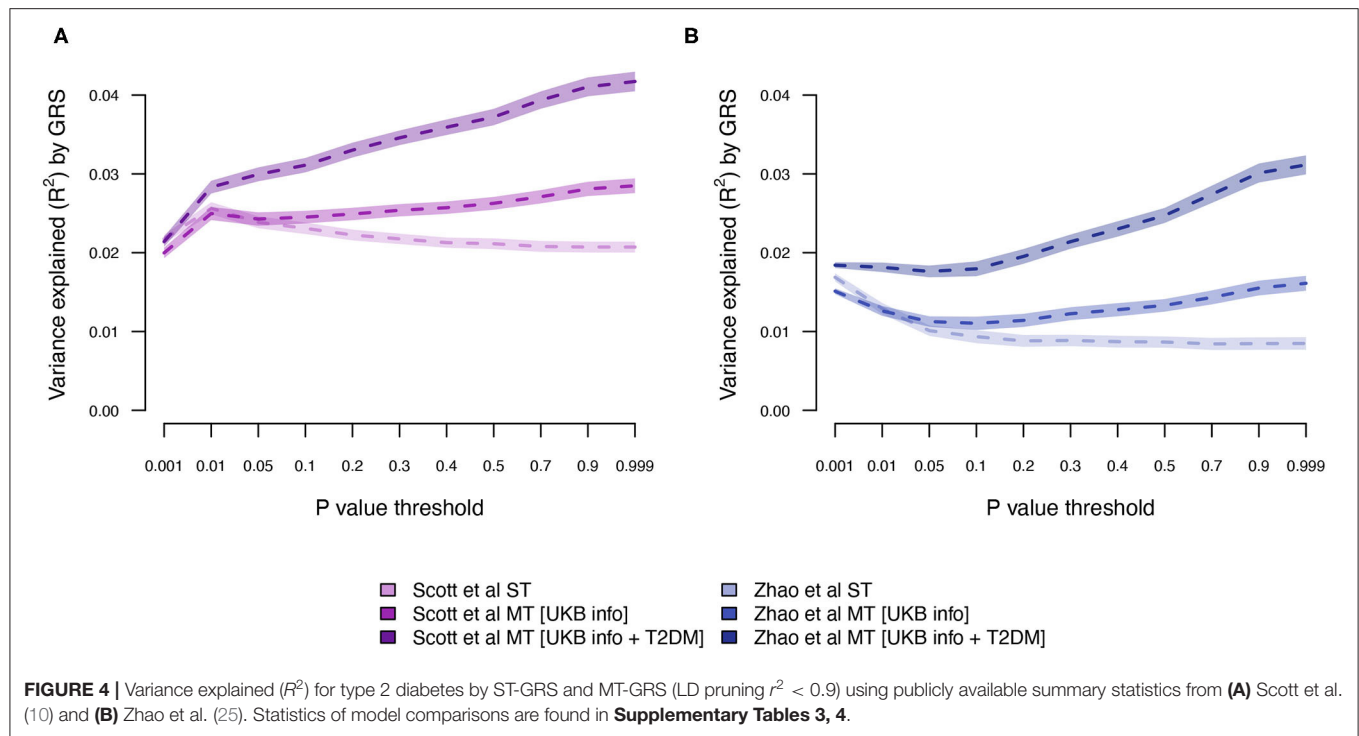
Next, we estimated the T2DM risk within the UKB using summary statistics from two independent external sets of summary statistics (**Figure 1**). Both external data sets [Scott et al. (10) and Zhao et al. (25)] showed low prediction accuracy when the GRS solely were computed using T2DM summary statistics [Scott et al. (10): $R^2 = 0.026$ at LD $r^2 = 0.9$ and $P < 0.01$; and Zhao et al. (25): $R^2 = 0.017$ at LD $r^2 = 0.9$ and $P < 0.001$; **Figure 4**; **Supplementary Tables 3, 4**; **Supplementary Figure 2**]. The external T2DM summary statistics were adjusted using summary statistics from the seven information traits obtained from the UKB (Scenario 2; **Figure 1**; **Supplementary Table 1**; **Supplementary Figure 3**), which for the summary statistics from Scott et al. (10) increased the prediction accuracy by 8%, but for Zhao et al. (25), a marginal drop in accuracy was observed when comparing the local maximum for ST-GRS with the local maximum for MT-GRS [$R^2 = 0.017$ ($r^2 = 0.9$, $P < 0.001$) vs. 0.016 [$R^2 = 0.016$ ($r^2 = 0.9$, $P < 0.999$); **Supplementary Table 4**]; however, comparing the accuracy within the P -value threshold, the accuracy of the MT-GRS model was superior over the ST (**Supplementary Table 4**). Extending the MT model to also include UKB T2DM summary statistics (Scenario 3, **Figure 1**), the accuracy was further increased by 50% (from 0.028 to 0.042; **Figure 4**) and 94% (from 0.016 to 0.031; **Figure 4**) using the summary statistics of Scott et al. (10) and Zhao et al. (25), respectively.

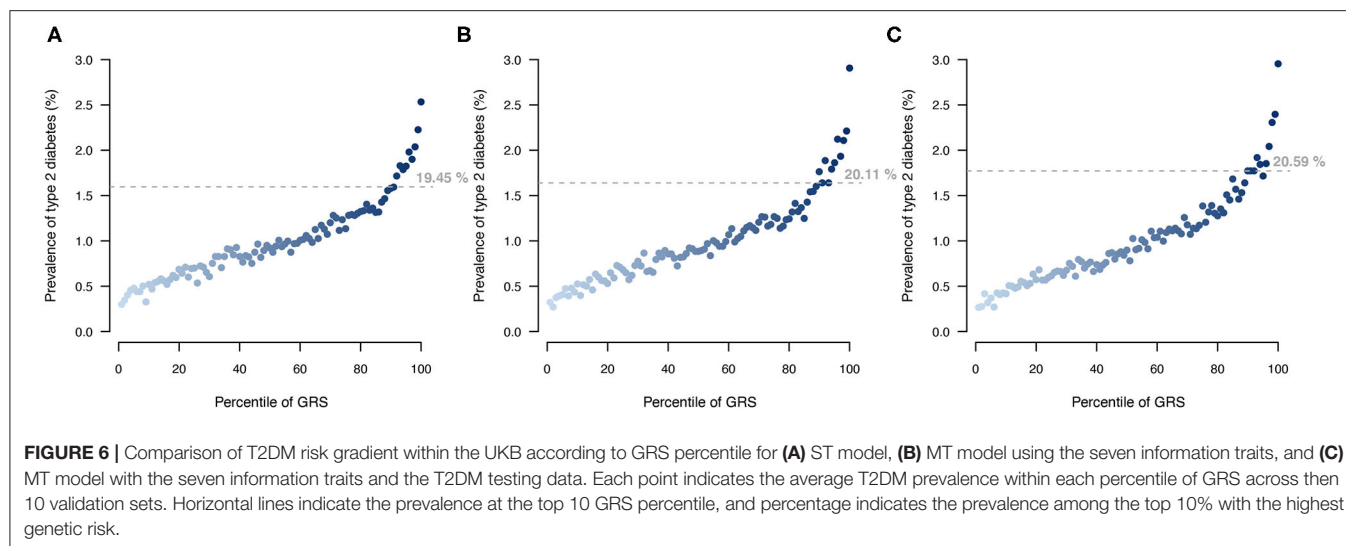
The MT model trained within the UKB was further extended to also include summary statistics from the two independent T2DM GWAS data sets (Scenario 4; **Figure 1**). Adjusting the UKB T2DM summary statistics by the seven information traits and the two independent T2DM GWAS data sets resulted in an increase in prediction accuracy from 0.032 to 0.043 (**Figure 5**; **Supplementary Table 2**), which is an increase of 34%.

T2DM Risk Stratification

Stratifying UKB participants based on their T2DM genetic risk showed that a larger proportion of individuals with a T2DM diagnosis were among the top 10% of individuals with highest genetic score when applying the MT strategy (**Figure 6**). The MT-GRS that in addition to the seven information traits also included information from the independent testing data gave a better stratification of cases by distributing a larger proportion of T2DM cases within the top risk (**Figure 6**), which also was apparent







with a large OR of the top 10% compared to the remaining (**Supplementary Figure 4**).

DISCUSSION

Precision medicine is predicted to change the way we prevent, diagnose, risk stratify individuals, and treat medical conditions (35, 36) through development of targeted preventive or treatment approaches based on the genetic background, biomarkers, environmental exposures, and lifestyle of the individual. Diagnosis and treatment plans based on genetic testing has been effectively applied to several monogenic disorders (37); however, for common complex diseases, genomic information has been far less incorporated. One reason for the lack of incorporating genomic information in disease prevention and diagnosis for complex diseases is because a large proportion of the underlying genetic variation remains unexplained (38, 39). In the current study, we investigated whether an MT-GRS approach provided more accurate risk stratification than traditional ST genetic scoring approaches.

Adjusting the UKB T2DM marker effects by the genomic correlation of the seven information traits increased the prediction accuracy from $R^2 = 0.032$ to 0.036, and further adjusted by the two UKB-independent T2DM studies increased the accuracy to $R^2 = 0.042$. The great improvement in prediction accuracy (31%) is achieved as a consequence of abundant genomic pleiotropy (18, 30) and the apparent genomic correlation with the selected traits. In comparison, Khera et al. (14) reported a prediction accuracy of ST-GRS of $R^2 = 0.028$ (14), and Maier et al. (30) obtained an accuracy of $R^2 < 0.01$ for both ST-GRS and MT-GRS (30). Although Maier et al. (30) showed increased prediction accuracy by combining the marker effects of selected traits (30), our reported prediction accuracies were greatly elevated compared with Maier et al. (30), most likely driven by differences in the included traits, and thereby in the optimal weights caused by differences in genomic correlation among the traits.

One of the information traits we included in the MT-GRS was the genetic liability to current medication use, which is the number of different medications the UKB participants have taken at the time of the verbal interview. Because most individuals that suffers from temporary or chronic diseases will undergo medical intervention and because of comorbidity many individuals will have multiple medical conditions, those individuals will be treated with a range of different medicines. Consequently, the total set of prescription and over-the-counter drugs is potentially an informative index of the current medical and health status of an individual. Wu et al. (40) performed genetic analysis of self-reported medication use within the UKB and found that categories of different types of medication were strongly genetically associated with a range of different diseases and traits (40). We found that the genetic correlation between T2DM and medication use was $r_g = 0.55$ (only the correlation between T2DM and BMI had higher estimate, $r_g = 0.58$). This is also evident by investigating the optimal weights (Equation 7), where BMI and medication use were the two information traits with the largest weights (**Supplementary Figure 5A**), besides T2DM itself. Including summary statistics from the two published T2DM association studies only marginally affected the optimal weights (**Supplementary Figure 5B**).

Although the exact level of prediction accuracy of T2DM was considerably lower when using external data from Zhao et al. (25) compared to data from Scott et al. (10) (**Figure 4**), the percentage increase when extending ST-GRS to the MT-GRS was higher for Zhao et al. (25) (82%) compared with Scott et al. (10) (62%), despite the much greater sample size by Zhao et al. (25) (**Table 1**). The discrepancy in prediction accuracy is most likely a consequence of different ancestries of the two external T2DM studies (10, 25), where the ancestry of the individuals in the study by Scott et al. (10) is more similar to the ancestry of the UKB (European) than the study by Zhao et al. (25) (mixed ancestry). It is well-established that across ancestry, risk prediction is very difficult because the LD between populations is very diverse (41–43).

The last decade has shown us that the sample size of human genetic association studies keeps increasing (44, 45), not only entailing more association signals but also providing more accurate effect estimates. This in conjunction with the increasingly accessibility of publicly available GWAS summary statistics (46, 47) implies that genomic prediction of complex diseases will continually improve, in particular if multivariate predictors are created by integrating information across studies. Although we have demonstrated increased prediction accuracy by constructing MT-GRS, our work has several limitations. Firstly, as our training data were the UKB and with a 10-fold cross-validation scheme, the number of cases became limited, meaning less accurate marker effect estimation and thereby less accurate risk stratification. Secondly, although we in addition to the UKB summary statistics from the 10-fold cross-validation obtained T2DM summary statistics from two independent studies (Table 1), we only had access to genotype information from the UKB and no other T2DM cohorts. Thirdly, we restricted the number of information traits to seven (Table 2), based on the criterion that it should be a type of information that is easy and accurate to measure and obtain; height, BMI, pulse rate, and diastolic and systolic blood pressure are things that we easily and accurately can measure, and smoking status and current medication use can easily be obtained by asking the participants. Accurate observations lead to more accurate estimation of marker effects and thereby better prediction accuracies. It is compelling to speculate whether other types of information traits would improve prediction accuracy even more, and additional studies are warranted for developing methods for identifying the set of information traits most important for a particular disease.

Genomic information has the potential to change the way we diagnose and treat individuals today and will be central for implementing preventive healthcare in the clinics. An important aspect of precision medicine is accurate prediction of genetic risk toward common diseases, as it may guide the general practitioners to better and earlier identify those individuals who have an inherent genetically lifetime high disease risk, and then to initiate lifestyle changes potentially before disease outcome. Moreover, precise stratification of T2DM patients not only based on their pathophysiological symptoms (13) but also on their genetic makeup may help the general practitioners to treat high-risk patients more aggressively, which has the potential to slow down disease progression, reduce symptoms, and prevent severe morbidity and mortality.

In conclusion, by incorporating information traits and two previously published T2DM GWAS results, the prediction accuracy for T2DM was increased by 31% (from $R^2 = 0.032$ to $R^2 = 0.042$), clearly demonstrating the added benefit of incorporating correlated information in the construction of GRS. Thus, incorporating genomic information on correlated traits and disease is advisable for obtaining improved individual genetic risk stratification.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: The genetic and phenotypic data were obtained from the UK Biobank Resource (ID 31269). Researchers can apply for access through: <https://www.ukbiobank.ac.uk/registerapply/>. Summary statistics for T2DM were obtained from published studies.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Ethics and Governance Framework (EGF) sets standards for the UK Biobank project so that all necessary safeguards are in place to ensure that the data and samples are only used for scientifically and ethically approved research. Participants of the UK Biobank have given their consent to participate which will apply throughout the lifetime of the UK Biobank unless the participants withdraw. Their consent involves the collection and storage of biological material (blood, saliva, urine samples) as well as collection of electronic health records (GP, hospitals, dental and prescription records). Information on the individual data level is anonymised for the researchers, and every research project has its own anonymised data. The ethics committee waived the requirement of written informed consent for participation.

AUTHOR CONTRIBUTIONS

PDR and PS conceived and designed the research project and performed the genetic analyses. PDR, PS, MN, and MK interpreted the results. All authors contributed to the preparation of the manuscript, read, edited, and approved the manuscript.

FUNDING

PDR has received funding from The Lundbeck Foundation (R287-2018-735).

ACKNOWLEDGMENTS

The data used in the presented study were obtained from the UKB Resource (project ID 31269). All of the computing for this project was performed on the GenomeDK HPC cluster. We would like to thank GenomeDK and Aarhus University for providing computational resources and support that contributed to these research results.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.711208/full#supplementary-material>

REFERENCES

1. The Emerging Risk Factors Collaboration, Sarwar N, Gao P, Kondapally Seshasai SR, Gobin R, Kaptoge S, et al. Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies. *The Lancet* (2010) 375:2215–22. doi: 10.1016/S0140-6736(10)60484-9
2. Guariguata L, Whiting DR, Hambleton I, Beagley J, Linnenkamp U, Shaw JE. Global estimates of diabetes prevalence for 2013 and projections for 2035. *Diabetes Res Clin Pract.* (2014) 103:137–49. doi: 10.1016/j.diabres.2013.11.002
3. Cho NH, Shaw JE, Karuranga S, Huang Y, da Rocha Fernandes JD, Ohlrogge AW, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res Clin Pract.* (2018) 138:271–81. doi: 10.1016/j.diabres.2018.02.023
4. Kolb H, Martin S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC Med.* (2017) 15:1–11. doi: 10.1186/s12916-017-0901-x
5. Flannick J, Florez JC. Type 2 diabetes: genetic data sharing to advance complex disease research. *Nat Rev Genet.* (2016) 17:535–49. doi: 10.1038/nrg.2016.56
6. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. The genetic architecture of type 2 diabetes. *Nature.* (2016) 536:41–7. doi: 10.1038/nature18642
7. Poulsen P, Ohm Kyvik K, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance - a population-based twin study. *Diabetologia.* (1999) 42:139–45. doi: 10.1007/s001250051131
8. Willemsen G, Ward KJ, Bell CG, Christensen K, Bowden J, Dalgård C, et al. The concordance and heritability of Type 2 diabetes in 34,166 twin pairs from international twin registers: The Discordant Twin (DISCOTWIN) Consortium. *Twin Res Hum Genet.* (2015) 18:762–71. doi: 10.1017/thg.2015.83
9. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè A V., Steinthorsdottir V, et al. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet.* (2012) 44:981–90. doi: 10.1038/ng.2383
10. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, et al. An expanded genome-wide association study of Type 2 diabetes in Europeans. *Diabetes.* (2017) 66:2888–902. doi: 10.2337/db16-1253
11. Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, Rayner NW, et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Genet.* (2018) 50:1505–13. doi: 10.1038/s41588-018-0241-6
12. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet.* (2020) 52:680–91. doi: 10.1101/19012690
13. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol.* (2018) 6:361–9. doi: 10.1016/S2213-8587(18)30051-2
14. Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet.* (2018) 50:1219–24. doi: 10.1038/s41588-018-0183-z
15. Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am J Hum Genet.* (2015) 97:576–92. doi: 10.1016/j.ajhg.2015.09.001
16. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc.* (2020) 15: doi: 10.1038/s41596-020-0353-1
17. Euesden J, Lewis CM, O'Reilly PF. PRSice: polygenic risk score software. *Bioinformatics.* (2015) 31:1466–8. doi: 10.1093/bioinformatics/btu848
18. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* (2015) 47:1236–41. doi: 10.1038/ng.3406
19. Iglay K, Hannachi H, Joseph Howie P, Xu J, Li X, Engel SS, et al. Prevalence and co-prevalence of comorbidities among patients with type 2 diabetes mellitus. *Curr Med Res Opin.* (2016) 32:1243–52. doi: 10.1185/03007795.2016.1168291
20. Goodarzi MO, Rotter JL. Genetics insights in the relationship between Type 2 diabetes and coronary heart disease. *Circ Res.* (2021) 126:1526–48. doi: 10.1161/CIRCRESAHA.119.316065
21. Danaei G, Lawes CM, Vander Hoorn S, Murray CJ, Ezzati M. Global and regional mortality from ischaemic heart disease and stroke attributable to higher-than-optimum blood glucose concentration: comparative risk assessment. *Lancet.* (2006) 368:1651–9. doi: 10.1016/S0140-6736(06)69700-6
22. Coresh J, Astor BC, Greene T, Eknoyan G, Levey AS. Prevalence of chronic kidney disease and decreased kidney function in the adult US population: third national health and nutrition examination survey. *Am J Kidney Dis.* (2003) 41:1–12. doi: 10.1053/ajkd.2003.50007
23. Dean J. Organising care for people with diabetes and renal disease. *J Ren Care.* (2012) 38:23–9. doi: 10.1111/j.1755-6686.2012.00272.x
24. Bycroft C, Elliott LT, Young A, Vukcevic D, Effingham M, Marchini J, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature.* (2018) 562:203–9. doi: 10.1038/s41586-018-0579-z
25. Zhao W, Rasheed A, Tikkanen E, Lee JJ, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. *Nat Genet.* (2017) 49:1450–7. doi: 10.1038/ng.3943
26. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* (2015) 4:1–16. doi: 10.1186/s13742-015-0047-8
27. Sofer T, Zheng X, Gogarten SM, Laurie CA, Grinde K, Shaffer JR, et al. A fully adjusted two-stage procedure for rank-normalization in genetic association studies. *Genet Epidemiol.* (2019) 43:263–75. doi: 10.1002/gepi.22188
28. Bulik-Sullivan B, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* (2015) 47:291–5. doi: 10.1038/ng.3211
29. Rohde PD, Fourie Sørensen I, Sørensen P. qqg: an R package for large-scale quantitative genetic analyses. *Bioinformatics.* (2020) 36:2614–5. doi: 10.1093/bioinformatics/btz955
30. Maier RM, Zhu Z, Lee SH, Trzaskowski M, Ruderfer DM, Stahl EA, et al. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun.* (2018) 9:989. doi: 10.1038/s41467-017-02769-6
31. Hu D, Wang C, O'Connor AM. A method of back-calculating the log odds ratio and standard error of the log odds ratio from the reported group-level risk of disease. *PLoS ONE.* (2020) 15:e0222690. doi: 10.1371/journal.pone.0222690
32. Hazel LN. The genetic basis for constructing selection indexes. *Genetics.* (1943) 28:476–90. doi: 10.1093/genetics/28.6.476
33. Wientjes YCJ, Bijma P, Veerkamp RF, Calus MPL. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. *Genetics.* (2016) 202:799–823. doi: 10.1534/genetics.115.183269
34. Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet.* (2011) 19:807–12. doi: 10.1038/ejhg.2011.39
35. Ginsburg GS, McCarthy JJ. Personalized medicine: revolutionizing drug discovery and patient care. *Trends Biotechnol.* (2001) 19:491–6. doi: 10.1016/S0167-7799(01)01814-5
36. Ashley EA. Towards precision medicine. *Nat Rev Genet.* (2016) 17:507–22. doi: 10.1038/nrg.2016.86
37. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet.* (2013) 14:415–26. doi: 10.1038/nrg3493
38. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet.* (2016) 17:392–406. doi: 10.1038/nrg.2016.27
39. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature.* (2009) 461:747–53. doi: 10.1038/nature08494

40. Wu Y, Byrne EM, Zheng Z, Kemper KE, Yengo L, Mallett AJ, et al. Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat Commun.* (2019) 10:1891. doi: 10.1038/s41467-019-09572-5
41. Kerminen S, Martin AR, Koskela J, Ruotsalainen SE, Havulinna AS, Surakka I, et al. Geographic variation and bias in the polygenic scores of complex diseases and traits in Finland. *Am J Hum Genet.* (2019) 104:1169–81. doi: 10.1016/j.ajhg.2019.05.001
42. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet.* (2017) 100:635–49. doi: 10.1016/j.ajhg.2017.03.004
43. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun.* (2019) 10:3328. doi: 10.1038/s41467-019-11112-0
44. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet.* (2012) 90:7–24. doi: 10.1016/j.ajhg.2011.11.029
45. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: biology, function, and translation. *Am J Hum Genet.* (2017) 101:5–22. doi: 10.1016/j.ajhg.2017.06.005
46. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet.* (2017) 18:117–27. doi: 10.1038/nrg.2016.142
47. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* (2019) 47:D1005–12. doi: 10.1093/nar/gky1120

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

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

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



OPEN ACCESS

Edited by:

Thirumal Kumar D.,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Ibrahim Jelaidan,
King Faisal Cardiac Center, King
Abdulaziz Medical City, Saudi Arabia
Prashantha Karunakar,
PES University, India

*Correspondence:

Ramu Elango
relango@kau.edu.sa
Osman O. Al-Radi
oradi@kau.edu.sa
Turki Alahmadi
tsalahmadi@kau.edu.sa

†These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 14 June 2021

Accepted: 10 August 2021

Published: 13 September 2021

Citation:

Alsafwani RS, Nasser KK, Shinawi T,
Banaganapalli B, ElSokary HA,
Zaher ZF, Shaik NA, Abdelmohsen G,
Al-Aama JY, Shapiro AJ, O. Al-Radi O,
Elango R and Alahmadi T (2021) Novel
MYO1D Missense Variant Identified
Through Whole Exome Sequencing
and Computational Biology Analysis
Expands the Spectrum of Causal
Genes of Laterality Defects.
Front. Med. 8:724826.
doi: 10.3389/fmed.2021.724826

Novel MYO1D Missense Variant Identified Through Whole Exome Sequencing and Computational Biology Analysis Expands the Spectrum of Causal Genes of Laterality Defects

Rabab Said Alsafwani^{1†}, Khalidah K. Nasser^{1,2†}, Thoraia Shinawi¹,
Babajan Banaganapalli^{2,3}, Hanan Abdelhalim ElSokary², Zhafer F. Zaher^{4,5},
Noor Ahmad Shaik^{2,3,6}, Gaser Abdelmohsen^{4,7}, Jumana Yousuf Al-Aama^{2,3},
Adam J. Shapiro⁸, Osman O. Al-Radi^{9*}, Ramu Elango^{2,3*} and Turki Alahmadi^{4,10*}

¹ Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, ² Princess Al-Jawhara Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia, ³ Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, ⁴ Department of Pediatrics, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, ⁵ Pediatric Cardiac Center of Excellence, King Abdulaziz University Hospital, King Abdulaziz University, Jeddah, Saudi Arabia, ⁶ Department of Genetics, Al Borg Medical Laboratories, Jeddah, Saudi Arabia, ⁷ Pediatric Cardiology Division, Department of Pediatrics, Cairo University, Kasr Al Ainy Faculty of Medicine, Cairo, Egypt, ⁸ Division of Pediatric Respiratory Medicine, McGill University Health Centre Research Institute, Montreal Children's Hospital, Montreal, QC, Canada, ⁹ Department of Surgery Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, ¹⁰ Pediatric Department, Faculty of Medicine in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia

Laterality defects (LDs) or asymmetrically positioned organs are a group of rare developmental disorders caused by environmental and/or genetic factors. However, the exact molecular pathophysiology of LD is not yet fully characterised. In this context, studying Arab population presents an ideal opportunity to discover the novel molecular basis of diseases owing to the high rate of consanguinity and genetic disorders. Therefore, in the present study, we studied the molecular basis of LD in Arab patients, using next-generation sequencing method. We discovered an extremely rare novel missense variant in MYO1D gene (Pro765Ser) presenting with visceral heterotaxy and left isomerism with polysplenia syndrome. The proband in this index family has inherited this homozygous variant from her heterozygous parents following the autosomal recessive pattern. This is the first report to show MYO1D genetic variant causing left-right axis defects in humans, besides previous known evidence from zebrafish, frog and Drosophila models. Moreover, our multilevel bioinformatics-based structural (protein variant structural modelling, divergence, and stability) analysis has suggested that Ser765 causes minor structural drifts and stability changes, potentially affecting the biophysical

and functional properties of *MYO1D* protein like calmodulin binding and microfilament motor activities. Functional bioinformatics analysis has shown that *MYO1D* is ubiquitously expressed across several human tissues and is reported to induce severe phenotypes in knockout mouse models. In conclusion, our findings show the expanded genetic spectrum of LD, which could potentially pave way for the novel drug target identification and development of personalised medicine for high-risk families.

Keywords: laterality defects, whole exome sequencing, microfilament, gene expression, variant

INTRODUCTION

Laterality defects (LDs) are a group of developmental diseases that affect internal organ positioning in the body. In general, human LDs can be divided into three categories: (1) *situs solitus* (SS) with normally expected organ arrangement; (2) *situs inversus* (SI) characterised by complete mirror image of organs; and (3) *situs ambiguus* (SA) with organ arrangement falling along a spectrum of various anomalies between SS and SI, including congenital heart defects (CHDs). Within SA, a subgroup of patients presents a severe and complex form of congenital heart disease, which is commonly known as *heterotaxy* (1). Defective left-right (LR) patterning of internal organs is associated with multiple congenital diseases affecting the cardiovascular system, kidneys, liver, and biliary tract (2, 3). According to the National Birth Defects Prevention Study (4), the estimated prevalence of LD is 1.1 per 10,000 in the United States. Despite the rare likelihood of LD, its incidence is expected to be higher among the Arab population due to their high rate of consanguinity and genetic disorders (5).

The aetiology of LD is complex and includes both environment (5–7) and genetic factors (8, 9). Disease-causing genetic variations are found in <20% of LD cases; and the remaining 80% of cases are due to unidentifiable causes (10, 11). Up to now, known LD genes were mostly associated with NODAL/TGF β signalling (*NODAL*, *CFC1*, *ACVR2B*, *LEFTYB*, *GDF1*, *TGFBR2*, and *FOXH1*), SHH signalling (*ZIC3* and *LZTFL1*), and monocilia function (*NPHP2*, *NPHP3*, *NPHP4*, *PKD2*, and *TTC8*) (10). Other genetic alterations associated with early cardiac development (*NKX2-5*, *CRELD1*, *MMP21*, and *PKD1L1*) were also implicated in LD development (10). The main functional roles of these genes were demonstrated in LR axis determination, controlling cardiac looping direction, nodal activity regulation in embryogenesis, protein interaction of primary cilia, and signalling involved in morphogenesis cascade (12–19). Hence, LDs occur in a variety of different diseases, affecting various cardiac, respiratory, and gastrointestinal organs, reflecting the complex genes involved in signalling pathways of organogenesis and ciliary function.

Genetic testing and molecular diagnostics are now regarded as an useful approach to discover molecular causes underlying the LD development. Whole-exome sequencing (WES) analysis proved to be a successful method to uncover novel candidate genes or novel variants in known candidate genes (10). To this end, there is an increasing need to study the rare developmental disorders across different ethnic populations

due to its potential in expanding the genetic spectrum of the disease. However, literature searches reveal sparse data on the Arab LD patients. We hypothesise that genetically investigating LD patients from a consanguineous Arab society will offer new insights into disease pathogenesis by identifying novel genes or novel variants in known genes, as demonstrated in other complex diseases (20). Therefore, the objective of the present study was to identify the genetic cause of LD in Arab patients, using WES and multilevel bioinformatics-based structural (protein variant structural modelling, divergence, and stability) and functional (gene expression and knockout mouse model) analysis approaches.

MATERIALS AND METHODS

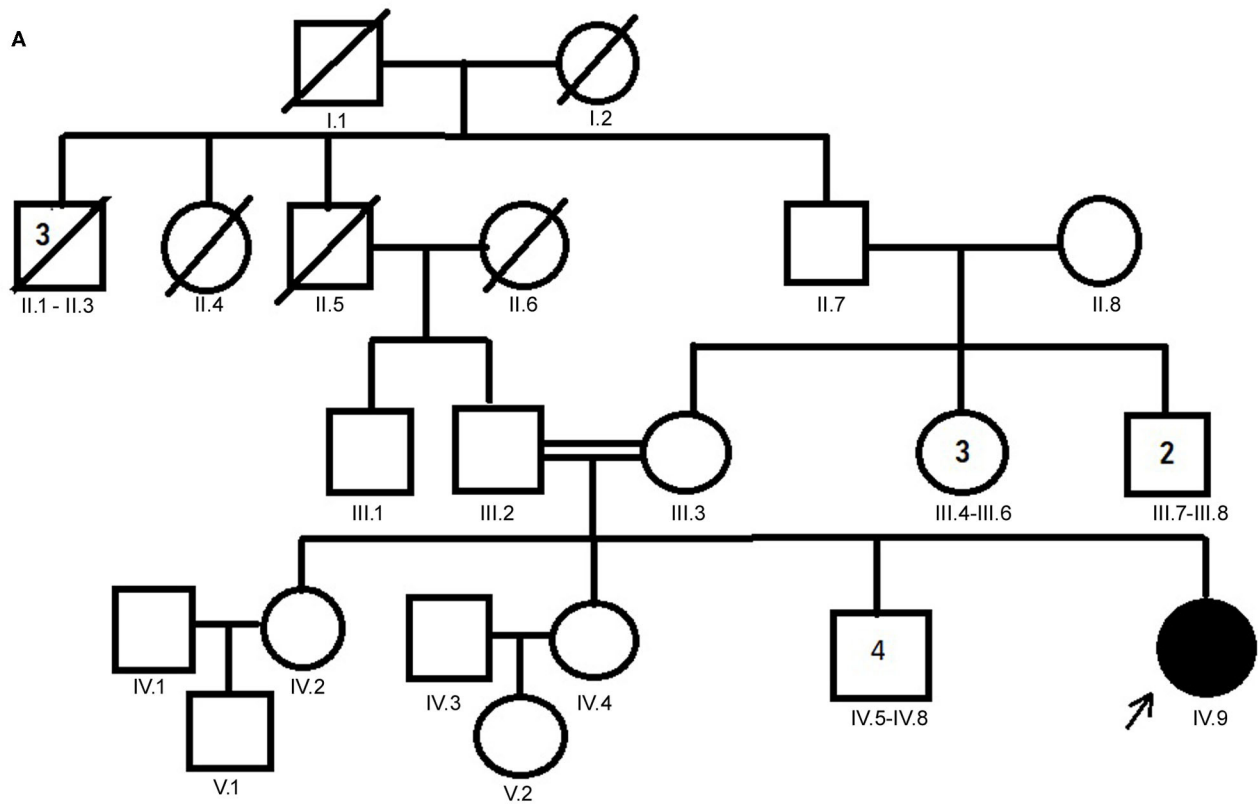
The ethical approval for the present study was obtained from the institutional ethics committee of King Abdulaziz University Hospital (KAUH), Jeddah, Saudi Arabia. Informed consent forms were collected from both adult parents and their children (parental consent and children's assent for those <18 years old) prior to blood sample collection and genetic testing. As per the National Birth Defects Prevention Study, LD participants were selected based on the following clinical criteria: *situs inversus*, CHDs (heterotaxy), isomerism of the lungs (bilateral two lobes/left-sidedness and bilateral three lobes/right-sidedness), abdominal situs abnormality (*abdominal situs inversus* and SA), and spleen abnormality (asplenia and polysplenia) (4). One LD index family composed of the proband (affected child) and both parents was recruited to paediatric cardiology, surgery, and pulmonology clinics in KAUH. Clinical, laboratory, and radiological results were independently assessed by both paediatric cardiology and pulmonology consultants. Family pedigree was drawn by interviewing the parents. Samples from two other LD families were screened for the presence of the identified candidate variants.

Molecular Testing Clinical Sampling

A total of 5 ml of whole EDTA peripheral blood samples was collected from each study participant.

DNA Extraction

The genomic DNA was isolated from circulating lymphocytes using QIAamp DNA blood Kit as per the manufacturer's protocol and quantified using a NanoDrop 2000 spectrophotometer. DNA integrity was checked on 2% agarose gel electrophoresis.



B

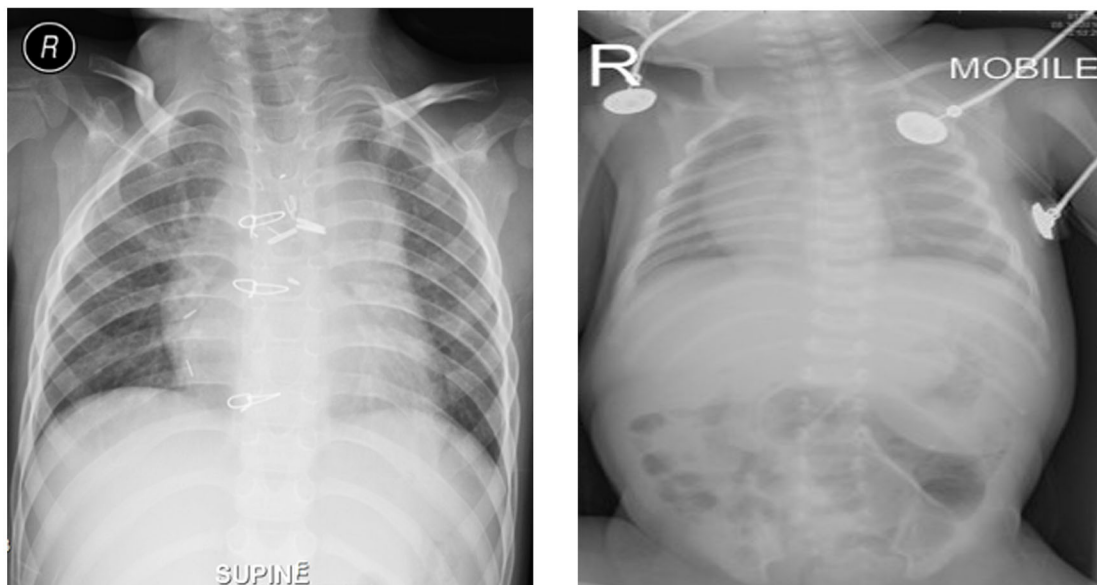


FIGURE 1 | (A) Family pedigree chart of laterality defect (LD) Arab family. Proband is indicated by the arrow. The affected proband (shaded circle) is homozygous of c.2293C>T mutation in *MYO1D* gene. Both parents were consanguineous (double horizontal lines) and heterozygous carriers of the identified mutation. **(B)** Chest X-rays showing left isomerism (heterotaxy).

TABLE 1 | List of LD potential candidate variants that showed autosomal recessive inheritance pattern.

	Gene name	Chrom	ref	Alt	Effect	HGVS.c	HGVS.p	dbSNP151_ID	1000G_AF	ExAC_AF	gnomAD_exomes_AF
(1)	ZC3H12A	chr1	G	T	Missense	c.99G>T	p.Arg33Ser	rs116208741	0.00139776	0.002405	0.002499959
(2)	ZC3H12A	chr1	C	T	Missense	c.95C>T	p.Pro32Leu	rs115805535	0.00119808	0.002117	0.002117
(3)	TMEM18A	chr7	C	T	Missense	c.116G>A	p.Gly39Glu	rs183593116	0.00219649	0.004772	0.00515989
(4)	CPNE7	chr16	C	T	Missense	c.898C>T	p.Pro300Ser	rs150443459	0.000399361	0.0002966	0.0003931566
(5)	ASXL2	chr2	C	T	Missense	c.1489G>A	p.Ala497Thr	rs192716734	0.00139776	0.003805	0.004480134
(6)	MYO1D	chr17	G	A	Missense	c.2293C>T	p.Pro765Ser	rs7209106	0.0043	0.002759	0.002424917
(7)	EPB42	chr15	C	T	Missense	c.1477G>A	p.Gly493Ser	rs148871144	0.000199681	0.0007578	0.0007836991
(8)	FAM220A	chr7	G	A	Missense	c.437C>T	p.Pro146Leu	rs75910050	0.00139776	0.0009884	0.0009261591
(9)	OR10A2	chr11	C	A	Missense	c.741C>A	p.Phe247Leu	rs150322658	0.00119808	0.0014	0.001421724

LD, laterality defect.

TABLE 2 | Computational pathogenicity prediction scores of the LD candidate variants.

S. no	Variant	Gene	CADD	FATHMM	MetaLR	MutationTaster	PROVEAN	REVEL
(1)	rs115805535	ZC3H12A	16.06	0.22678	0.0431	0.08975	0.30964	0.027
(2)	rs116208741	ZC3H12A	–	–	–	–	–	–
(3)	rs150443459	CPNE7	–	–	–	–	–	–
(4)	rs192716734	ASXL2	14.94	0.18248	0.025	0.25126	0.24026	0.079
(5)	rs148871144	EPB42	0.135	0.78537	0.1928	0.08975	0.23156	0.121
(6)	rs75910050	FAM220A	0.524	0.06931	0.0102	0.08975	0.58248	0.049
(7)	rs150322658	OR10A2	–	–	–	–	–	–
(8)	rs183593116	TMEM18A	9.855	0.43279	0.0355	0.08975	0.07008	0.039
(9)	rs7209106	MYO1D	23.2	0.88298	0.6557	0.81001	0.80682	0.553

LD, laterality defect.

CADD: >25 = damaging; <25 = tolerated; FATHMM: >0.5 = damaging; <0.5 = tolerated; MetaLR: > 0.5 = damaging; <0.5 = tolerated; Mutation Tester: >0.5 = damaging; <0.5 = tolerated; PROVEAN: >0.5 = damaging; <0.5 = tolerated; REVEL: >0.5 = damaging; <0.5 = tolerated.

Whole-Exome Sequencing Analysis

The DNA library was prepared using Agilent Sure Select Target Enrichment Kit. DNA library was captured using ultralong 120 mer biotinylated cRNA baits. The library was sequenced using HiSeq2000 Next Generation Sequencer (Illumina, San Diego, CA, USA). The FASTQ format sequence was obtained, and reads were aligned using Burrows-Wheeler Aligner (BWA) software (Version bwa-0.7.12) against human genome reference sequence build 38 (GRCH38.p12). Variant calling was conducted using the genome analysis tool kit (GATK). The filtration pipeline was applied as follows: all coding variants that passed quality control (phred > 30 score) were included. All variants with a minor allele frequency (MAF) <0.015% were included. Known candidate gene variants were filtered based on the function of the gene and their role in LD development. Genes that are related to LD disease were collected through the Coremine Medical™ tool, National Center for Biotechnology Information (NCBI), OMIM, and literature review.

Variant Validation Using Sanger Sequencing Method

The potential LD candidate variant was validated using Sanger sequencer ABI 3500 Genetic Analyzer. The primers were designed using NCBI Primer Blast to capture targeted mutation in candidate gene. The sequence files (chromatogram) were analysed using BioEdit software.

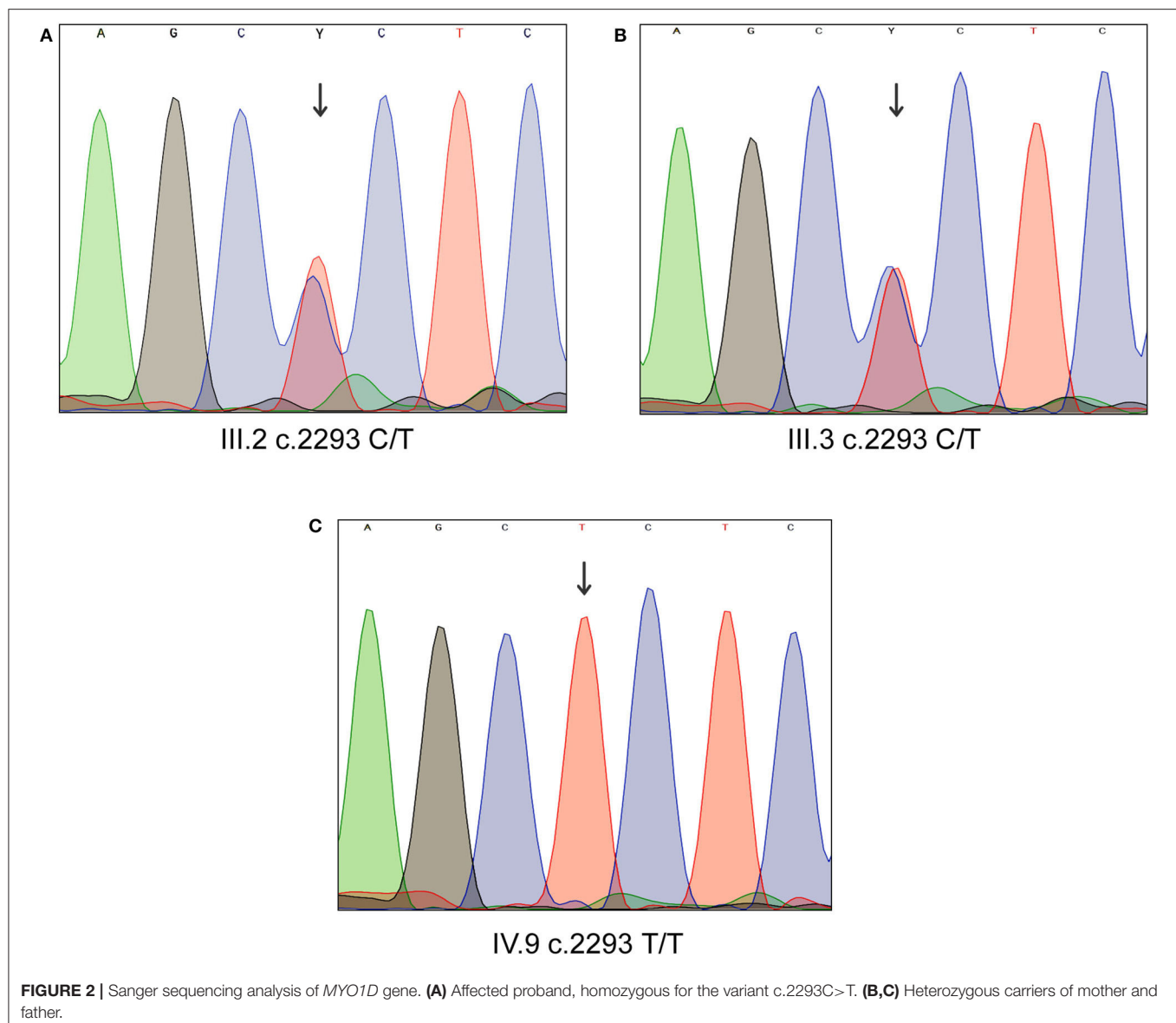
Functional Analysis of Laterality Defect Variant Using Computational Methods

Amino Acid Conserved Domains

The functional relevance of LD candidate genetic variant on candidate proteins was predicted by searching the nucleotide and amino acid sequences against the functional domains of concerned protein as per the listing available in Conserved Domain Database (CDD). To estimate the sequence conservation characteristics of the functional domains in the candidate protein, CDD tool uses RPS-BLAST, which rapidly scans the query protein for pre-computed position-specific scoring matrices (PSSMs). The output file demonstrates the links between protein domains with annotations against the query input sequence together with imagining choices (21).

3D Protein Modelling

The pathogenic effects of amino acid variants on disease candidate proteins can be best understood when they are studied at the structural level. Therefore, the potential effect of LD variant on the tertiary structural features was explored through 3D simulation of the candidate protein. Based on the availability of the X-ray crystallographic structure of the query protein, either a combination of *ab initio* approaches or homology modelling approaches were followed (22, 23). The 3D simulated structure



of the native protein was then used to construct the mutated version of candidate protein, which was then energy minimised and then analysed for structural deformities like amino acid or whole structure level deviations using YASARA software (24). The impact of candidate variant on the stability of protein structure was estimated using DUET webserver, which contains Protein Data Bank (PDB) structures of query proteins to predict the Gibbs free energy (G) values (25).

RNA Expression, Gene Ontology, and Mouse Gene Knockout Model

The Human Protein Atlas (HPA) (<https://www.proteinatlas.org/>) database was used to determine the RNA expression status of the

LD candidate gene. This database provides the expression profile of the query gene or protein based on primary antibody staining data in a series of immunohistochemistry pictures of clinical specimens. The functional enrichment analysis of the potential LD candidate gene was done using gene ontology (GO) webtool hosted in Ensembl web browser. Moreover, Mouse Genome Informatics (MGI) database (<http://www.informatics.jax.org/>) was used to better understand the functional role of potential LD gene on phenotype characteristics of knockout mouse models. The MGI resource provides a comprehensive set of data, tools, and analysis designed specifically for use in mouse laboratory model. It accepts input data in the form of a gene symbol and provides output corresponding to the physiological condition of knockout mice.

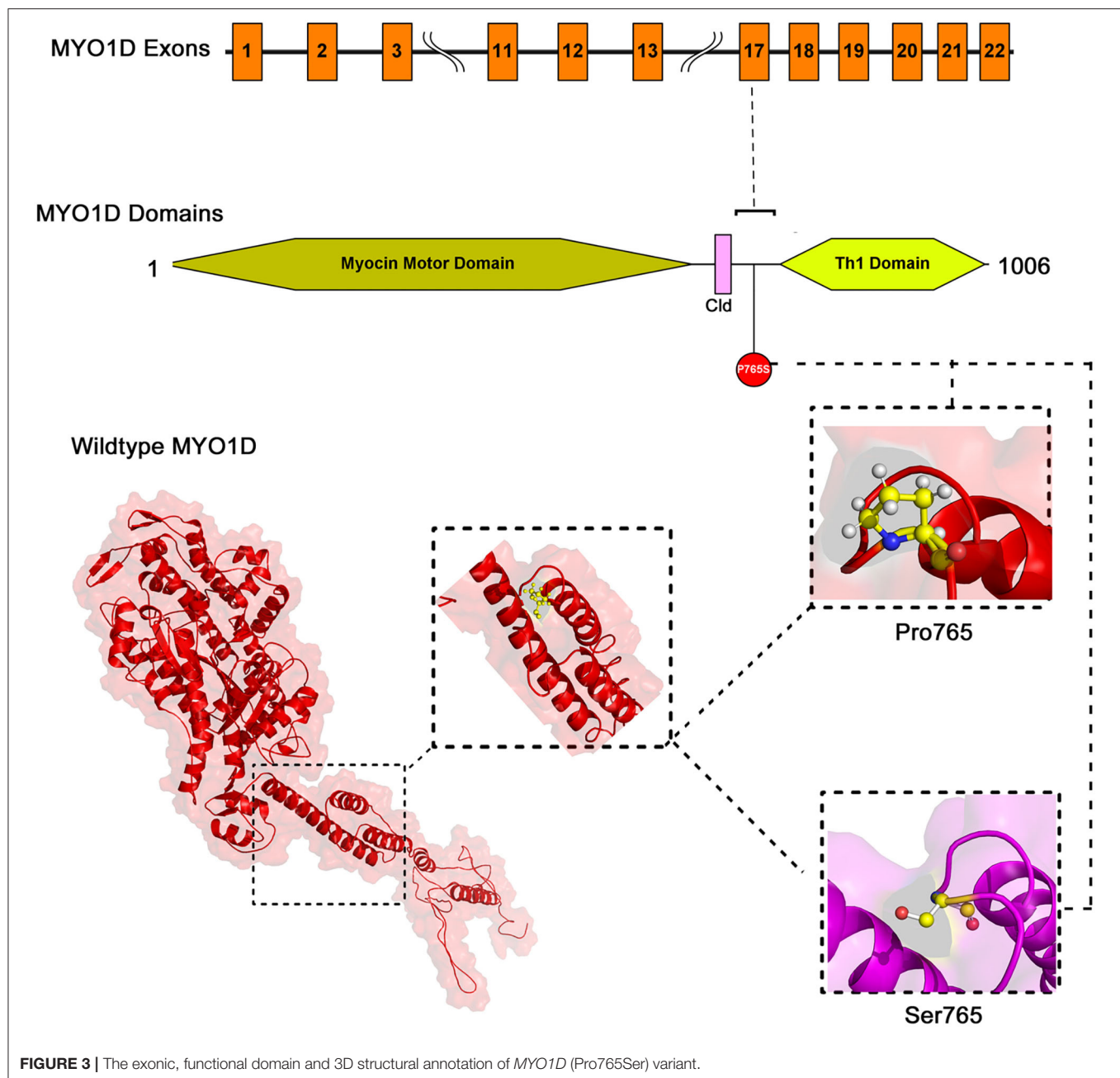


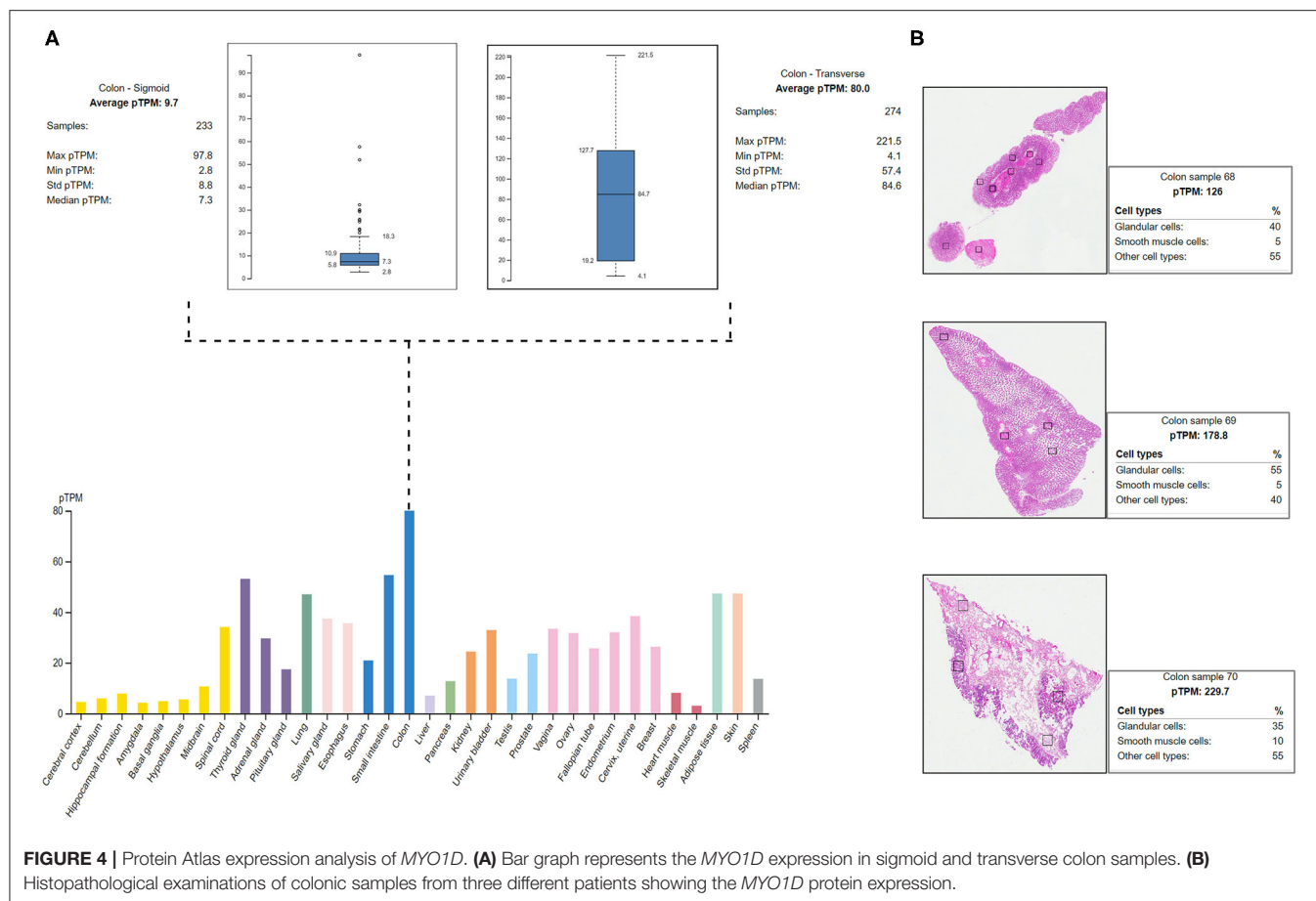
FIGURE 3 | The exonic, functional domain and 3D structural annotation of *MYO1D* (Pro765Ser) variant.

RESULTS

Clinical Assessment

The proband aged 4 years 6 months at the time of clinical diagnosis was born to an apparently healthy consanguineous parents of Arab origin (**Figure 1A**). The proband exhibited a spectrum of phenotypes including visceral heterotaxy (abnormal arrangements of thoracoabdominal organs) (**Figure 1B**), congenital cyanotic heart disease in the form of single ventricle physiology, left isomerism with polysplenia syndrome, double inlet atrioventricular connection (a heart defect that affects the valves and chambers), pulmonary atresia, interrupted

inferior vena cava with absent supra-renal segment, and azygos continuation (a rare congenital abnormality often combined with cardiovascular and visceral malformations). At the age of 10 months, the proband underwent thorough palliative cardiac procedures in the form of ductal stenting in the neonatal period followed by Kawashima cavo-pulmonary shunt (a palliative surgical procedure performed in cases of left isomerism and azygos continuation of the inferior vena cava, and common atrioventricular valve with or without regurgitation and pulmonary stenosis) in addition to left pulmonary artery (LPA) balloon dilatation procedure. At 3 years of age, Fontan completion was performed *via* incorporation of hepatic veins to



pulmonary artery correcting thereby blood flow from the lower body parts directly to the lungs.

Genetic Analysis

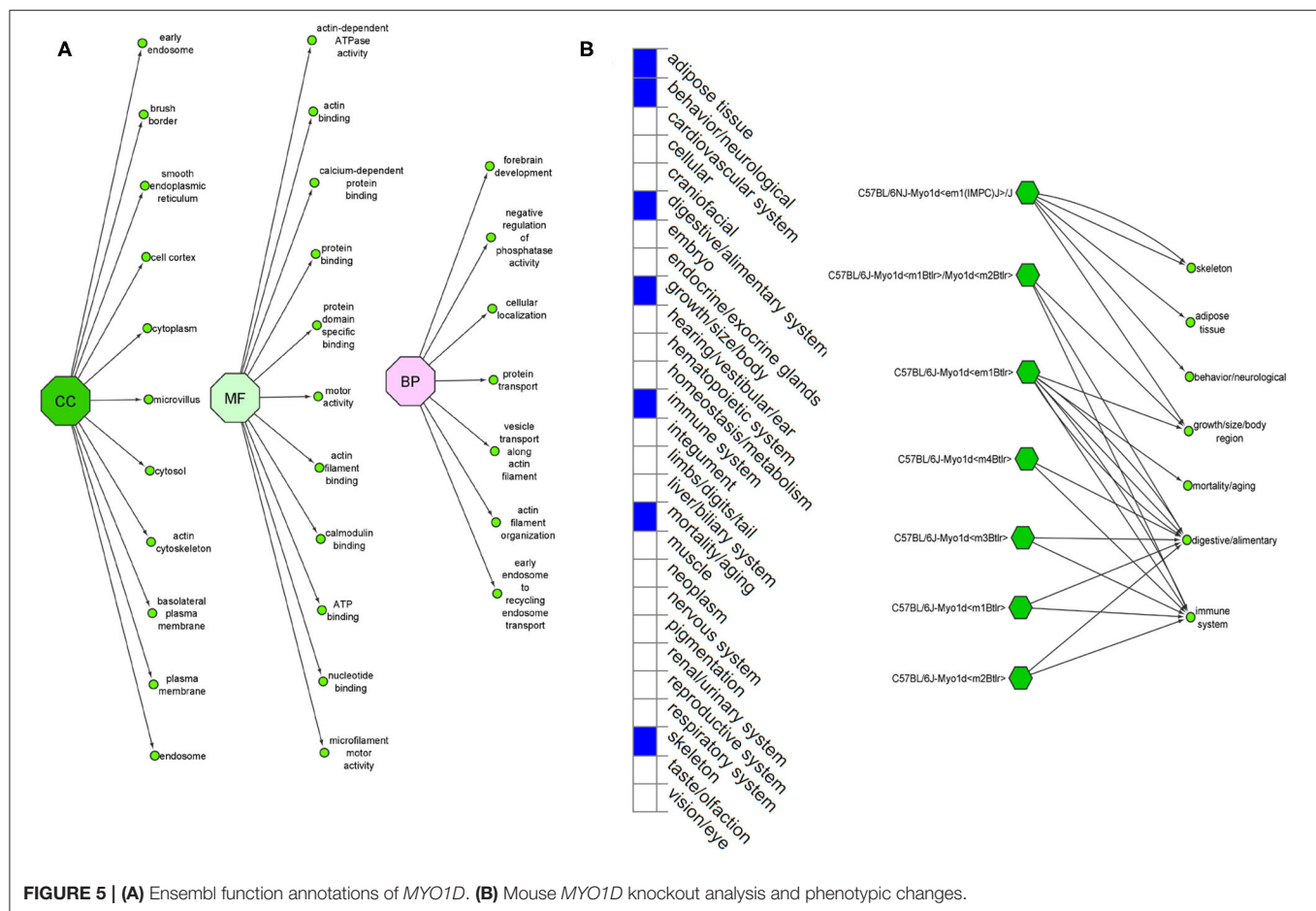
Whole-Exome Sequencing Variant Filtering and Novel Gene Identification

The sequencing of the index case generated approximately 98,000 variants, including 12,150 synonymous variants, 13,000 missense variants, and 11,500 indels. Variant filtration was based on its rare frequency, deleterious potential, autosomal recessive mode of inheritance, and functional relevance to disease (LD, primary ciliary dyskinesia (PCD), congenital heart disease, and heterotaxy). Nine genetic variants were identified as potential candidates (Table 1). Among these variants, only one missense variant (rs7209106: NM_015194.2:c.2293C>T; p.Pro765Ser) in *MYO1D* novel gene has survived our variant filtration criteria. This allele is absent in local databases like GME (Greater Middle East) (<http://igm.ucsd.edu/gme/>), DALIA (Disease Alleles in Arabs) (<http://clingen.igib.res.in/dalia/index>), and Saudi Human Genome Program (SHGP) (<https://shgp.kacst.edu.sa/index.en.html#home>). The MAF of this variant in international databases like 1,000 Genomes and gnomAD databases is 0.005 and 0.002, respectively. Although it has an allele frequency of 0.013 in the African population, only eight individuals are

reported as homozygous for this variant in the gnomAD. But their clinical details are not provided in the gnomAD database. In the index family studied here, both parents were heterozygous and do not have any symptoms associated with LD, confirming the autosomal recessive inheritance pattern as we initially deduced from their pedigree analysis. Moreover, more than 80% (5/6; 83.34) of the computational prediction methods like CADD, FATHMM, MetaLR, Mutation Taster, PROVEAN, and REVEL have attributed pathogenicity scores to this *MYO1D* (p.Pro765Ser) variant (Table 2). Functional biology data available from model organisms like *Drosophila*, zebrafish, and frog have proved the functional role of *MYO1D* gene in LDs.

Sanger Sequencing Validation

Sanger sequencing analysis confirmed that the LD patient is homozygous for c.2293C>T variant in *MYO1D* gene (1V.9, Figures 1A, 2), whereas the mother and father were heterozygous carriers (111.2, 111.3, Figures 1A, 2). This variant was absent in apparently healthy siblings and were homozygous for the T allele (1V.2, 1V.4, and 1V.5–8; Figure 1A). Two additional clinically diagnosed LD families were screened for this variant, and none carries this mutation, suggesting that *MYO1D* (p.Pro765Ser) variant is a rare private mutation in this family.



Computational Functional Analysis

Variant Mapping on MYO1D Protein Domain

The mapping of conserved amino acid domains is a vital step in deducing the association between the nucleotide sequence, protein structure, and function of disease-causing proteins. The CDD analysis showed that *MYO1D* protein is made up of three domains, namely, motor (11–682 amino acids), IQ calmodulin-binding motif (699–719 amino acids), and *Myosin TH1* (803–1,000 amino acids) domains. The Pro765Ser variant is located between the *Myosin TH1* and calmodulin-binding domains (Figure 3).

MYO1D 3D Model Construction

The PDB database search revealed the availability of partial 707 aa (between 10 and 717 of 1,006 AA long *MYO1D* protein) X-ray crystal protein model (4L79) with 2.3-Å resolution. Hence, the remaining 306 aa long chain was simulated with iterative threading assembly modification (I-TASSER) webserver following an *ab initio* approach. From the I-TASSER output, the best *MYO1D* model was chosen based on its polypeptide prediction quality scores like confidence score ($C = -1.52$), template modelling ($TM = 0.53 \pm 0.15$), and root-mean-square deviation (RMSD) (12.7 ± 4.3 Å) scores. These quality metrics indicate the very good structural similarity between the query

and template proteins (Figure 3). The stereochemical evaluation of the energy-minimised *MYO1D* protein model revealed that 96.2% of the amino acids are in the allowed portion of the protein, whereas 3.8% are in the non-allowed region. As per the above outlined processes, the native *MYO1D* model was used as a template to create a mutant variant by manually substituting proline for serine at the 765th position. PyMOL was used to depict native and mutant proteins.

Structural Deviation and Stability Findings

We have used YASARA tool to analyze α -atom coordinates of native and mutant *MYO1D* 3D structures to evaluate their structural drifts (in terms of RMSD) at residue and whole structure levels. RMSD value is used to quantitatively measure the structural similarity between two atomic coordinates when superimposed on each other. The impact of substitution mutations on amino acid structures can be calculated when there is a divergence at the polypeptide chain level. We noticed minor structural drifts in *MYO1D* structure only at 765th residue position due to the RMSD value difference (2.28) induced by the substitution of proline with serine (Figure 3). The DUET analysis of the *MYO1D* (P765S) variant predicted Gibbs free energy ($\Delta\Delta G$) alterations shifting the energy equilibrium to negative value, i.e., -0.959 kcal/mol, suggesting that the queried

TABLE 3 | Phenotypes and genetic data of LD and/or PCD among Arabs.

Nationality	Situs inversus (SI)	PCD	Dextrocardia	Heterotaxy of abdominal organs	Recurrent chest infections	Chronic rhinitis	Bronchiectasis	Chronic or recurrent otitis media	Gene	Reference
Yemeni	+	–	+	Visceral heterotaxy, polysplenia syndrome	–	–	–	–	<i>MYO1D</i>	Present study
Saudi (6 patients)	+6/6	–	NR	NR	–	NR	NR	NR	<i>DNAH5, PKD1L1, DNAAF5/CYP21A2, DNAI1</i>	(26)
Arab	+	NR	+	NR	NR	NR	NR	NR	<i>GDF1</i>	(27)
Saudi	+	NR	+	Abdominal ultrasound revealed that the liver and gallbladder were located in the left hypochondrium, spleen on the right side	NR	NR	NR	NR	NR	(28)
Arab	+	NR	+	+	NR	NR	NR	NR	<i>HES7</i>	(29)
Saudi (75 patients)	+	+73/75	NR	NR	+15/75	NR	NR	NR	<i>CCDC151, CCDC39, CCDC40, DNAH11, GOLGA3, RSPH9, CCNO, RSPH9, ITCH, MCIDAS, RSPH4A, DNAH5, CEP164, GOLGA3</i>	(26)
Morocco	+	+	+	+	+	+	+	–	NR	(30)
Saudi	+	+	+	NR	NR	+	+	+	<i>DNAH1</i>	(31)
Kuwaiti	+	+		NR	NR	+	NR	NR	<i>DNAH5</i>	(32),
Saudi	NR	+	NR	NR	+	+	+	NR	<i>CCDC151</i>	(33)
Palestine	NR	+	+	NR	NR	NR	+	+	<i>LRR6</i>	(34)
Saudi	NR	+	NR	NR	NR	NR	+	NR	<i>RSPH9</i>	(35)
UAE	NR	+	NR	NR	+	NR	+	NR	<i>RSPH9</i>	(36)
UAE	–	+	NR	NR	+	+	+	+	<i>RSPH9</i>	(37)
Saudi	+	+	NR	NR	NR	NR	+	NR	<i>19q13.3</i>	(27)
Saudi	+	+	–	Situs inversus of cardiac shadow and gastric air bubble	+	NR	+	–	NR	(38)

+, presence of symptoms; –, absence of symptoms; NR, not reported; PCD, primary ciliary dyskinesia.

variant is potentially deleterious to the protein stability owing to its destabilising behaviour.

RNA Expression Analysis

The HPA shows the positive expression status of *MYO1D* gene in different tissues and organs of the human body like the colon, lungs, and thyroid gland. In particular, the highest expression was seen in the digestive system with the colon, where the 274 transverse colon samples showed a maximum of 221.5 protein transcripts per million (pTPM) and 233 sigmoid colon samples showed a maximum of 97.8 pTPM. The RNA-Seq analysis of immunohistochemistry tissue specimens from three control specimens showed that glandular cells showed the highest pTPM status of *MYO1D* gene when compared with smooth muscle cells and other cell types in the colon (**Figure 4**).

Gene Ontology and Gene Knockout Analysis

The Ensembl GO analysis of *MYO1D* gene showed its involvement in 43 GO terms, including those connected to biological processes (seven GO terms), molecular functions (11 GO terms), and cellular component (25 GO terms) (**Supplementary Table 1**). All the annotations collectively highlight that *MYO1D* is localised in the cytoplasm (cellular component) and plays an important role in actin filament organisation (biological process) as well as microfilament motor activity function (molecular function). **Supplementary Table 2** shows details of disease phenotypes corresponding to the *MYO1D* genetic background in different knockout mice models.

DISCUSSION

Recent evolution and easy accessibility of next-generation sequencing resulted in accurate molecular diagnosis of variety of genetic diseases from around the globe (39, 40). Owing to the genetic heterogeneity, identification of specific molecular cause of LD is very challenging in up to 80% of the cases. Majority of known LD causative genes are structural proteins of the cilia and are known for their involvement in NODAL/TGF β or SHH signalling pathways. During embryonic development, these genes play an important role in symmetrical LR positioning of the organs (10).

MYO1D gene consists of 27 exons mapped to chromosome 17q11.2. Myosin heavy chain class 1 is a member of the myosin superfamily, playing essential roles in cytoskeletal structure, mechanical signal-transduction membrane dynamics (41), and endosome processing (42). In the present study, we identified the first case of a homozygous missense (c.2293C>T) mutation in *MYO1D* gene causing LD in the proband of an Arab consanguineous family. Further screening of LD participants from two additional Arab families did not reveal any mutations in this gene. In Middle Eastern Arab databases (with more than 10,000 exome data combined) such as SHGP, GMC, and DALIA, not a single case was recorded for this variant. Also, this variant was extremely rare (<0.005) in 1,000 Genomes and 0.013 in gnomAD across all ethnic groups. In gnomAD, only eight individuals were reported as homozygous (six males and two females), but no clinical data were available.

Various studies suggested the role of *MYO1D* gene in laterality disease in *Drosophila*, zebrafish, and frog (43–47). *MYO1D* has a role in organ asymmetry in *Drosophila*, which lacks cilia and nodal pathway while developing LD by using polar cell polarity (PCP), *MYO1D*, and *HOX* gene *Abd-B* (48, 49). Another study demonstrated the function of *MYO1D* in *Xenopus laevis* and influence the orientation of the cilia on the LR organiser (LRO) through planar cell polarity pathway as implicated in *Drosophila* (45). In zebrafish, *MYO1D* plays a fundamental role in the LR organisation (43, 50). Aside from the above initial reports, our understanding of *MYO1D* function in the context of human LR patterning remains largely unexplored. The *situs inversus* (SI) phenotype is reported in approximately 50% of PCD cases with congenital cardiac defects (51, 52). Approximately 3–7% of LD patients have CHDs (53). Many studies (54, 55) in a variety of species failed to identify a unifying mechanism for LR patterning. However, recent studies (43, 50) provided the first evidence of a shared origin of laterality in both arthropods and chordates through *MYO1D* gene. The clinical features are consistent with previous observations in zebrafish and *Drosophila*, indicating that *MYO1D* has an important role in LR patterning during embryogenesis. Moreover, asymmetric clustering of cilia was disrupted in ependymal cells of *MYO1D* KO rat models, consistent with LD (56).

MYO1D gene knockout in different mouse strains (seven different strains) is shown to demonstrate a variety of phenotypes like decreased body fat amount (adipose tissue), decreased startle reflex (behaviour/neurological), increased susceptibility to colitis (digestive/alimentary), decreased bone mineral content (skeleton), and increased susceptibility to weight loss (growth size/body weight and immune system and mortality or ageing) (**Figure 5**). Our study is the first one to report the association of defective *MYO1D* to LD in humans, confirming that the function is evolutionarily conserved from *Drosophila*, to zebrafish, to frog, to humans. Thus, *MYO1D* gene can be considered as the new causal gene for LD in humans. Though the *Drosophila* and zebrafish models clearly showed the visceral heterotaxy, mouse KO phenotypes were surprisingly not showing any LD-related phenotypes.

Extensive computational analysis of the protein structure and function adds the supporting evidence for *MYO1D* in LD. *MYO1D* protein is 1,006 aa long with a molecular weight of 116 kDa. It consists of a large, highly conserved Myocin Motor Domain (671 aa), short calmodulin-binding motif (20 aa), and a basic C-terminal tail homology-1 (TH1) domain (197 aa). The amino acid residue level structural deviation observed with the variant serine (Pro765Ser) in *MYO1D* is likely to disturb the primary, secondary, tertiary, and quaternary structural features in the protein. Numerous studies have shown the strong correlation between deviations in residue level RMSD score and structural properties for the disease-causing variants (23, 57–59). Disease causative pathogenic mutations have often changed the energy equilibrium, which is required to maintain the protein stability (60). Given the close physical proximity of P765S variant between calmodulin-binding motif and TH1 domains, the conformational and stability changes in *MYO1D* protein are likely to impact its main biological functions such as calmodulin

binding, actin-dependent ATPase activity, calcium-dependent protein binding, and microfilament motor activities (61).

LD is a complex disease, and its clinical phenotype presentations often overlap with PCD symptoms. A recent study from Saudi Arabia reported the overlapping clinical symptoms between PCD and LD patients (26). This report investigated a total of 81 patients, including 58 patients with sinopulmonary infections (SPIs), 15 patients with combined LD with SPIs, and six patients with LD alone. They reported mutations in the known PCD genes as follows: *RSPH9*, *CCNO*, *DNAAF5*, *RSPH4A*, *MCIDAS*, and *CCDC40* gene mutations in PCD patients with SPIs; *CCDC151*, *DNAH11*, *CCDC40*, *DNAH5*, and *CCDC39* gene mutations in LD patients with SPIs; *PKD1L1* and *DNAAF5* gene mutations in LD patients; and *RSPH9* and *MCIDAS* gene mutations in neonatal respiratory distress. Additionally, they have also identified gene mutations in *ITCH* and *CEP164* in two patients, demonstrating ITCH-related syndrome and Bardet-Biedl syndrome. Sparrow et al. (29) reported *HES7* as a cause of spondylocostal dysostosis with SI and dextrocardia. Molecular diagnosis of LD and PCD in Arab patients has revealed a spectrum of mutations in many genes with variable clinical presentations (35), which is summarised in Table 3.

CONCLUSION

In conclusion, we discovered missense mutation in *MYO1D* gene (c.2293C>T) in an Arab patient presenting with visceral heterotaxy and left isomerism with polysplenia syndrome by using higher-throughput WES technology. This is the first report to establish the relationship between *MYO1D* variants and LD, supporting the previous findings in *Drosophila zebrafish*, and frog. This exciting finding may support the critical role of *MYO1D* gene for LR patterning in humans. This study has some sincere limitations, as this is the first case identified with *MYO1D* mutation potentially contributing to LD phenotypes, and there are no reported cases with *MYO1D* variants to compare our data with. Therefore, testing *MYO1D* variants for LD patients in large cohort studies is recommended to verify our findings. Future functional studies are also recommended to investigate the specific molecular role and therapeutic prospects of targeting *MYO1D* genetic variants in patients demonstrating LD phenotypes.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because (a) participants' refusal to store or distribute the genomic

data in the public domain and (b) as per the local Institutional Ethics committee approval and Saudi national policy on genomic data sharing in the public domain outside the country. Requests to access the data should be directed to RE or TS.

ETHICS STATEMENT

The Ethical approval for the present study was obtained from institutional Ethics Committee of King Abdulaziz University Hospital (KAUH), Jeddah, Saudi Arabia and informed consent forms were collected from both adult parents and their children (guardian consent for Language evaluation).

CONSENT TO PARTICIPATE

Informed consent was obtained from all subjects involved in the study.

AUTHOR CONTRIBUTIONS

TA, RE, and KN: conceptualisation. KN, RA, BB, HE, and RE: methodology. BB: software and visualisation. KN, RA, BB, NS, and RE: formal analysis. RA, KN, TS, and RE: investigation. KN, TS, and BB: resources. RA, KN, TS, BB, ZZ, GA, NS, JA-A, AS, OA-R, RE, and TA: writing—original draft preparation. KN, TS, NS, and RE: writing—review and editing. KN, BB, and RE: supervision. TS: project administration and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This research work was funded by Institutional Fund Projects under grant no. (IFPRC-132-290-2020). Therefore, authors gratefully acknowledge technical and financial support from the Ministry of Education and King Abdulaziz University, Jeddah, Saudi Arabia.

ACKNOWLEDGMENTS

The authors gratefully acknowledge technical and financial support from the Ministry of Education and King Abdulaziz University, Jeddah, Saudi Arabia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.724826/full#supplementary-material>

REFERENCES

- Shapiro AJ, Davis SD, Ferkol T, Dell SD, Rosenfeld M, Olivier KN, et al. Laterality defects other than situs inversus totalis in primary ciliary dyskinesia: insights into situs ambiguus and heterotaxy. *Chest*. (2014) 146:1176–86. doi: 10.1378/chest.13-1704
- Kosaki K, Casey B. Genetics of human left-right axis malformations. *Semin Cell Dev Biol*. (1998) 9:89–99. doi: 10.1006/scdb.1997.0187

3. Lopez KN, Li AH, Hanchard N, Azamian M, Lalani S, Dickerson H, et al. Whole exome sequencing in congenital heart disease reveals variants in left-right patterning genes previously associated with heterotaxy syndrome and primary ciliary dyskinesia. *J Am Coll Cardiol*. (2015) 65(10 Suppl.):A490. doi: 10.1016/S0735-1097(15)60490-9
4. Lin AE, Krikov S, Riehle-Colarusso T, Frias JL, Belmont J, Anderka M, et al. Laterality defects in the national birth defects prevention study (1998-2007): birth prevalence and descriptive epidemiology. *Am J Med Genet A*. (2014) 164A:2581-91. doi: 10.1002/ajmg.a.36695
5. Best S, Shoemark A, Rubbo B, Patel MP, Fassad MR, Dixon M, et al. Risk factors for situs defects and congenital heart disease in primary ciliary dyskinesia. *Thorax*. (2019) 74:203-5. doi: 10.1136/thoraxjnl-2018-212104
6. Kuehl KS, Loffredo C. Risk factors for heart disease associated with abnormal sidedness. *Teratology*. (2002) 66:242-8. doi: 10.1002/tera.10099
7. Kuehl KS, Loffredo CA. Population-based study of l-transposition of the great arteries: possible associations with environmental factors. *Birth Defects Res A Clin Mol Teratol*. (2003) 67:162-7. doi: 10.1002/bdra.10015
8. Sutherland MJ, Ware SM. Disorders of left-right asymmetry: heterotaxy and situs inversus. *Am J Med Genet C*. (2009) 151C:307-17. doi: 10.1002/ajmg.c.30228
9. Deng H, Xia H, Deng S. Genetic basis of human left-right asymmetry disorders. *Expert Rev Mol Med*. (2015) 16:e19. doi: 10.1017/erm.2014.22
10. Li AH, Hanchard NA, Azamian M, D'alessandro LCA, Coban-Akdemir Z, Lopez KN, et al. Genetic architecture of laterality defects revealed by whole exome sequencing. *Eur J Hum Genet*. (2019) 27:563-73. doi: 10.1038/s41431-018-0307-z
11. Ware SM, Jefferies JL. New genetic insights into congenital heart disease. *J Clin Exp Cardiol*. (2012) S8:003. doi: 10.4172/2155-9880.S8-003
12. Collignon J, Varlet I, Robertson EJ. Relationship between asymmetric nodal expression and the direction of embryonic turning. *Nature*. (1996) 381:155-8. doi: 10.1038/381155a0
13. Olson EN, Srivastava D. Molecular pathways controlling heart development. *Science*. (1996) 272:671-6. doi: 10.1126/science.272.5262.671
14. Iratni R, Yan YT, Chen C, Ding J, Zhang Y, Price SM, et al. Inhibition of excess nodal signaling during mouse gastrulation by the transcriptional corepressor DRAP1. *Science*. (2002) 298:1996-9. doi: 10.1126/science.1073405
15. Otto EA, Schermer B, Obara T, O'toole JF, Hiller KS, Mueller AM, et al. Mutations in INVS encoding inversin cause nephronophthisis type 2, linking renal cystic disease to the function of primary cilia and left-right axis determination. *Nat Genet*. (2003) 34:413-20. doi: 10.1038/ng1217
16. Qian F, Germino FJ, Cai Y, Zhang X, Somlo S, Germino GG. PKD1 interacts with PKD2 through a probable coiled-coil domain. *Nat Genet*. (1997) 16:179-83. doi: 10.1038/ng0697-179
17. Yoshida S, Shiratori H, Kuo IY, Kawasumi A, Shinohara K, Nonaka S, et al. Cilia at the node of mouse embryos sense fluid flow for left-right determination via Pkd2. *Science*. (2012) 338:226-31. doi: 10.1126/science.1222538
18. Calvet JP. Ciliary signaling goes down the tubes. *Nat Genet*. (2003) 33:113-4. doi: 10.1038/ng1078
19. Grimes DT, Keynton JL, Buenavista MT, Jin X, Patel SH, Kyosuke S, et al. Genetic analysis reveals a hierarchy of interactions between polycystin-encoding genes and genes controlling cilia function during left-right determination. *PLoS Genet*. (2016) 12:e1006070. doi: 10.1371/journal.pgen.1006070
20. Saadah OI, Banaganapalli B, Kamal NM, Sahly AN, Alsufyani HA, Mohammed A, et al. Identification of a rare exon 19 skipping mutation in ALMS1 gene in alström syndrome patients from two unrelated saudi families. *Front Pediatr*. (2021) 9:652011. doi: 10.3389/fped.2021.652011
21. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, Deweese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res*. (2010) 39:D225-D9. doi: 10.1093/nar/gkq1189
22. Ahmed Awan Z, Bima A, Rashidi OM, Jamil K, Khan IA, Almukadi HS, et al. Low resolution protein mapping and KB-R7943 drug-protein molecular interaction analysis of long-QT syndrome linked KCNH2 mutations. *All Life*. (2020) 13:183-93. doi: 10.1080/26895293.2020.1737249
23. Awan ZA, Bahattab R, Kutbi HI, Noor AOF, Al-Nasser MS, Shaik NA, et al. Structural and molecular interaction studies on familial hypercholesterolemia causative PCSK9 functional domain mutations reveals binding affinity alterations with LDLR. *Int J Pept Res Ther*. (2021) 27:719-33. doi: 10.1007/s10989-020-10121-8
24. Shaik NA, Bokhari HA, Masoodi TA, Shetty PJ, Ajabnoor GMA, Elango R, et al. Molecular modelling and dynamics of CA2 missense mutations causative to carbonic anhydrase 2 deficiency syndrome. *J Biomol Struct Dyn*. (2020) 38:4067-80. doi: 10.1080/07391102.2019.1671899
25. Pires DE, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res*. (2014) 42:W314-319. doi: 10.1093/nar/gku411
26. Shamseldin HE, Al Mogarri I, Alqwaiee MM, Alharbi AS, Baqais K, Alsaadi M, et al. An exome-first approach to aid in the diagnosis of primary ciliary dyskinesia. *Hum Genet*. (2020) 139:1273-83. doi: 10.1007/s00439-020-02170-2
27. Marek-Yagel D, Bolkier Y, Barel O, Vardi A, Mishali D, Katz U, et al. A founder truncating variant in GDF1 causes autosomal-recessive right isomerism and associated congenital heart defects in multiplex Arab kindreds. *Am J Med Genet A*. (2020) 182:987-93. doi: 10.1002/ajmg.a.61509
28. Saleh A. Dextrocardia with situs inversus totalis in a boy: a case report. *Int J Contemp Pediatr*. (2016) 1096-11098. doi: 10.18203/2349-3291.ijcp20162398
29. Sparrow DB, Faqeh EA, Sallout B, Alswaid A, Ababneh F, Al-Sayed M, et al. Mutation of HES7 in a large extended family with spondylocostal dysostosis and dextrocardia with situs inversus. *Am J Med Genet A*. (2013) 161A:2244-9. doi: 10.1002/ajmg.a.36073
30. Raoufi M, Sator H, Lahma J, El Ayoubi A, Nitassi S, Oujilal A, et al. A case of Kartagener syndrome with rhinolalia clausa. *Pan Afr Med J*. (2016) 23:159. doi: 10.11604/pamj.2016.23.159.8664
31. Imtiaz F, Allam R, Ramzan K, Al-Sayed M. Variation in DNAH1 may contribute to primary ciliary dyskinesia. *BMC Med Genet*. (2015) 16:14. doi: 10.1186/s12881-015-0162-5
32. Marafie MJ, Al Suliman IS, Redha AM, Alshati AM. Primary ciliary dyskinesia: Kartagener syndrome in a family with a novel DNAH5 gene mutation and variable phenotypes. *Egypt J Med Hum Genet*. (2015) 16:95-9. doi: 10.1016/j.ejmhg.2014.08.001
33. Hjeij R, Onoufriadi A, Watson CM, Slagle CE, Klena NT, Dougherty GW, et al. CCDC151 mutations cause primary ciliary dyskinesia by disruption of the outer dynein arm docking complex formation. *Am J Hum Genet*. (2014) 95:257-74. doi: 10.1016/j.ajhg.2014.08.005
34. Horani A, Ferkol TW, Shoseyov D, Wasserman MG, Oren YS, Kerem B, et al. LRRC6 mutation causes primary ciliary dyskinesia with dynein arm defects. *PLoS One*. (2013) 8:e59436. doi: 10.1371/journal.pone.0059436
35. Alsaadi MM, Gaunt TR, Boustred CR, Guthrie PA, Liu X, Lenzi L, et al. From a single whole exome read to notions of clinical screening: primary ciliary dyskinesia and RSPH9 p.Lys268del in the Arabian Peninsula. *Ann Hum Genet*. (2012) 76:211-20. doi: 10.1111/j.1469-1809.2012.00704.x
36. Reish O, Slatkin M, Chapman-Shimshoni D, Elizur A, Chioza B, Castleman V, et al. Founder mutation(s) in the RSPH9 gene leading to primary ciliary dyskinesia in two inbred Bedouin families. *Ann Hum Genet*. (2010) 74:117-25. doi: 10.1111/j.1469-1809.2009.00559.x
37. Castleman VH, Romio L, Chodhari R, Hirst RA, De Castro SC, Parker KA, et al. Mutations in radial spoke head protein genes RSPH9 and RSPH4A cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *Am J Hum Genet*. (2009) 84:197-209. doi: 10.1016/j.ajhg.2009.01.011
38. Bashi S, Khan MA, Guirjis A, Johariy IA, Abid MA. Immotile-cilia syndrome with azoospermia: a case report and review of the literature. *Br J Dis Chest*. (1988) 82:194-6. doi: 10.1016/0007-0971(88)90043-5
39. Bokhari HA, Shaik NA, Banaganapalli B, Nasser KK, Ageel HI, Al Shamrani AS, et al. Whole exome sequencing of a Saudi family and systems biology analysis identifies CPED1 as a putative causative gene to Celiac Disease. *Saudi J Biol Sci*. (2020) 27:1494-502. doi: 10.1016/j.sjbs.2020.04.011
40. Gaboon NEA, Banaganapalli B, Nasser K, Razeeth M, Alsaedi MS, Rashidi OM, et al. Exome sequencing and metabolomic analysis of a chronic kidney disease and hearing loss patient family revealed RMND1 mutation induced sphingolipid metabolism defects. *Saudi J Biol Sci*. (2020) 27:324-34. doi: 10.1016/j.sjbs.2019.10.001
41. Huber LA, Fialka I, Paiha K, Hunziker W, Sacks DB, Bähler M, et al. Both calmodulin and the unconventional myosin Myr4 regulate membrane

- trafficking along the recycling pathway of MDCK cells. *Traffic*. (2000) 1:494–503. doi: 10.1034/j.1600-0854.2000.010607.x
42. De La Cruz EM, Ostap EM. Relating biochemistry and function in the myosin superfamily. *Curr Opin Cell Biol*. (2004) 16:61–7. doi: 10.1016/j.ceb.2003.11.011
43. Juan T, Geminard C, Coutelis JB, Cerezo D, Poles S, Noselli S, et al. Myosin1D is an evolutionarily conserved regulator of animal left-right asymmetry. *Nat Commun*. (2018) 9:1942. doi: 10.1038/s41467-018-04284-8
44. Lebreton G, Geminard C, Lapraz F, Pyrpasopoulos S, Cerezo D, Speder P, et al. Molecular to organismal chirality is induced by the conserved myosin 1D. *Science*. (2018) 362:949–52. doi: 10.1126/science.aat8642
45. Tingle M, Kurz S, Maerker M, Ott T, Fuhl F, Schweickert A, et al. A conserved role of the unconventional myosin 1D in laterality determination. *Curr Biol*. (2018) 28:810–16.e813. doi: 10.1016/j.cub.2018.01.075
46. Yuan S, Brueckner M. Left-right asymmetry: myosin 1D at the center. *Curr Biol*. (2018) 28:R567–R9. doi: 10.1016/j.cub.2018.03.01
47. Blum M, Ott T. Mechanical strain, novel genes and evolutionary insights: news from the frog left-right organizer. *Curr Opin Genet Dev*. (2019) 56:8–14. doi: 10.1016/j.gde.2019.05.005
48. Spéder P, Ádám G, Noselli S. Type 1D unconventional myosin controls left-right asymmetry in *Drosophila*. *Nature*. (2006) 440:803–7. doi: 10.1038/nature04623
49. González-Morales N, Geminard C, Lebreton G, Cerezo D, Coutelis J-B, Noselli S. The atypical cadherin dachsous controls left-right asymmetry in *Drosophila*. *Dev Cell*. (2015) 33:675–89. doi: 10.1016/j.devcel.2015.04.026
50. Blum M, Ott T. Animal left-right asymmetry. *Curr Biol*. (2018) 28:R301–R4. doi: 10.1016/j.cub.2018.02.073
51. Mirra V, Werner C, Santamaria F. Primary ciliary dyskinesia: an update on clinical aspects, genetics, diagnosis, and future treatment strategies. *Front Pediatr*. (2017) 5:135. doi: 10.3389/fped.2017.00135
52. Boon M, Smits A, Cuppens H, Jaspers M, Proesmans M, Dupont LJ, et al. Primary ciliary dyskinesia: critical evaluation of clinical symptoms and diagnosis in patients with normal and abnormal ultrastructure. *Orphanet J Rare Dis*. (2014) 9:11. doi: 10.1186/1750-1172-9-11
53. Escobar-Diaz MC, Friedman K, Salem Y, Marx GR, Kalish BT, Lafranchi T, et al. Perinatal and infant outcomes of prenatal diagnosis of heterotaxy syndrome (asplenia and polysplenia). *Am J Cardiol*. (2014) 114:612–7. doi: 10.1016/j.amjcard.2014.05.042
54. Blum M, Feistel K, Thumberger T, Schweickert A. The evolution and conservation of left-right patterning mechanisms. *Development*. (2014) 141:1603–13. doi: 10.1242/dev.100560
55. Coutelis JB, González-Morales N, Geminard C, Noselli S. Diversity and convergence in the mechanisms establishing L/R asymmetry in metazoa. *EMBO Rep*. (2014) 15:926–37. doi: 10.15252/embr.201438972
56. Hegan PS, Ostertag E, Geurts AM, Mooseker MS. Myosin Id is required for planar cell polarity in ciliated tracheal and ependymal epithelial cells. *Cytoskeleton (Hoboken)*. (2015) 72:503–16. doi: 10.1002/cm.21259
57. Nasser KK, Banaganapalli B, Shinawi T, Elango R, Shaik NA. Molecular profiling of lamellar ichthyosis pathogenic missense mutations on the structural and stability aspects of TGM1 protein. *J Biomol Struct Dyn*. (2020) 1–11. doi: 10.1080/07391102.2020.1782770
58. Ibrahim AZ, Thirumal Kumar D, Abunada T, Younes S, George Priya Doss C, Zaki OK, et al. Investigating the structural impacts of a novel missense variant identified with whole exome sequencing in an Egyptian patient with propionic acidemia. *Mol Genet Metab Rep*. (2020) 25:100645. doi: 10.1016/j.ymgmr.2020.100645
59. Thirumal Kumar D, Udhaya Kumar S, Nishaat Laeeque AS, Apurva Abhay S, Bithia R, Magesh R, et al. Computational model to analyze and characterize the functional mutations of NOD2 protein causing inflammatory disorder - Blau syndrome. *Adv Protein Chem Struct Biol*. (2020) 120:379–408. doi: 10.1016/bs.apcsb.2019.11.005
60. Shaik NA, Nasser KK, Alruwaili MM, Alallasi SR, Elango R, Banaganapalli B. Molecular modelling and dynamic simulations of sequestosome 1 (SQSTM1) missense mutations linked to Paget disease of bone. *J Biomol Struct Dyn*. (2021) 39:2873–84. doi: 10.1080/07391102.2020.1758212
61. Ko Y-S, Bae JA, Kim KY, Kim SJ, Sun EG, Lee KH, et al. MYO1D binds with kinase domain of the EGFR family to anchor them to plasma membrane before their activation and contributes carcinogenesis. *Oncogene*. (2019) 38:7416–32. doi: 10.1038/s41388-019-0954-8

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

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

Copyright © 2021 Alsafwani, Nasser, Shinawi, Banaganapalli, ElSokary, Zaher, Shaik, Abdelmohsen, Al-Aama, Shapiro, O. Al-Radi, Elango and Alahmadi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Development and Verification of an Autophagy-Related lncRNA Signature to Predict Clinical Outcomes and Therapeutic Responses in Ovarian Cancer

Yan Li^{1†}, Juan Wang^{2†}, Fang Wang², Chengzhen Gao¹, Yuanyuan Cao¹ and Jianhua Wang^{3*}

¹ Department of Obstetrics and Gynecology, The Yancheng Clinical College of Xuzhou Medical University, The First People's Hospital of Yancheng, Yancheng, China, ² Department of Obstetrics and Gynecology, The First Affiliated Hospital of Soochow University, Suzhou, China, ³ Department of Gastroenterology, The Yancheng Clinical College of Xuzhou Medical University, The First People's Hospital of Yancheng, Yancheng, China

OPEN ACCESS

Edited by:

Balu Kamaraj,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia

Reviewed by:

Abdul Azeez Sayed,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia
Sabeena Mustafa,
King Abdullah International Medical
Research Center (KAIMRC),
Saudi Arabia

*Correspondence:

Jianhua Wang
jjahuw@163.com

[†]These authors have contributed
equally to this work

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 26 May 2021

Accepted: 30 August 2021

Published: 04 October 2021

Citation:

Li Y, Wang J, Wang F, Gao C, Cao Y
and Wang J (2021) Development and
Verification of an Autophagy-Related
lncRNA Signature to Predict Clinical
Outcomes and Therapeutic
Responses in Ovarian Cancer.
Front. Med. 8:715250.
doi: 10.3389/fmed.2021.715250

Objective: Long noncoding RNAs (lncRNAs) are key regulators during ovarian cancer initiation and progression and are involved in mediating autophagy. In this study, we aimed to develop a prognostic autophagy-related lncRNA signature for ovarian cancer.

Methods: Autophagy-related abnormally expressed lncRNAs were screened in ovarian cancer with the criteria values of |correlation coefficient| > 0.4 and $p < 0.001$. Based on them, a prognostic lncRNA signature was established. The Kaplan–Meier overall survival analysis was conducted in high- and low-risk samples in the training, verification, and entire sets, followed by receiver operating characteristics (ROCs) of 7-year survival. Multivariate Cox regression analysis was used for assessing the predictive independency of this signature after adjusting other clinical features. The associations between the risk scores and immune cell infiltration, PD-L1 expression, and sensitivity of chemotherapy drugs were assessed in ovarian cancer.

Results: A total of 66 autophagy-related abnormally expressed lncRNAs were identified in ovarian cancer. An autophagy-related lncRNA signature was constructed for ovarian cancer. High-risk scores were indicative of poorer prognosis compared with the low-risk scores in the training, verification, and entire sets. ROCs of 7-year survival confirmed the well-predictive efficacy of this model. Following multivariate Cox regression analysis, this model was an independent prognostic factor. There were distinct differences in infiltrations of immune cells, PD-L1 expression, and sensitivity of chemotherapy drugs between high- and low-risk samples.

Conclusions: This study constructed an autophagy-related lncRNA signature that was capable of predicting clinical outcomes and also therapeutic responses for ovarian cancer.

Keywords: ovarian cancer, autophagy, lncRNAs, signature, prognosis, tumor microenvironment

INTRODUCTION

Ovarian cancer represents the major cause of death among gynecological malignancies (1). Surgery followed by chemotherapy (such as platinum and taxane) remained the first-line therapeutic strategy (2). Approximately 80% of the subjects originally respond to this therapy. Nevertheless, the majority of the subjects in late stages usually experienced recurrence following chemotherapy, thereby leading to an undesirable prognosis (5-year survival < 50%). Hence, it is significant to probe into the pathogenesis of ovarian cancer and also predictive indicators for prognostic stratification.

Alterations in gene expression profiling have become fundamental laboratory tools for improving tumor diagnoses, survival outcomes, and also treatment responses, which overcome weaknesses of typical clinical and imaging features due to heterogeneity at the genetic and molecular levels (3). Dysregulated long non-coding RNAs (lncRNAs) such as LINC00189, CACNA1G-AS1, and CHRM3-AS2 have been implicated in ovarian cancer initiation and the progress, highlighting their promising functions as markers of precision medicine (4). Their correlations to survival outcomes and treatment responses are still indistinct in ovarian cancer. A few lncRNA-based expression signatures have been developed for predicting survival outcomes, status, and chemosensitivity in ovarian cancer. For instance, Zheng et al. developed a three-lncRNA signature (LOC101927151, LINC00861, and LEMD1-AS1) for predicting clinical outcomes of the subjects with ovarian cancer on the basis of copy number variation (5). Zhang et al. proposed a three-lncRNA signature (LINC01619, DLX6-AS1, and AC004943.2) that can predict survival and response of chemotherapy in ovarian cancer (6). In comparison to abundant lncRNAs identified by genome-wide studies, functionally lncRNAs require better characterization in ovarian cancer.

Autophagy, an evolutionarily conserved process, maintains cell homeostasis through a lysosomal degradation system that supports cell survival and also maintains homeostasis under various types of stress (7). Autophagy-based cell deaths provide molecular mechanisms and clinical implications upon ovarian cancer therapy (8). Thus, it is of significance to identify key regulators of autophagy for theoretical basis and clinical practice. Several studies have found that autophagy can be mediated by lncRNA regulators in ovarian cancer (9–12). For instance, GAS8-AS1 inhibits the progress of ovarian cancer by activation of autophagy through binding with Beclin1 (9). Moreover, lncRNA TUG1 induces autophagy-related paclitaxel resistance *via* sponging miR-29b-3p in ovarian cancer (10). Also, silencing HOTAIR enhances the sensitivity to cisplatin in ovarian cancer *via* inhibition of cisplatin-induced autophagy (11). lncRNA highly upregulated in liver cancer exerts a carcinogenic effect

via targeting autophagy-related genes ATG7 and ITGB1 in an epithelial ovarian cancer (12). Thus, autophagy-related lncRNAs possess potential as prognostic indicators and therapeutic targets.

Ovarian cancer represents an insidious malignancy, which usually develops asymptotically to late stages along with metastases, chemoresistance, and undesirable clinical outcomes (13). Autophagy is a key bioprocess during the initiation and progression of ovarian cancer (14). lncRNA regulators are involved in the autophagy process. There is still a lack of systematic analysis for identifying autophagy-related lncRNA signature for prediction of the survival outcomes of the patients with ovarian cancer. Herein, this study developed a prognostic autophagy-related lncRNA signature for ovarian cancer.

MATERIALS AND METHODS

Data Retrieval and Pre-processing

Among all databases, only the Cancer Genome Atlas (TCGA; <https://portal.gdc.cancer.gov/>) database has the RNA-seq expression profiling and corresponding clinical and prognostic information of ovarian cancer. Thus, we curated transcriptome data and clinical information containing age, survival time, recurrence, survival status, histologic grade, and pathologic stage of 379 ovarian cancer tissues from the TCGA database. Mutation annotation format (MAF) files of somatic mutation data of ovarian cancer samples were also downloaded from the TCGA database. Meanwhile, the RNA-seq expression profiles of 133 normal ovarian samples were obtained from the genotype tissue expression (GTEx; https://toil.xenahubs.net/download/GTEX_phenotype.gz) database (15). The RNA-seq counts values from the TCGA and GTEx databases were normalized and preprocessed by the TCGAbiolinks package (16). Based on the HUGO Gene Nomenclature Committee (HGNC; <http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/searchgenes.pl>), lncRNAs and mRNAs were annotated (17).

Acquisition of Abnormally Expressed lncRNAs

Abnormally expressed lncRNAs between ovarian cancer and normal ovarian specimens were screened utilizing edgeR package (<http://bioconductor.org>) based on gene expression data (18). Adjusted $p \leq 0.05$ and $|\log_2 \text{fold change (FC)}| \geq 1$ were set as the criteria values of abnormally expressed lncRNAs.

Acquisition of Autophagy-Related lncRNAs

A total of 232 autophagy-related genes were retrieved from the Human Autophagy Database (HADb; <http://www.autophagy.lu/>). The correlation between abnormally expressed lncRNAs and autophagy-related genes was analyzed by psych package (<https://CRAN.R-project.org/package=psych>) with Pearson's correlation analysis. Autophagy-related lncRNAs were screened with the criteria values of $|\text{correlation coefficient}| > 0.4$ and $p < 0.001$.

Construction of a LASSO Regression Model

By adopting the least absolute shrinkage and selection operator (LASSO) Cox regression analysis, this study constructed

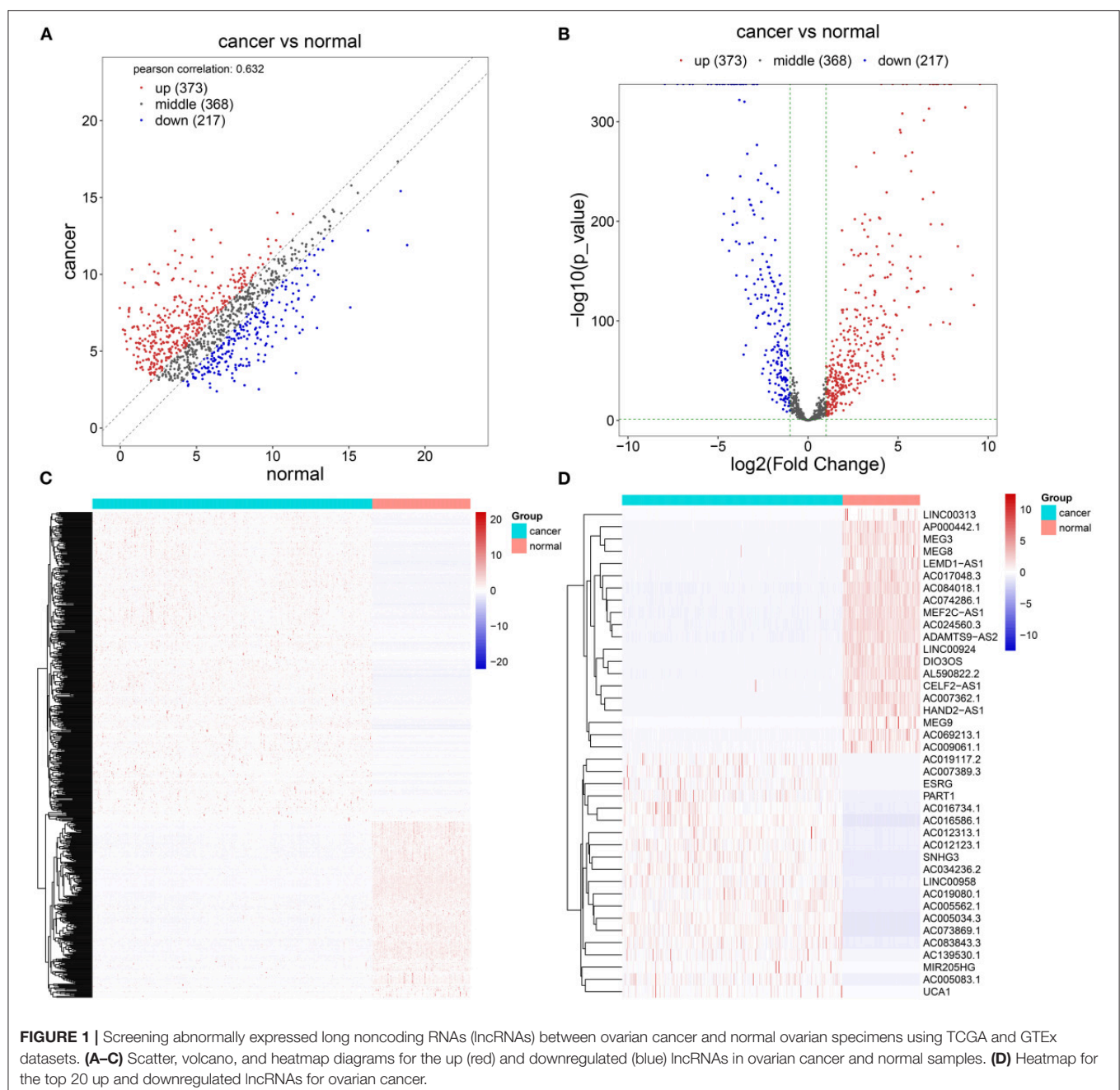
Abbreviations: lncRNAs, long non-coding RNAs; TCGA, The Cancer Genome Atlas; GTEx, Genotype-Tissue Expression; FC, fold change; LASSO, least absolute shrinkage and selection operator; OS, overall survival; HR, hazard ratio; ROC, receiver operating characteristic; ESTIMATE, Estimation of STromal and Immune cells in Malignant Tumor tissues using Expression data.

an autophagy-related lncRNA model utilizing the glmnet package (19). The association between the lncRNAs in this model and overall survival (OS) was assessed by univariate Cox regression analysis. lncRNAs with hazard ratio (HR) > 1 and $p < 0.05$ were risk factors, while those with HR < 1 and $p < 0.05$ were protective factors. The risk score of each sample was determined. The formula was as follows: risk score = coefficient (lncRNA1) × expression (lncRNA1) + coefficient (lncRNA2) × expression (lncRNA2) + ... + coefficient (lncRNA n) × expression (lncRNA n). The distributions of risk scores, survival

status, and disease progress were assessed in the samples of ovarian cancer.

Assessment of the Prognostic Model

All the subjects with ovarian cancer were randomly separated into the training set and the validation set at a ratio of 1:1. The subjects with ovarian cancer were separated into high- and low-risk groups. The Kaplan–Meier survival analyses were carried out to assess the OS differences between high- and low-risk groups with survival package in the training set, validation set, and the whole dataset. Time-dependent receiver operating



characteristic (ROCs) curves were depicted for estimating the predictive efficacy of survival time through risk score and other clinical features (age, pathologic stage, and histologic grade) utilizing the survival ROC package. Furthermore, the differences in the risk scores between patients with different

histologic grades or pathologic stages were analyzed by one-way ANOVA. Multivariate Cox regression analysis was presented for evaluating whether the risk score could be independent of other clinical features including pathologic stage and histologic grade.

TABLE 1 | The top 20 upregulated lncRNAs in ovarian cancer.

Gene name	log ₂ FC	P-value	Q-value	Cancer	Normal
AC012313.1	9.547183939	0	0	10.31895834	0.771774397
UCA1	9.216655018	1.2272E-116	6.3207E-116	12.81661174	3.59995672
AC007389.3	9.147273622	1.8891E-146	1.321E-145	9.442887756	0.295614134
LINC00958	8.733161562	0	0	10.64514913	1.911987563
AC139530.1	8.320256095	1.2565E-175	1.0943E-174	9.16409823	0.843842135
AC019117.2	7.940574318	1.5611E-132	9.4059E-132	10.17086201	2.230287692
AC073869.1	7.93835612	0	0	11.53335685	3.595000733
ESRG	7.88342569	1.29304E-97	5.43305E-97	9.509050129	1.625624439
AC034236.2	7.840261744	0	0	7.803144412	-0.037117331
AC012123.1	7.575128788	0	0	10.06414233	2.489013539
MIR205HG	7.508354236	2.9168E-99	1.25304E-98	10.64232063	3.133966398
PART1	7.450907876	1.1583E-197	1.1681E-196	12.24750023	4.796592357
AC005562.1	7.220023466	0	0	7.809237126	0.58921366
AC016586.1	7.200189856	0	0	11.25391667	4.053726816
AC016734.1	7.136519504	0	0	7.696281571	0.559762067
AC083843.3	7.119307279	0	0	9.449229653	2.329922374
AC019080.1	6.989829409	0	0	7.558546463	0.568717055
AC005034.3	6.970382156	0	0	9.402961692	2.432579536
AC005083.1	6.968474931	9.7814E-230	1.2628E-228	9.411351502	2.442876571
SNHG3	6.926312525	0	0	12.89759844	5.971285912

TABLE 2 | The 20 downregulated lncRNAs in ovarian cancer.

Gene name	log ₂ FC	P-value	Q-value	Cancer	Normal
AC007362.1	-7.953839402	0	0	3.573056805	11.52689621
DIO3OS	-7.25010061	0	0	7.8433415	15.09344211
MEG3	-6.926193554	0	0	11.89962972	18.82582328
LINC00313	-6.572300069	0	0	2.518410795	9.090710864
HAND2-AS1	-6.400004123	0	0	6.508966669	12.90897079
MEG9	-5.584769193	5.2013E-247	7.3277E-246	6.183950367	11.76871956
AC024560.3	-5.562626374	0	0	6.396148069	11.95877444
AC074286.1	-5.375749227	0	0	2.776822389	8.152571616
AL590822.2	-5.097997242	0	0	4.32297655	9.420973791
LINC00924	-5.04563661	0	0	5.818215591	10.8638522
ADAMTS9-AS2	-4.904513345	0	0	6.606317799	11.51083114
MEG8	-4.756706248	5.3224E-182	4.9503E-181	4.749036786	9.505743034
LEMD1-AS1	-4.748134967	0	0	7.224444409	11.97257938
AC017048.3	-4.676146545	3.2389E-208	3.6939E-207	5.295737783	9.971884328
AC069213.1	-4.643176366	0	0	4.407805217	9.050981583
AP000442.1	-4.428047311	0	0	3.056508739	7.48455605
CELF2-AS1	-4.38133152	5.0736E-171	4.3013E-170	4.075613874	8.456945394
MEF2C-AS1	-4.276829555	0	0	4.792378869	9.069208425
AC009061.1	-4.255822376	0	0	4.029896192	8.285718568
AC084018.1	-4.237903564	0	0	8.332160874	12.57006444

CIBERSORT

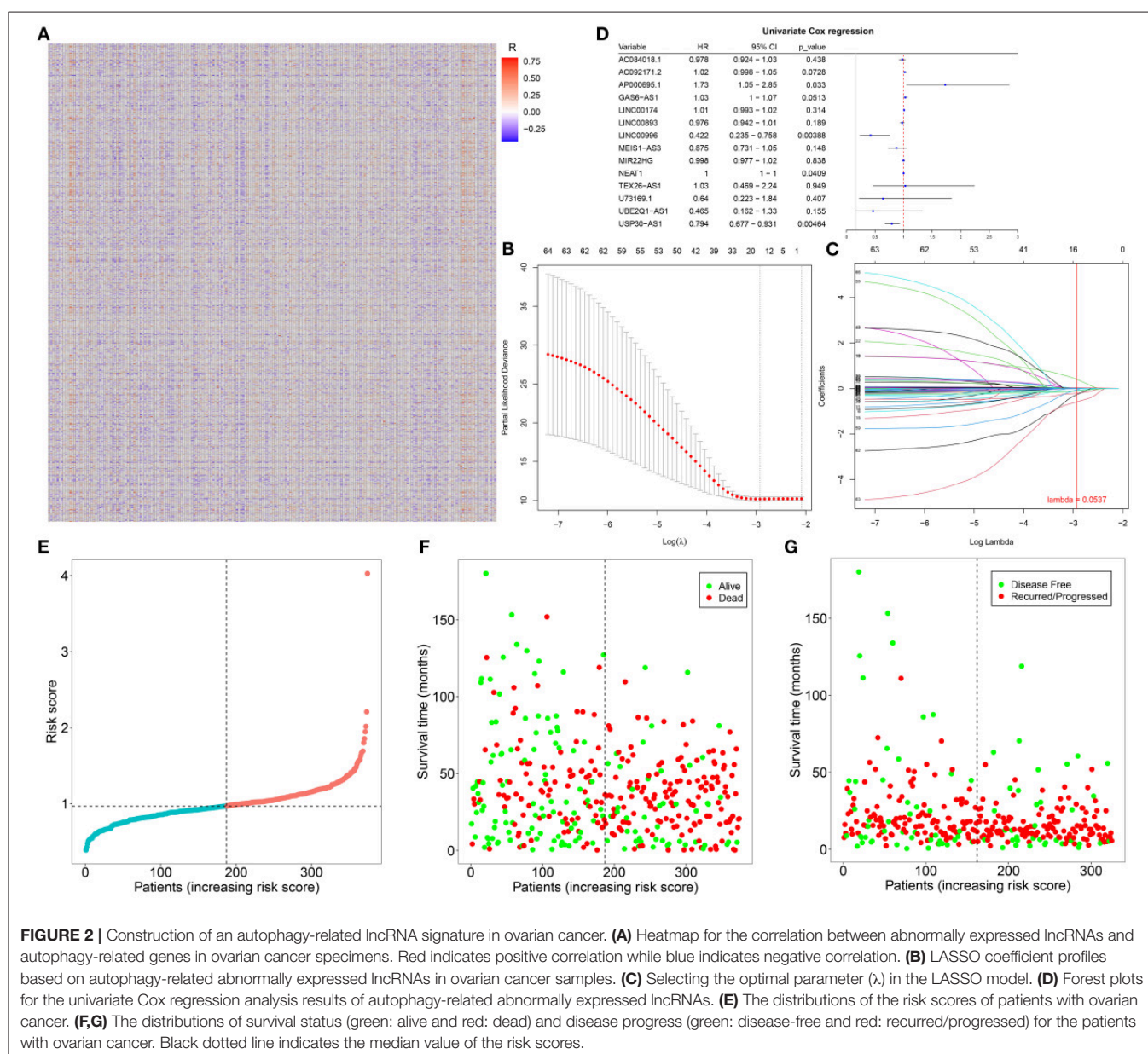
CIBERSORT algorithm (<http://cibersort.stanford.edu/>) can characterize cell compositions of complex tissues based on the gene expression profiles (20). This algorithm is superior to other methods in terms of noise, unknown mixture content, and closely related cell types. CIBERSORT algorithm was employed to characterize immune cell compositions (including B cells naïve, B cells memory, plasma cells, T cells CD8, T cells CD4, naïve T cells, CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, T cells regulatory (Tregs), T cells gamma delta, NK cells resting, NK cells activated, monocytes, macrophages M0, macrophages M1, macrophages M2, dendritic cells resting, dendritic cells activated, mast cells resting, mast cells activated, eosinophils and neutrophils) in ovarian cancer tissues.

Estimation of Stromal and Immune Cells in Malignant Tumor Tissues Using Expression Data

The fractions of stromal and immune cells were inferred in each ovarian cancer sample using the ESTIMATE algorithm (<https://sourceforge.net/projects/estimateproject/>) (21). The differences in immune and stromal scores were assessed between high- and low-risk samples of ovarian cancer by Student's *t*-test.

Sensitivity to Chemotherapy Drugs

A total of 94 ovarian cancer-related chemotherapy drugs were collected from the Cancer Genome Project (CGP; version 2014). The pRRophetic package was employed to predict the clinical chemotherapeutic responses based on the gene expression



profiles of ovarian cancer (22). The IC50 values of chemotherapy drugs were compared between high- and low-risk groups utilizing Student's *t*-test.

Somatic Mutation Analysis

The “Masked Somatic Mutation” data were obtained and processed with VarScan software (version 2), a method for the detection of somatic mutation and copy number variation in exome data (23). The maftools package (24) was adopted to analyze the MAF of the somatic variants. The “titv” function classified single nucleotide polymorphisms (SNPs) into transitions (Ti) and transversions (Tv). The overall distribution of the six different SNVs and the proportion of transitions in each sample were evaluated. The oncoplot of the top 30 mutated genes was depicted by the “ComplexHeatmap” function. The “somaticInteractions” function was used to detect mutually exclusive or co-occurring genomes, which were tested by Fisher's exact test.

Gene Set Enrichment Analysis

Pathways underlying the autophagy-related lncRNA signature were evaluated through GSEA package (25). The “c5.bp.v6.2.symbols.gm” gene set was curated from the Molecular Signatures Database, which was employed as a reference set. Terms with $|\text{nominal enrichment score (NES)}| > 1.7$ and nominal $p < 0.05$ were significantly enriched.

TABLE 3 | A total of 66 autophagy-related abnormally expressed lncRNAs.

LncRNAs	LncRNAs	LncRNAs
AC005387.2	C9orf163	LINC00996
AC006538.1	CACTIN-AS1	MAP3K14-AS1
AC007292.3	CCDC39	MEG3
AC007382.1	CDKN2B-AS1	MEIS1-AS3
AC009093.1	DHRS4-AS1	MIR22HG
AC010336.1	EHMT2-AS1	MKNK1-AS1
AC073869.1	ERVK13-1	NARF-IT1
AC084018.1	EXTL3-AS1	NEAT1
AC092171.2	FAM13A-AS1	PCED1B-AS1
AC105020.1	FLNB-AS1	PRKQ-AS1
AC137932.1	GAS6-AS1	RNF216P1
ACTA2-AS1	HCP5	SH3BP5-AS1
AL136115.1	LINC00174	SPANXA2-OT1
AL137127.1	LINC00265	SRGAP3-AS2
AL590822.2	LINC00342	TEX26-AS1
AP000695.1	LINC00624	TNK2-AS1
AP005482.1	LINC00677	TTL10-AS1
AP006621.1	LINC00702	U73169.1
AP4B1-AS1	LINC00893	UBE2Q1-AS1
ARAP1-AS2	LINC00894	USP30-AS1
C10orf55	LINC00954	YEATS2-AS1
C1orf195	LINC00968	ZSWIM8-AS1

RESULTS

Screening Abnormally Expressed lncRNAs

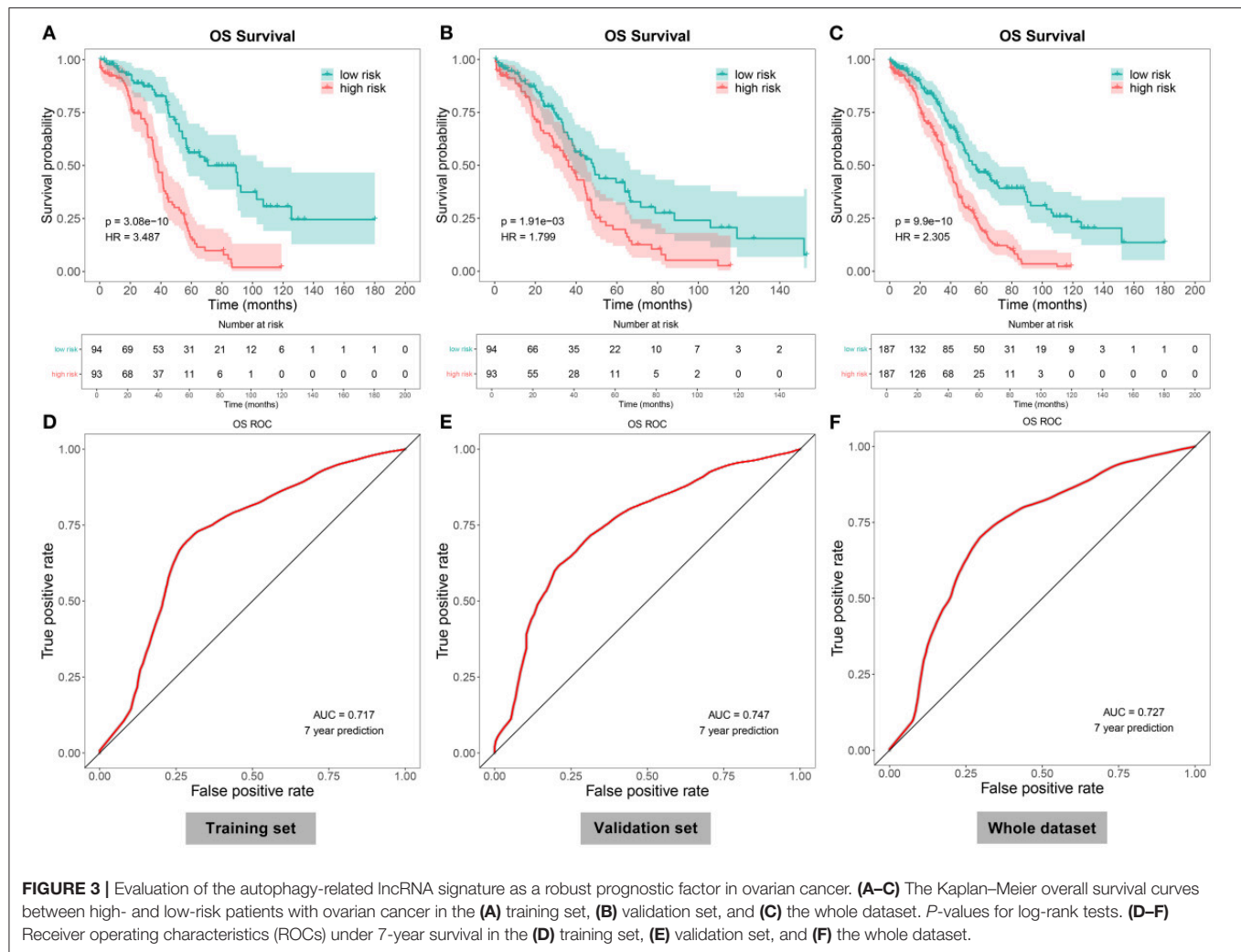
Herein, we retrieved RNA expression profiles of ovarian cancer and normal ovarian tissues from TCGA and GTEx datasets. Totally, 590 lncRNAs with adjusted $p \leq 0.05$ and $|\log_2\text{FC}| \geq 1$ were abnormally expressed in 379 specimens of ovarian cancer in comparison with the normal ovarian cancer (Figures 1A–C). There were 373 upregulated (Supplementary Table 1) and 217 downregulated (Supplementary Table 2) lncRNAs for ovarian cancer. We separately displayed the top 20 upregulated (Table 1) and downregulated (Table 2) lncRNAs for ovarian cancer (Figure 1D).

Construction of an Autophagy-Related lncRNA Prognostic Signature in Ovarian Cancer

The correlation between abnormally expressed lncRNAs and autophagy-related genes was evaluated by Pearson's correlation analysis, which was visualized into the heat map (Figure 2A; Supplementary Table 3). With the criteria values of $|\text{correlation coefficient}| > 0.4$ and $p < 0.001$, we selected 66 autophagy-related abnormally expressed lncRNAs, as shown in Table 3. Using LASSO regression analysis, we determined the regression coefficients of the 14 lncRNAs in the model (Figures 2B,C). The risk score of each specimen of ovarian cancer was calculated as follows: AC084018.1 expression * (−0.008315282) + AC092171.2 expression * 2.84E−05 + AP000695.1 * 0.395692194 + GAS6-AS1 expression * 0.028077942 + LINC00174 expression * 0.010226196 + LINC00893 expression * (−0.013179597) + LINC00996 expression * (−0.214479586) + MEIS1-AS3 expression * (−0.093561265) + MIR22HG expression * (−0.007661823) + NEAT1 expression * (0.000320475) + TEX26-AS1 expression * (−0.061717576) + U73169.1 expression * (−0.263275905) + UBE2Q1-AS1 expression * (−0.589381398) + USP30-AS1 expression * (−0.114206463). Univariate Cox regression analysis results showed that AP000695.1 (HR: 1.73, 95%

TABLE 4 | The clinical characteristics of patients with ovarian cancer.

Variables	The whole set (N = 374)	Training set (N = 187)	Validation set (N = 187)
Age	59.54 ± 11.4	60.34 ± 10.63	58.74 ± 12.09
Status			
Alive	145 (38.77)	73 (39.04)	72 (38.5)
Dead	229 (61.23)	114 (60.96)	115 (61.5)
Pathologic stage			
Stage I	1 (0.27)	0 (0)	1 (0.53)
Stage II	22 (5.88)	9 (4.81)	13 (6.95)
Stage III	291 (77.81)	139 (74.33)	152 (81.28)
Stage IV	57 (15.24)	36 (19.25)	21 (11.23)
Unknow	3 (0.8)	3 (1.6)	0 (0)



CI: 1.05–2.85, *p*-value: 0.033) was a risk factor of ovarian cancer while LINC00996 (HR: 0.422, 95% CI: 0.235–0.758, *p*-value: 0.00388) and USP30-AS1 (HR: 0.794, 95% CI: 0.677–0.931, *p*-value: 0.00464) were protective factors of ovarian cancer (Figure 2D). In Figure 2E, we visualized the distributions of the risk scores of the patients with ovarian cancer. Also, we found that the risk scores displayed associations with survival status (Figure 2F) and disease progression (Figure 2G).

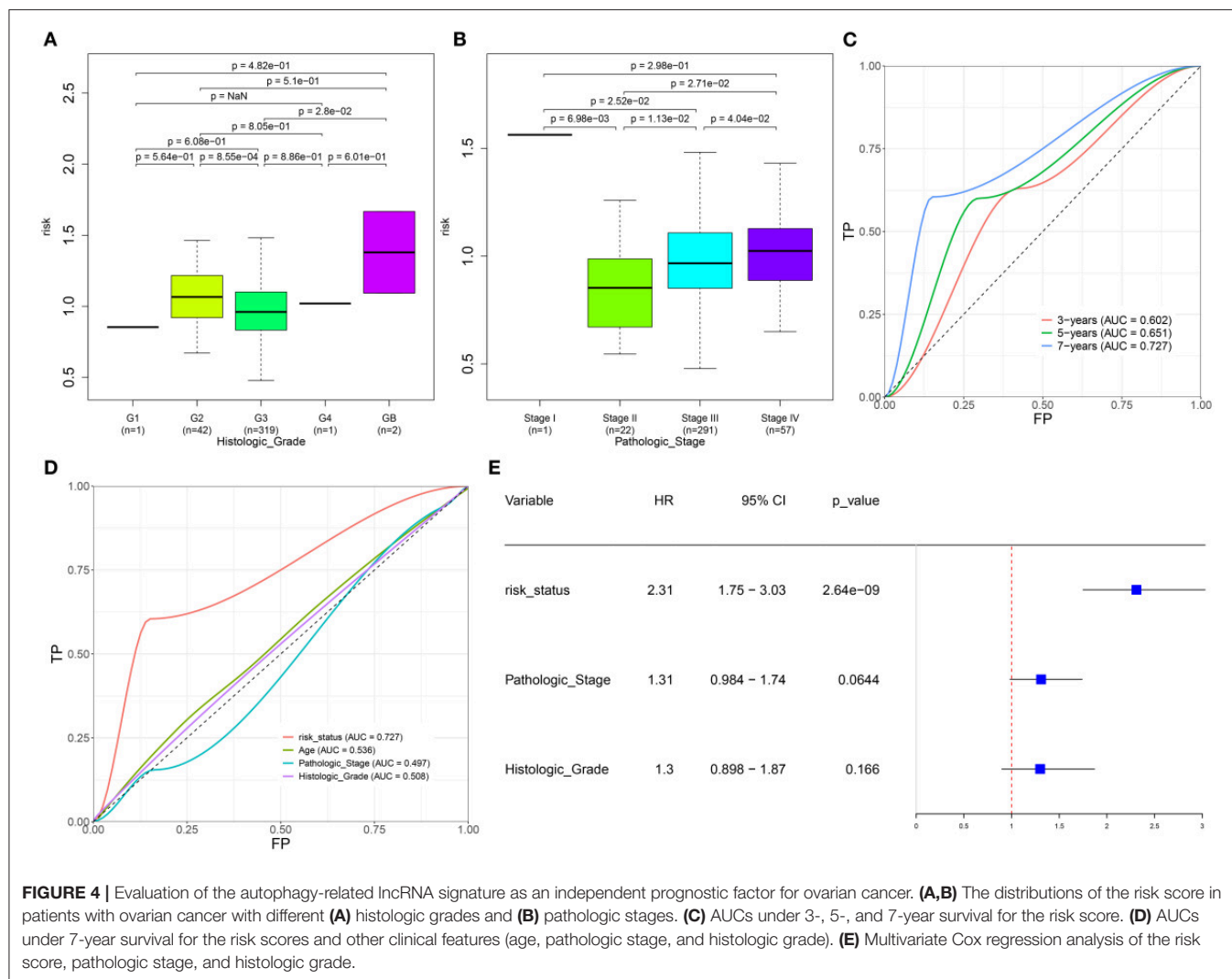
Evaluation of the Autophagy-Related lncRNA Signature as a Robust Prognostic Factor in Ovarian Cancer

Herein, all the patients with ovarian cancer were randomly separated into the training set and validation set (both *n* = 187) in Table 4. Based on the median value of the risk scores, the patients were divided into high- and low-risk groups. Our data showed that there were distinct differences in OS time between the two groups in the training set (*p* = 3.08e-10), validation set (*p* = 1.91e-03), and the whole dataset (*p* =

9.9e-10; Figures 3A–C). Patients in the low-risk group had prolonged OS duration in comparison with those in the high-risk group. ROCs were conducted to validate the predictive performance of this signature. The area under the curves under 7-year survival were 0.717, 0.747, and 0.727 in the training set, validation set, and the whole dataset, respectively (Figures 3D–F).

The Autophagy-Related lncRNA Signature as an Independent Prognostic Factor for Ovarian Cancer

We further evaluated the correlations between the risk scores and other clinical features in specimens with ovarian cancer. As a result, lowered risk scores were detected in grade 3 than grade 2 (*p* = 8.55e-04; Figure 4A). Meanwhile, we found that the risk scores increased gradually as the pathologic stages increased (Figure 4B). The AUCs under 3-, 5-, and 7-year survival time were 0.602, 0.651, and 0.727 for the patients with ovarian cancer, indicating the well-predictive performance of this signature (Figure 4C). Furthermore, compared with the



other clinical characteristics including age (AUC = 0.536), pathologic stage (AUC = 0.497), and histologic grade (AUC = 0.508), there was a higher AUC value under 7-year OS time for the risk score (Figure 4D). The data suggested that this risk score might possess higher sensitivity and accuracy in predicting the prognosis of the patients with ovarian cancer. Multivariate Cox regression analysis demonstrated that this risk score could be independently predictive of the prognosis of the patients (HR: 2.31, 95% CI: 1.75–3.03, $p = 2.64 \times 10^{-9}$; Figure 4E).

The Autophagy-Related lncRNA Signature Is Associated With Immune Cell Infiltration in Ovarian Cancer

Here, we adopted the CIBERSORT algorithm to infer the immune cell infiltration in samples with ovarian cancer. Figure 5A visualized the proportions of 22 kinds of immune cell components in ovarian cancer tissues. Also, we analyzed the

correlations between distinct immune cells. In Figure 5B, there were strong correlations between B cells *naïve* and macrophages M2 ($r = 0.93$), between B cells memory and monocytes ($r = 0.98$), between T cells CD4 *naïve* and T cells CD4 memory resting ($r = 0.98$), between T cells follicular helper and Tregs ($r = 0.97$), between T cells follicular helper and T cells gamma delta ($r = 0.96$), between Tregs and macrophages M0 ($r = 0.97$), between T cells gamma delta and mast cells activated ($r = 0.92$), and between monocytes and mast cells activated ($r = 0.95$) in ovarian cancer tissues. Also, heatmap visualized the differences in the immune cell infiltrations between high- and low-risk samples of the ovarian cancer (Figure 5C). Compared with the low-risk group, there were lowered infiltration levels of macrophages M1 ($p < 0.0001$), mast cells resting ($p = 0.007$), plasma cells ($p = 0.003$), T-cell CD8 ($p = 0.016$), and T-cell follicular helper ($p = 0.001$) and also higher infiltration levels of macrophages M2 ($p = 0.003$), mast cells activated ($p < 0.0001$), and neutrophils ($p = 0.036$) in the high-risk group (Figure 5D).

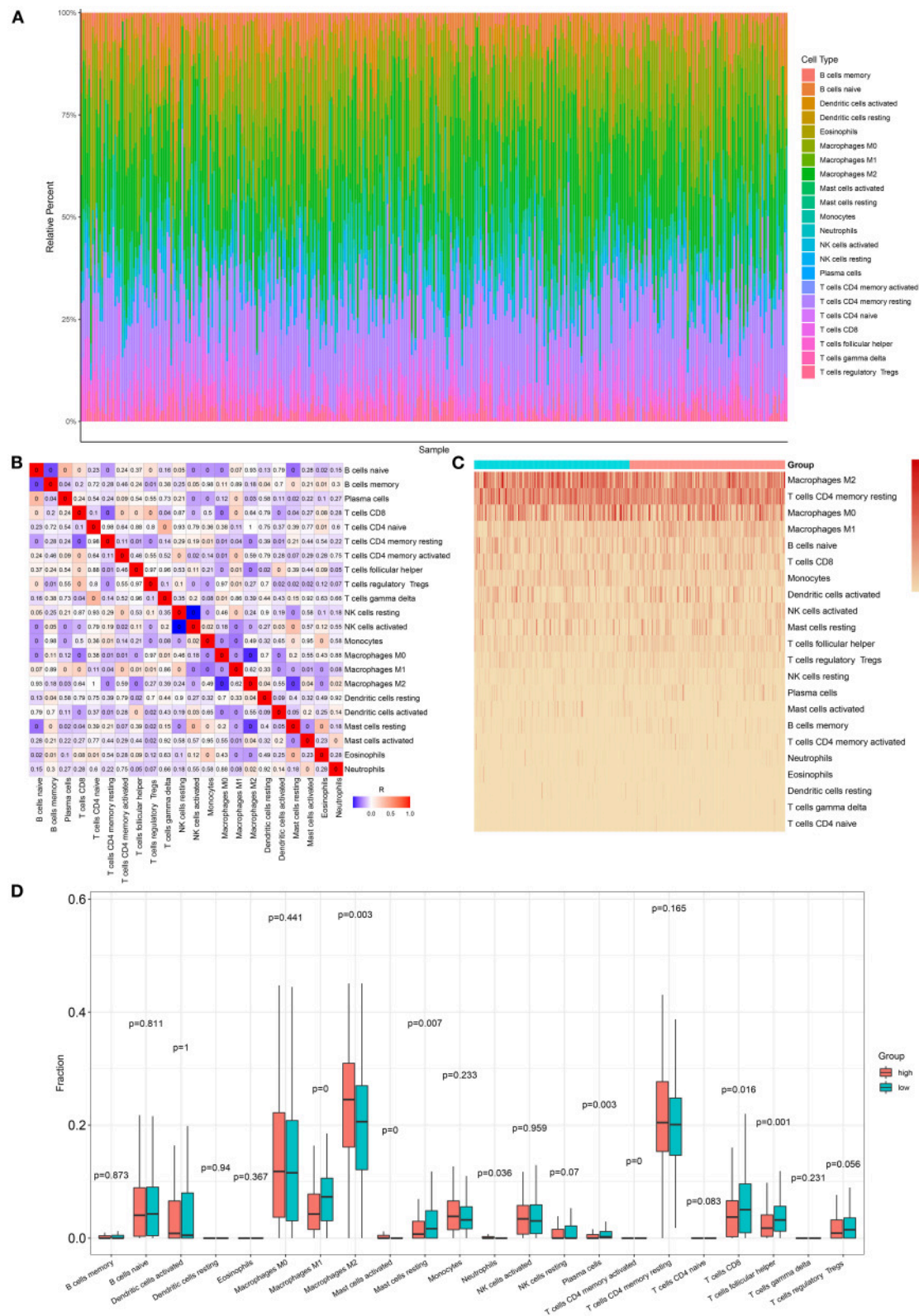
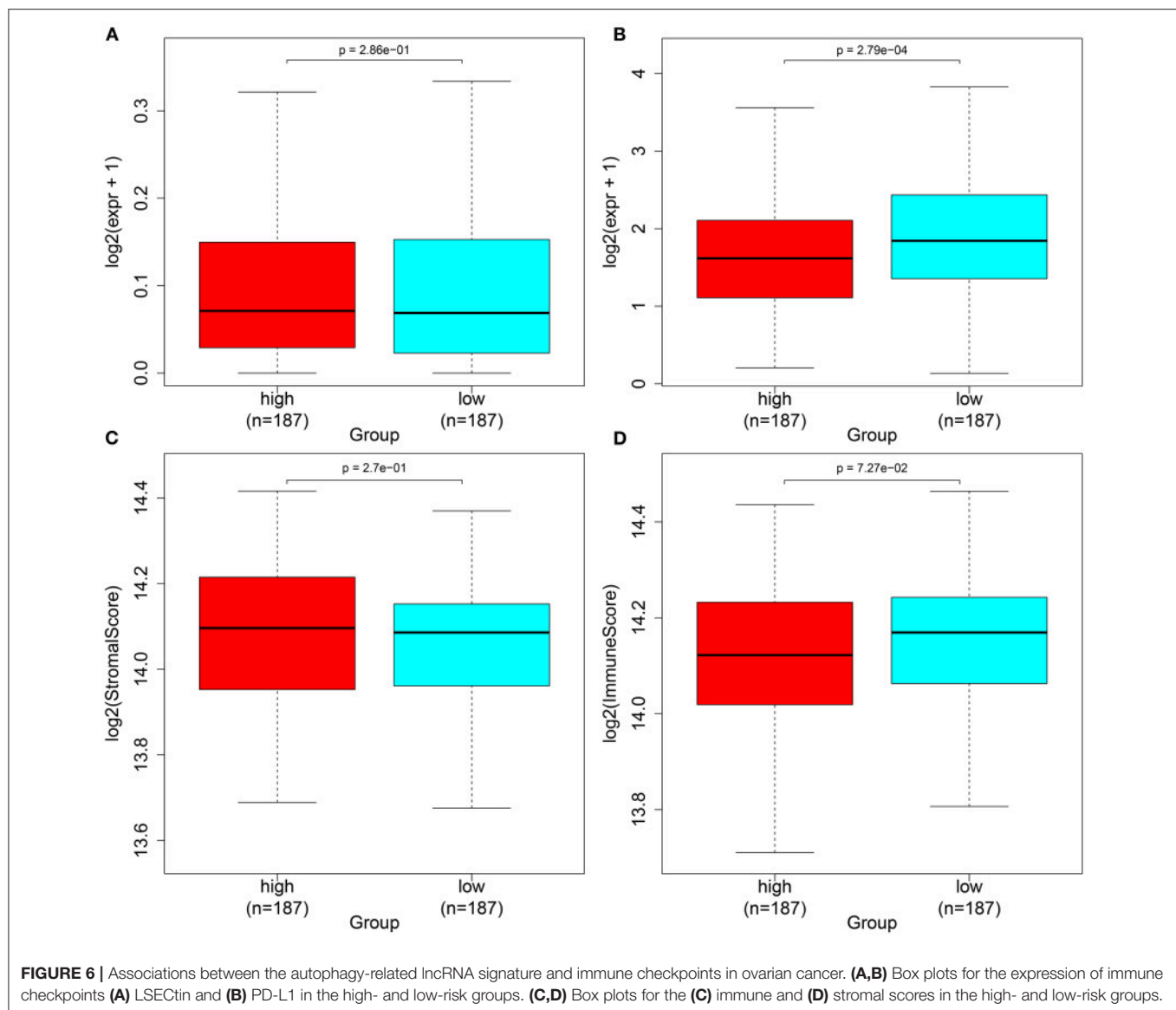


FIGURE 5 | Association between the autophagy-related lncRNA signature and immune cell infiltration in ovarian cancer. **(A)** The proportions of different immune cell components in ovarian cancer tissues. **(B)** Heatmap for the correlations between different immune cells. The darker the color, the greater the [correlation coefficient]. Red: positive correlation and blue: negative correlation. **(C)** Heatmap for the infiltration levels of immune cells in the high- and low-risk samples with ovarian cancer. **(D)** Box plots for the differences in infiltration levels of immune cells between the high- and low-risk samples with ovarian cancer.

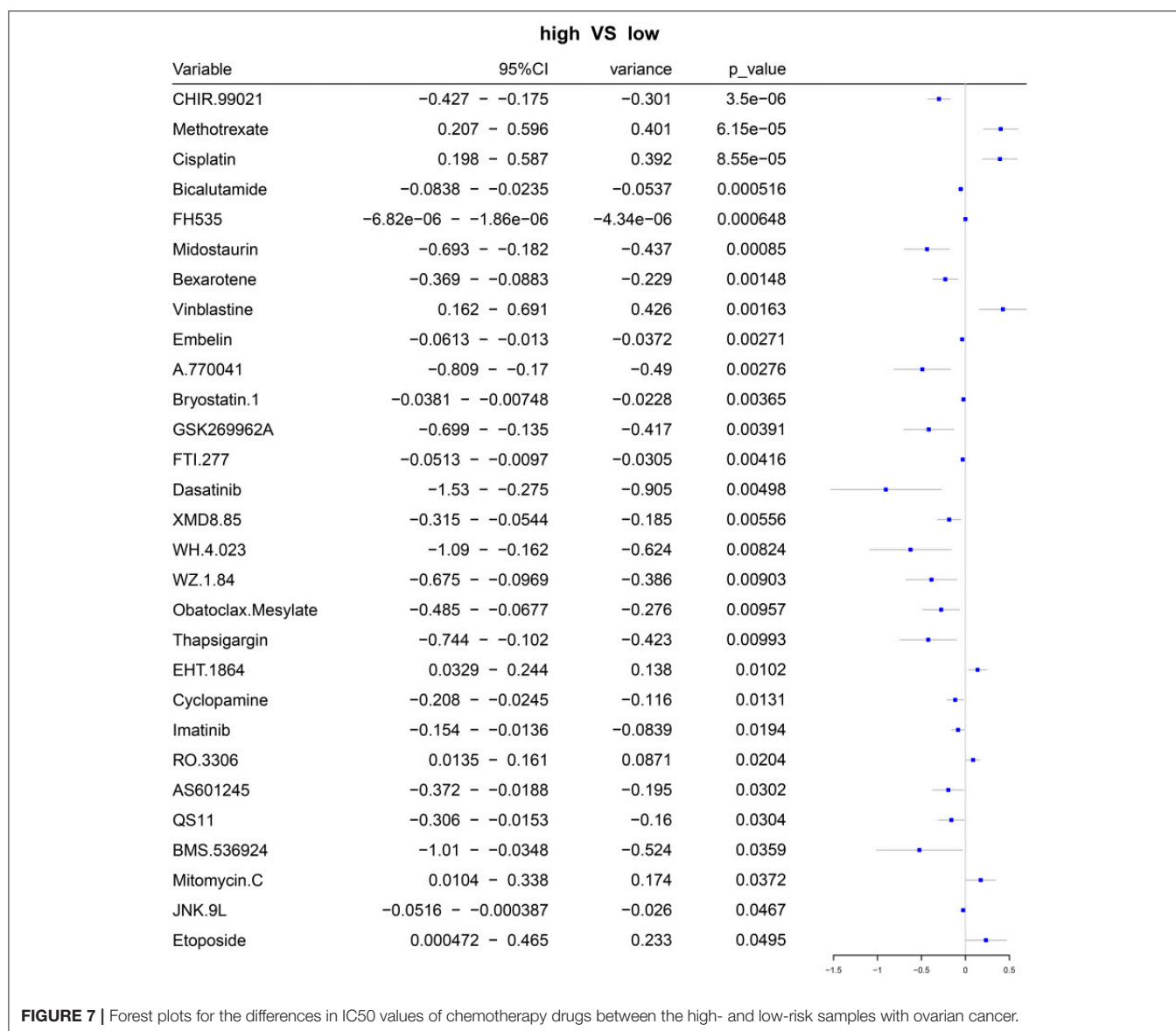


The Autophagy-Related lncRNA Signature Is Associated With Immune Checkpoints in Ovarian Cancer

Herein, we observed whether the autophagy-related lncRNA signature was associated with the expression of immune checkpoints in ovarian cancer tissues. Higher LSECtin expression was found in the high-risk group compared with the low-risk group (**Figure 6A**). Furthermore, there was distinctly decreased PD-L1 expression in the low-risk group compared with the high-risk group ($p = 2.79e-04$; **Figure 6B**). We also assessed the correlations between the risk scores and immune or stromal scores. As a result, higher immune scores and lower stromal scores were detected in the high-risk samples compared with the low-risk samples, which were not statistically significant (**Figures 6C,D**).

Prediction of the Sensitivity to Chemotherapy Drugs Based on the Autophagy-Related lncRNA Signature

Based on the gene expression profiles, we predicted the responses to 94 chemotherapy drugs in each patient with ovarian cancer. Among them, there were significant differences in responses to 29 chemotherapy drugs with $p < 0.05$ between high- and low-risk patients with ovarian cancer, including CHIR.99021, methotrexate, cisplatin, bicalutamide, FH535, midostaurin, bexarotene, vinblastine, embelin, A.770041, bryostatins.1, GSK269962A, FTI.277, dasatinib, XMD8.85, WH.4.023, WZ.1.84, Obatoclax.Mesylate, thapsigargin, EHT.1864, cyclopamine, imatinib, RO.3306, AS601245, QS11, BMS.536924, mitomycin.C, JNK.9L, and etoposide (**Figure 7; Supplementary Table 4**).



Somatic Mutation Landscapes in Ovarian Cancer

We further analyzed the somatic mutation landscapes in specimens of ovarian cancer. Both in the high- (Figure 8A) and low-risk (Figure 8B) samples, missense mutation was the most common mutation type. TP53 was the top-ranked mutated gene, followed by TTN. C > T mutation had the highest proportion in the two groups (Figures 8C,D). Furthermore, potential druggable categories targeted mutated genes were predicted, such as the druggable genome, clinically actionable, and kinase in the high- (Figure 8E) and low-risk (Figure 8F) samples. Approximately 98.48% of samples occurred genetic mutations in the high-risk samples (Figure 9A) and 99.28% of the occurred mutations in the low-risk samples (Figure 9B). Many genes co-occur in cancer or show strong exclusivity in their mutation patterns. Here, we visualized the close interactions

between the mutated genes in the high- and low-risk samples (Figures 9C,D).

Significant Signaling Pathways Underlying the Autophagy-Related lncRNA Signature

To explore the significant signaling pathways underlying the autophagy-related lncRNA signature, this study carried out GSEA by comparing high- and low-risk groups. As shown in Figure 10, we observed that ribosome (NES = 1.90 and nominal $p = 0.019$), oxidative phosphorylation (NES = 1.76 and nominal $p = 0.037$), and Parkinson's disease (NES = 1.74 and nominal $p = 0.035$) were markedly activated in the high-risk group. Meanwhile, various types of N-glycan biosynthesis (NES = -1.84 and nominal $p = 0.002$), lysosome (NES = -1.87 and nominal $p = 0.002$), and circadian rhythm (NES = -2.08 and nominal $p < 0.0001$) were enriched significantly in the low-risk group.

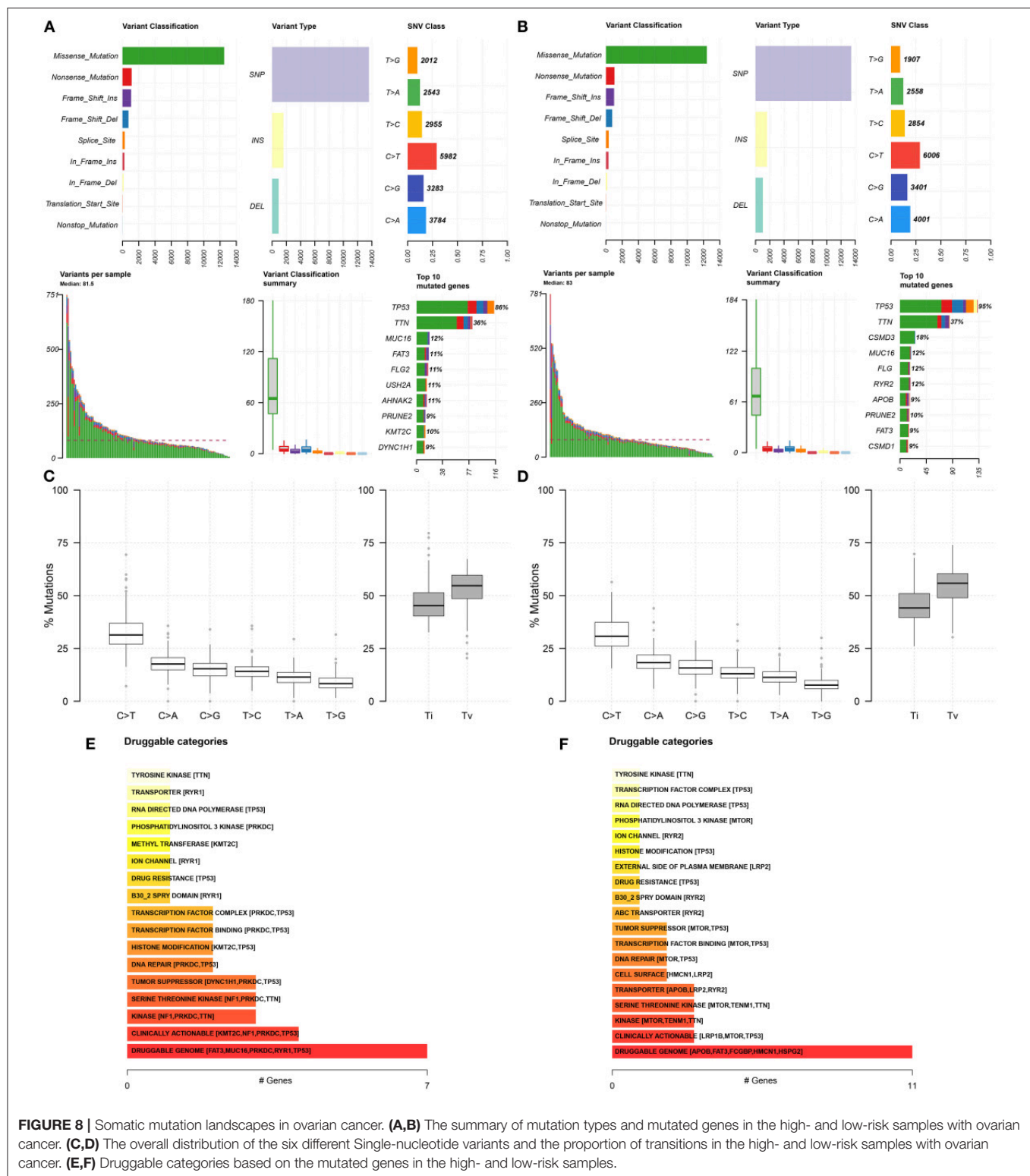


FIGURE 8 | Somatic mutation landscapes in ovarian cancer. **(A,B)** The summary of mutation types and mutated genes in the high- and low-risk samples with ovarian cancer. **(C,D)** The overall distribution of the six different Single-nucleotide variants and the proportion of transitions in the high- and low-risk samples with ovarian cancer. **(E,F)** Druggable categories based on the mutated genes in the high- and low-risk samples.

DISCUSSION

As per a previous study, 17 autophagy-related lncRNAs have been identified as prognostic predictors of ovarian cancer

(26). Nevertheless, a single autophagy-related lncRNA often has limited predictive power. Furthermore, the established clinical prognostic biomarkers have limited accuracy and specificity. It has been confirmed that gene models exhibit higher

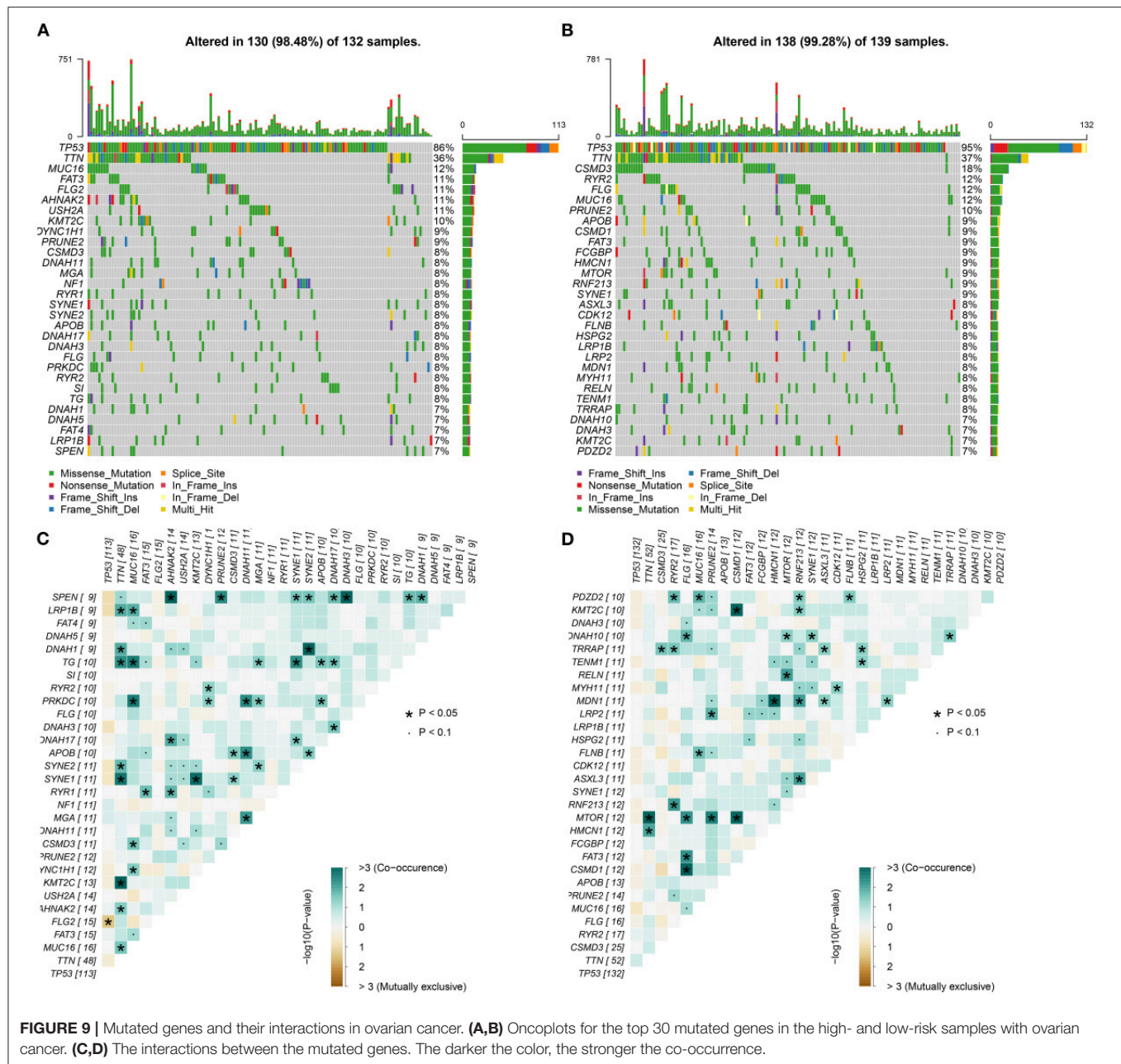


FIGURE 9 | Mutated genes and their interactions in ovarian cancer. **(A,B)** Oncoplots for the top 30 mutated genes in the high- and low-risk samples with ovarian cancer. **(C,D)** The interactions between the mutated genes. The darker the color, the stronger the co-occurrence.

predictive power than a single gene (5). This study employed the LASSO regression method to establish an autophagy-related lncRNA signature containing AC084018.1, AC092171.2, AP000695.1, GAS6-AS1, LINC00174, LINC00893, LINC00996, MEIS1-AS3, MIR22HG, NEAT1, TEX26-AS1, U73169.1, UBE2Q1-AS1, and USP30-AS1 for ovarian cancer based on abnormally expressed autophagy-related lncRNA profiles. After verification, this signature robustly and independently predicted the survival outcomes of the patients with ovarian cancer. Among the 14 lncRNAs in this signature, GAS6-AS1 exerts a carcinogenic effect in the breast cancer (27) and hepatocellular carcinoma (28). LINC00174 promotes glioma

progression *via* miR-152-3p/SLC2A1 axis (29). LINC00893 suppresses papillary thyroid cancer by inactivating the AKT pathway and stabilizing PTEN (30). USP30-AS1 expression is related to autophagy in the bladder urothelial carcinoma (31). More experiments require to verify the regulatory roles of these lncRNAs on autophagy.

As immunogenic cancers, the spontaneous anticancer immune responses may increase survival duration, and the immune escape may cut down survival duration (32). In the recent years, several immune-based strategies such as immune checkpoint inhibition, vaccination, and antigen-specific active immunotherapy have been developed in ovarian

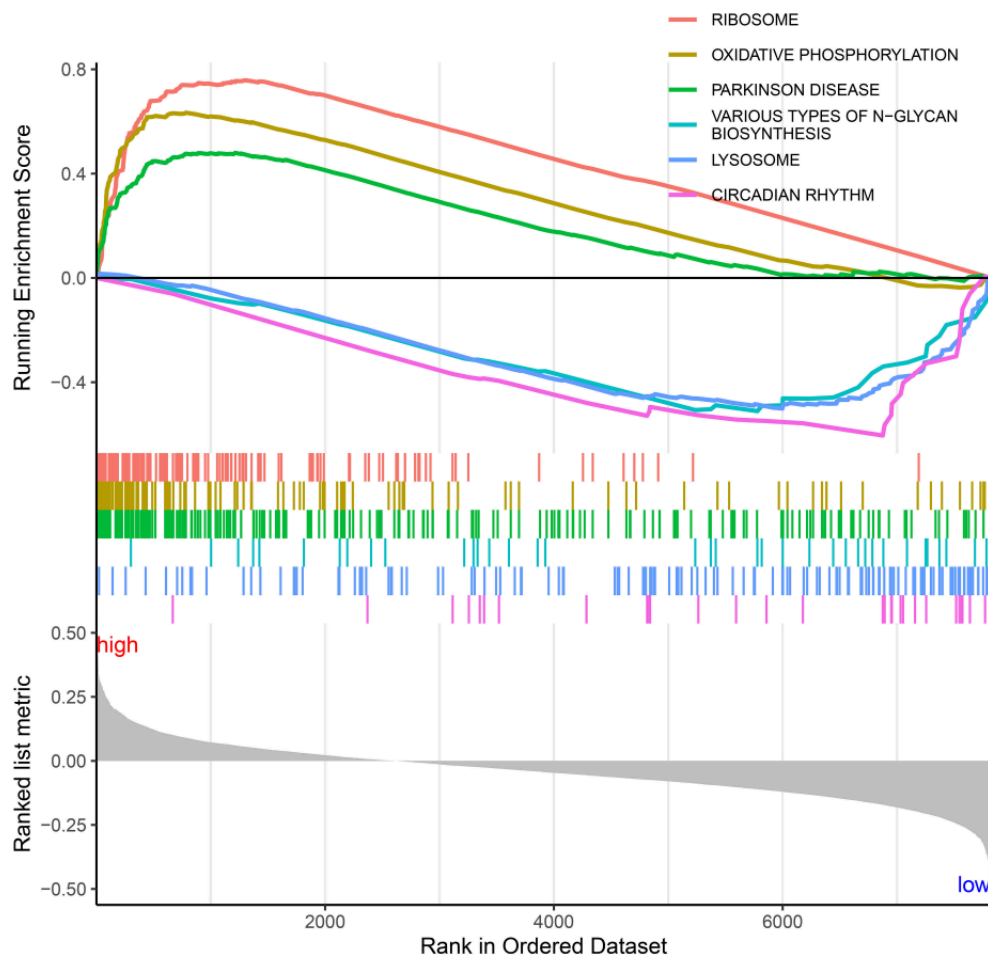


FIGURE 10 | Significant signaling pathways underlying the autophagy-related lncRNA signature by Gene set enrichment analysis.

cancer (33). The immune-suppressive networks in the tumor microenvironment have been considered for immunotherapy implementation (34). The immune-related markers may be utilized for predicting the responses to immunotherapy (35). Hence, the interactions between molecules and tumor microenvironment require in-depth exploration. Our study demonstrated that the autophagy-related lncRNA signature was in relation to infiltrations of macrophages M1, mast cells resting, plasma cells, T-cell CD8, T-cell follicular helper, macrophages M2, mast cells activated, and neutrophils, indicating the crosstalk between these autophagy-related lncRNAs and immune cells. Consistently, Deng et al. identified a novel autophagy-related lncRNA model that was distinctly related to the infiltrations of immune cells in pancreatic cancer (36). Ovarian cancer represents a low immune-reactive malignancy with restricted immune cell infiltrations and extensive immunosuppressive T cell infiltrations (37). Tumors positive for PD-L1 usually exhibit a higher response to immune checkpoint inhibition therapy, and highly expressed PD-L1 is a predictor of undesirable clinical outcomes (37). Here, our data showed that the low-risk ovarian cancer patients

had higher PD-L1 expression in comparison to those with high risk.

Chemotherapy (platinum and taxanes) plays a fundamental role in adjuvant therapy against ovarian cancer (38). Despite the initial response to this therapy, most of the patients diagnosed with ovarian cancer developed chemotherapy resistance (39). This resistance may be driven by a range of mechanisms. Hence, it is of importance to develop individual markers for the prediction of the sensitivity to chemotherapy. Our data showed that the risk scores were in relation to 29 chemotherapy drugs such as cisplatin, indicating that the autophagy-related lncRNAs were involved in chemotherapy sensitivity, as a previous study (11).

CONCLUSIONS

Collectively, this study provided a knowledge base of novel autophagy-related lncRNAs in ovarian cancer, improving the understanding of the functions of lncRNA on the regulation

of autophagy in ovarian cancer. We established an autophagy-related lncRNA signature as a robust prognostic marker for the prediction of survival outcomes, immunotherapy response, and chemotherapy sensitivity. Our findings may assist to precisely guide therapeutic strategies for individual patients with ovarian cancer in clinical practice.

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

JiW conceived and designed the study. YL and JuW conducted most of the experiments and data

analysis and wrote the manuscript. FW, CG, and YC participated in collecting data and helped to draft the manuscript. All authors reviewed and approved the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.715250/full#supplementary-material>

Supplementary Table 1 | The list of upregulated lncRNAs in ovarian cancer tissues.

Supplementary Table 2 | The list of downregulated lncRNAs in ovarian cancer tissues.

Supplementary Table 3 | The correlation analysis results between abnormally expressed lncRNAs and autophagy-related genes.

Supplementary Table 4 | The list of chemotherapy drugs.

REFERENCES

- Herrera FG, Irving M, Kandalaft LE, Coukos G. Rational combinations of immunotherapy with radiotherapy in ovarian cancer. *Lancet Oncol.* (2019) 20:e417–e33. doi: 10.1016/S1470-2045(19)30401-2
- Das T, Anand U, Pandey SK, Ashby CR, Jr., Assaraf YG, et al. Therapeutic strategies to overcome taxane resistance in cancer. *Drug Resist Updat.* (2021) 55:100754. doi: 10.1016/j.drug.2021.100754
- Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. *Genome Med.* (2020) 12:8. doi: 10.1186/s13073-019-0703-1
- Yang H, Gao L, Zhang M, Ning N, Wang Y, Wu D, et al. Identification and analysis of an epigenetically regulated five-lncRNA signature associated with outcome and chemotherapy response in ovarian cancer. *Front Cell Dev Biol.* (2021) 9:644940. doi: 10.3389/fcell.2021.644940
- Zheng M, Hu Y, Gou R, Nie X, Li X, Liu J, et al. Identification three lncRNA prognostic signature of ovarian cancer based on genome-wide copy number variation. *Biomed Pharmacother.* (2020) 124:109810. doi: 10.1016/j.biopha.2019.109810
- Zhang M, Wang G, Zhu Y, Wu D. Characterization of BRCA1/2-Directed ceRNA network identifies a novel three-lncRNA signature to predict prognosis and chemo-response in ovarian cancer patients with wild-type BRCA1/2. *Front Cell Dev Biol.* (2020) 8:680. doi: 10.3389/fcell.2020.00680
- Levy JMM, Towers CG, Thorburn A. Targeting autophagy in cancer. *Nat Rev Cancer.* (2017) 17:528–42. doi: 10.1038/nrc.2017.53
- Zhou J, Jiang YY, Chen H, Wu YC, Zhang L. Tanshinone I attenuates the malignant biological properties of ovarian cancer by inducing apoptosis and autophagy via the inactivation of PI3K/AKT/mTOR pathway. *Cell Prolif.* (2020) 53:e12739. doi: 10.1111/cpr.12739
- Fang YJ, Jiang P, Zhai H, Dong JS. lncRNA GAS8-AS1 inhibits ovarian cancer progression through activating beclin1-mediated autophagy. *Oncotargets Ther.* (2020) 13:10431–40. doi: 10.2147/OTT.S266389
- Gu L, Li Q, Liu H, Lu X, Zhu M. Long noncoding RNA TUG1 Promotes autophagy-associated paclitaxel resistance by sponging miR-29b-3p in ovarian cancer cells. *Oncotargets Ther.* (2020) 13:2007–19. doi: 10.2147/OTT.S240434
- Yu Y, Zhang X, Tian H, Zhang Z, Tian Y. Knockdown of long non-coding RNA HOTAIR increases cisplatin sensitivity in ovarian cancer by inhibiting cisplatin-induced autophagy. *J Buon.* (2018) 23:1396–401.
- Chen S, Wu DD, Sang XB, Wang LL, Zong ZH, Sun KX, et al. The lncRNA HULC functions as an oncogene by targeting ATG7 and ITGB1 in epithelial ovarian carcinoma. *Cell Death Dis.* (2017) 8:e3118. doi: 10.1038/cddis.2017.486
- Braga EA, Fridman MV, Moscovtsev AA, Filippova EA, Dmitriev AA, Kushlinskii NE. lncRNAs in ovarian cancer progression, metastasis, and main pathways: ceRNA and alternative mechanisms. *Int J Mol Sci.* (2020) 21:8855. doi: 10.3390/ijms21228855
- Zhu H, Gan X, Jiang X, Diao S, Wu H, Hu J. ALKBH5 inhibited autophagy of epithelial ovarian cancer through miR-7 and BCL-2. *J Exp Clin Cancer Res.* (2019) 38:163. doi: 10.1186/s13046-019-1159-2
- Human Genomics. The genotype-tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* (2015) 348:648–60. doi: 10.1126/science.1262110
- Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* (2016) 44:e71. doi: 10.1093/nar/gkv1507
- Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO gene nomenclature database, 2006 updates. *Nucleic Acids Res.* (2006) 34 (Database issue):D319–21. doi: 10.1093/nar/gkj147
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
- Engelbrechtsen S, Böhlin J. Statistical predictions with glmnet. *Clin Epigenetics.* (2019) 11:123. doi: 10.1186/s13148-019-0730-1
- Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* (2015) 12:453–7. doi: 10.1038/nmeth.3337
- Yoshihara K, Shahmoradgol M, Martínez E, Vegesna R, Kim H, Torres-García W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun.* (2013) 4:2612. doi: 10.1038/ncomms3612
- Geeleher P, Cox N, Huang RS. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS ONE.* (2014) 9:e107468. doi: 10.1371/journal.pone.0107468
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* (2012) 22:568–76. doi: 10.1101/gr.129684.111
- Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* (2018) 28:1747–56. doi: 10.1101/gr.239244.118
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* (2005) 102:15545–50. doi: 10.1073/pnas.0506580102
- Meng C, Zhou JQ, Liao YS. Autophagy-related long non-coding RNA signature for ovarian cancer. *J Int Med Res.* (2020) 48:300060520970761. doi: 10.1177/0300060520970761

27. Li S, Jia H, Zhang Z, Wu D. lncRNA GAS6-AS1 facilitates the progression of breast cancer by targeting the miR-324-3p/SETD1A axis to activate the PI3K/AKT pathway. *Eur J Cell Biol.* (2020) 99:151124. doi: 10.1016/j.ejcb.2020.151124
28. Ai J, Sun J, Zhou G, Zhu T, Jing L. Long non-coding RNA GAS6-AS1 acts as a ceRNA for microRNA-585, thereby increasing EIF5A2 expression and facilitating hepatocellular carcinoma oncogenicity. *Cell Cycle.* (2020) 19:742–57. doi: 10.1080/15384101.2020.1729323
29. Shi J, Zhang Y, Qin B, Wang Y, Zhu X. Long non-coding RNA LINC00174 promotes glycolysis and tumor progression by regulating miR-152-3p/SLC2A1 axis in glioma. *J Exp Clin Cancer Res.* (2019) 38:395. doi: 10.1186/s13046-019-1390-x
30. Li S, Zhang Y, Dong J, Li R, Yu B, Zhao W, et al. LINC00893 inhibits papillary thyroid cancer by suppressing AKT pathway via stabilizing PTEN. *Cancer Biomark.* (2020) 30:277–286. doi: 10.3233/CBM-190543
31. Sun Z, Jing C, Xiao C, Li T. An autophagy-related long non-coding RNA prognostic signature accurately predicts survival outcomes in bladder urothelial carcinoma patients. *Aging.* (2020) 12:15624–37. doi: 10.18632/aging.103718
32. Kandalaft LE, Powell DJ, Jr., Singh N, Coukos G. Immunotherapy for ovarian cancer: what's next? *J Clin Oncol.* (2011) 29:925–33. doi: 10.1200/JCO.2009.27.2369
33. Kandalaft LE, Odunsi K, Coukos G. Immunotherapy in Ovarian cancer: are we there yet? *J Clin Oncol.* (2019) 37:2460–71. doi: 10.1200/JCO.19.00508
34. Shimizu K, Iyoda T, Okada M, Yamasaki S, Fujii SI. Immune suppression and reversal of the suppressive tumor microenvironment. *Int Immunol.* (2018) 30:445–54. doi: 10.1093/intimm/dxy042
35. Darvin P, Toor SM, Sasidharan Nair V, Elkord E. Immune checkpoint inhibitors: recent progress and potential biomarkers. *Exp Mol Med.* (2018) 50:1–11. doi: 10.1038/s12276-018-0191-1
36. Deng Z, Li X, Shi Y, Lu Y, Yao W, Wang J. A novel autophagy-related lncRNAs signature for prognostic prediction and clinical value in patients with pancreatic cancer. *Front Cell Dev Biol.* (2020) 8:606817. doi: 10.3389/fcell.2020.606817
37. Majidpoor J, Mortezaee K. The efficacy of PD-1/PD-L1 blockade in cold cancers and future perspectives. *Clin Immunol.* (2021) 226:108707. doi: 10.1016/j.clim.2021.108707
38. Perez-Fidalgo JA, Grau F, Fariñas L, Oaknin A. Systemic treatment of newly diagnosed advanced epithelial ovarian cancer: from chemotherapy to precision medicine. *Crit Rev Oncol Hematol.* (2021) 158:103209. doi: 10.1016/j.critrevonc.2020.103209
39. Christie EL, Bowtell DDL. Acquired chemotherapy resistance in ovarian cancer. *Ann Oncol.* (2017) 28:viii13–5. doi: 10.1093/annonc/mdx446

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

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

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



Alpha-Synuclein Aggregation in Parkinson's Disease

E. Srinivasan^{1,2}, G. Chandrasekhar¹, P. Chandrasekar¹, K. Anbarasu², A. S. Vickram³, Rohini Karunakaran⁴, R. Rajasekaran¹ and P. S. Srikumar^{5*}

¹ Bioinformatics Lab, Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology (Deemed to be University), Vellore, India, ² Department of Bioinformatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India, ³ Department of Biotechnology, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Chennai, India, ⁴ Unit of Biochemistry, Faculty of Medicine, AIMST University, Bedong, Malaysia, ⁵ Unit of Psychiatry, Faculty of Medicine, AIMST University, Bedong, Malaysia

OPEN ACCESS

Edited by:

Thirumal Kumar D,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Manigandan Venkatesan,
The University of Texas Health Science
Center at San Antonio, United States
Jannathul Firdous,
University of Kuala Lumpur, Malaysia

*Correspondence:

P. S. Srikumar
gooddoc15@gmail.com

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 06 July 2021

Accepted: 01 September 2021

Published: 18 October 2021

Citation:

Srinivasan E, Chandrasekhar G,
Chandrasekar P, Anbarasu K,
Vickram AS, Karunakaran R,
Rajasekaran R and Srikumar PS
(2021) Alpha-Synuclein Aggregation in
Parkinson's Disease.
Front. Med. 8:736978.
doi: 10.3389/fmed.2021.736978

Parkinson's disease (PD), a neurodegenerative disorder characterized by distinct aging-independent loss of dopaminergic neurons in substantia nigra pars compacta (SNpc) region urging toward neuronal loss. Over the decade, various key findings from clinical perspective to molecular pathogenesis have aided in understanding the genetics with assorted genes related with PD. Subsequently, several pathways have been incriminated in the pathogenesis of PD, involving mitochondrial dysfunction, protein aggregation, and misfolding. On the other hand, the sporadic form of PD cases is found with no genetic linkage, which still remain an unanswered question? The exertion in ascertaining vulnerability factors in PD considering the genetic factors are to be further dissevered in the forthcoming decades with advancement in research studies. One of the major proponents behind the prognosis of PD is the pathogenic transmutation of aberrant alpha-synuclein protein into amyloid fibrillar structures, which actuates neurodegeneration. Alpha-synuclein, transcribed by SNCA gene is a neuroprotein found predominantly in brain. It is implicated in the modulation of synaptic vesicle transport and eventual release of neurotransmitters. Due to genetic mutations and other elusive factors, the alpha-synuclein misfolds into its amyloid form. Therefore, this review aims in briefing the molecular understanding of the alpha-synuclein associated with PD.

Keywords: Parkinson's, genes, alpha-synuclein, mutation, protein aggregation

INTRODUCTION

Parkinson's disease (PD) is the second most common neurodegenerative disorder that falls under the category of synucleinopathy (1). PD is characterized by distinct aging-independent loss of dopaminergic neurons in substantia nigra pars compacta (SNpc) region and the decrease in dopamine levels (2). Possibly, PD leads to the loss of terminal ends of striatum, which occurs before the neuronal loss in SNpc and; it seems to be more significant in disease pathogenesis (3). About 95% of PD cases are sporadic with no genetic linkage. Mostly, PD has its mean age of onset at 55 years with increased incidences with aging (1). PD is the most common disorder in a range of disorders classified as Parkinsonism characterized by dopamine deficiency and striatal damage.

In early stages of PD, the patients are pre-symptomatic with possible pathological changes of dorsal motor nucleus of vagus in medulla and the anterior olfactory nucleus of olfactory bulb followed by changes in locus ceruleus neurons of pons and dopaminergic neurons of substantia

nigra. Hence, the smell and taste disturbances may be early clinical features of PD. During later stages, the pathology spreads to amygdala, basal forebrain, and medial temporal lobe structures. Neocortex is affected in final stages of the disease. Wherein, these stages are based on Braak's PD staging scheme (2, 4).

PD is characterized by motor symptoms such as bradykinesia, hypokinesia, akinesia, hypomimia, hypophonia, drooling, swallowing problems, micrographia, decreased stride length during walking, rigidity (stiffness of limbs), postural instability, and resting tremors (5). Some of the non-motor symptoms of PD include sleep disorders, depression, memory impairment, lack of initiative, delayed response, slowed cognition, passiveness, psychosis, and confusion (6–9). Pain is the most common non-motor symptom in PD patients and it may occur even before the motor symptoms. Hence, these symptoms may not share the same pathogenic pathways. Nociceptive dysfunction in the peripheral primary afferent nerves leading to abnormal sensory input and degeneration has been hypothesized as a possible reason for the early-stage pain symptoms and impaired response to pain stimuli (10).

Other parkinsonian disorders and related atypical parkinsonian disorders are caused by multiple system atrophy (MSA), tauopathies, or progressive supranuclear palsy (PSP) and cortico-basal degeneration (CBD) (7).

The most pathological hallmark of PD is Lewy bodies (LB). Lewy bodies are intraneuronal inclusions that contain immunoreactive alpha-synuclein aggregates which may also contain various neurofilament proteins as well as proteins involved in proteolysis such as ubiquitin. Predominantly, the cell death is caused by disruption of nuclear membrane integrity and release of alpha synuclein aggregation promoting nuclear factors like histones. Alpha synuclein may spread to other cells by direct or indirect means once aggregation starts. When compared with unaffected normal individuals, around 50–70% of neurons are lost in this region, at the time of death in patients with PD (11–13). Some studies suggest that LBs are the cell's defensive mechanism to prevent intracellular protein aggregate accumulation, while other studies suggest LBs to have a pathogenic role in PD. Therefore, LBs are an area of controversy in PD. LB formation may activate pathways for neuronal dysfunction and cell death (2, 4).

Mutations in the alpha-synuclein gene are responsible for some familial cases of PD with LB, whereas mutations in the Parkin gene cause a parkinsonian syndrome without LB in early-onset cases. Furthermore, Parkin proteins cause ubiquitination of the alpha-synuclein by interacting with synphilin-1 and thus promote the formation of LB (14–16). Mutations in genes coding the proteins involved in ubiquitin-proteasome system (UPS) and some deubiquitinating enzymes have been linked to PD. UPS removes unwanted proteins inside the cell and maintains many intracellular processes for cell viability (5).

This review offers an analytical assessment of the literature designating the possible role of the genes involved in causing PD. The information was collected from the molecular, cellular, and computational studies from various library databases and search engines. Hence, this article helps in providing a better

understanding over the impact of the alpha synuclein and its mutations causing PD in terms of the molecular and neurological perspectives.

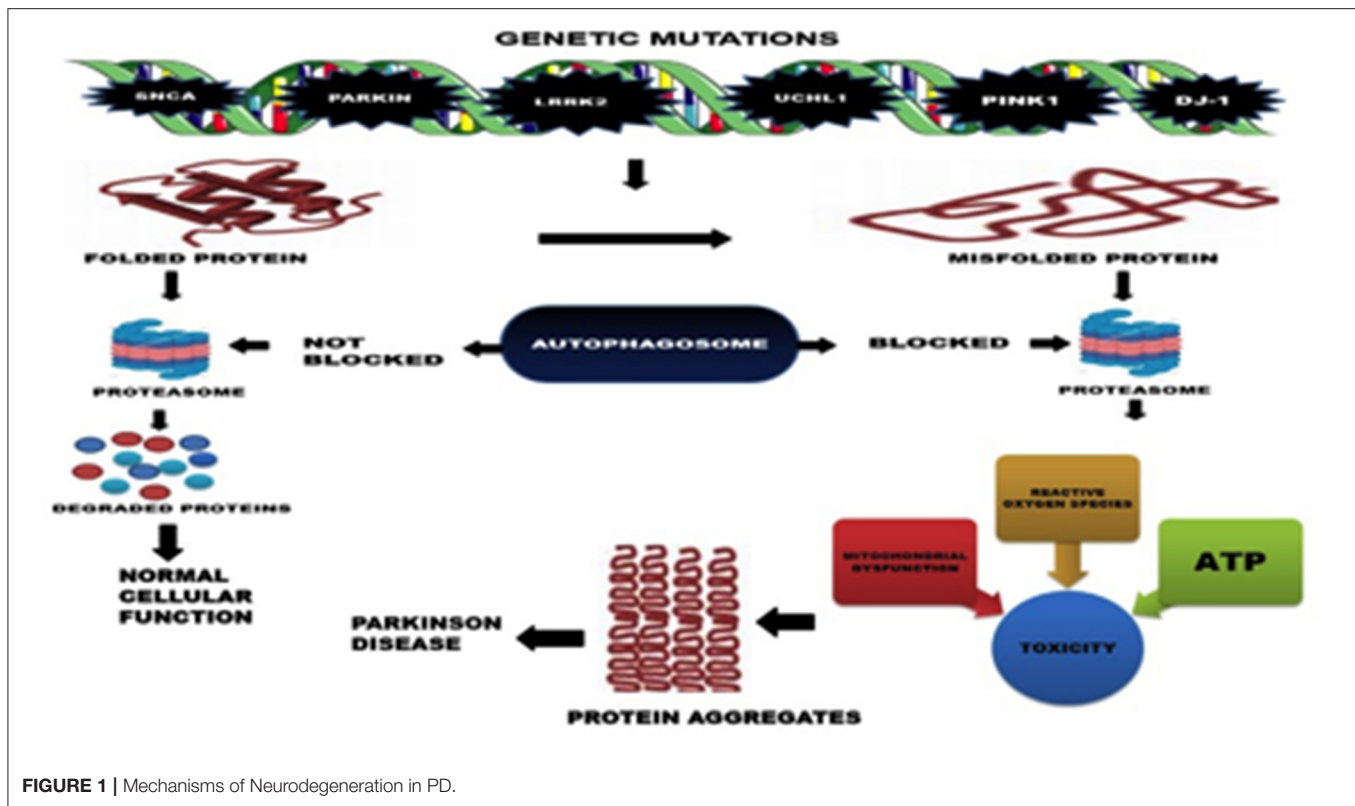
OVERVIEW ABOUT THE AMYLOIDOSIS IN NEURODEGENERATIVE DISORDERS

The pathological commonality among predominant neurodegenerative disorders such as Alzheimer's disease (AD), Huntington's disease (HD), PD, prion disorders, and amyotrophic lateral sclerosis (ALS) are protein aggregation and the formation of inclusion bodies. The aggregates also called as amyloids, are composed of fibers and fibrils consisting aberrant misfolded proteins rich in beta-sheets. Amyloid deposition is directly associated with cellular deterioration and neuronal malfunction. It is also insinuated to effectuate endoplasmic reticulum, oxidative stress, mitochondrial, and proteasomal dysfunction, thus eventually causing neuronal waning. While mutations are culpable behind protein misfolding, sporadic causes attributed to other genetic and environmental variables. Increasing reports from various studies have elucidated the better comprehension of biochemical pathways associated with protein aggregation. Chief pathways suspected to cause protein aggregation include unfolded states/unfolding intermediates mediated physical aggregation, aggregation propelled chemical linkages, or protein self-association and aggregation mediated via chemical degradations (17–20).

DISORDERED PROTEINS INVOLVED IN PATHOGENESIS OF PARKINSON'S

Primarily, the accumulation of misfolded proteins has been suggested to be the cause of PD. Mutations in alpha-synuclein (SNCA), ATP13A2, GBA, FBX07, VPS35, PLA2G6, DNAJC6, SYNJ1, UCHL1, parkin (PRKN), LRRK2, PINK1, and DJ-1 genes has been identified to cause familial early onset PD by causing abnormal protein conformations and disrupting the ability of cellular machinery to clear the misfolded proteins (**Figure 1**) (6, 21). Pdr-1 (Parkin gene) and PINK1 mutants in *C. elegans* have been found to exhibit defective dopamine dependent behavior.

ATP13A2 encodes cation-transporting ATPase 13A2 that partakes in maintaining mitochondrial, lysosomal, and neuronal integrity. Its chief function involves transporting divalent transition metal cations. On the other hand, Glucosylceramidase beta encoded by GBA gene, is implicated in glucocerebroside's hydrolyzation. While, FBX07 gene translates into F-box only protein 7 that is involved in mitophagy. VPS35 encodes Vacuolar protein sorting-associated protein 35, which participates in autophagy by transporting proteins across Golgi apparatus and vesicular structures. Next, gene PLA2G6 translates into an 85–88 kDa protein, which is calcium-independent phospholipase A2. It is also involved in regulating phospholipid remodeling and fatty acids' release from phospholipids. In addition, it has been implicated in prostaglandin and leukotriene production, and in nitric oxide/vasopressin mediated arachidonic acid release. DNAJC6, on the other hand, encodes putative tyrosine-protein



phosphatase auxilin that partakes in neuronal clathrin-mediated endocytosis. Further, *SYNJ1* translates into Synptojanin-1, which is a phosphoinositide phosphatase enzyme that modulates PIP2 in membrane. Next, Ubiquitin carboxyl-terminal hydrolase isozyme L1 is encoded by gene *UCHL1*. It is a deubiquitinating enzyme that produces monomers of ubiquitin, suspected to regulate monoubiquitin's degradation in lysosomes. And, *PRKN* translates to E3 ubiquitin-protein ligase parkin, a ubiquitin ligase that participates in effacing damaged or misfolded proteins (22).

The *pdr-1* and *PINK1* mutants had greater accumulation of dysfunctional mitochondria with age, leading to the activation of mitochondrial ubiquitin proteasome response. By this way, the prevention of upregulation of ubiquitin protein response was found to reduce dopaminergic neuron lifespan and lead to their loss in *C.elegans* (23). *PINK1* and Parkin were found to regulate the mitochondrial quality control in neurons by removing damaged mitochondria (with reduced membrane potential, increased ROS production, defective electron transport chain, or accumulation of unfolded proteins) through autophagy and thus replacing the damaged mitochondria. Mutations in either *PINK1* or Parkin can alter this mechanism and lead to mitochondrial dysfunction, which is found to be prevalent in PD neurons (24). *LRRK2* is the most common gene involved in sporadic PD, while the other gene defects cause only a small number of familial PD cases; however, they provide information on the proteins involved and disease mechanism (25). Point mutations in alpha-synuclein gene has been identified to cause early-onset PD in an autosomal-dominant way and

the overexpression of gene has been found to cause late-onset or sporadic cases of PD (26). *LRRK2* gene codes for the dardarin protein and this gene is the most common cause of familial or sporadic PD. LBs have been found in PD cases involving *LRRK2* (27). *LRRK2* G2019S mutation (substitution of glycine to serine at codon 2019) accounts for the majority of familial cases and 1.6% of sporadic cases of PD even though its prevalence is variable (28). Single gene mutations in Parkin and DJ-1 cause early-onset PD and these mutations are inherited by autosomal-recessive pattern. Abnormalities in mitochondrial Complex I of oxidative phosphorylation pathway have been consistently found to cause mitochondrial defects by inducing oxidative stress, increasing the production of reactive oxygen species along with superoxides that can target the electron transport chain to accelerate their own production and energy failure leading to PD pathogenesis (29). Dopaminergic neurons are particularly vulnerable to oxidative stress reactive oxygen species mediated mitochondrial dysfunction because the metabolism of dopamine produces superoxides radicals, hydrogen peroxide, and DA-quinone (produced by auto-oxidation of dopamine) which damage the cells. However, the direct links between ROS generation, defects in oxidative phosphorylation, and PD pathology are not strong and convincing, because of the rare occurrence of Parkinsonism in patients with mutations affecting oxidative phosphorylation. Mitochondrial defects have been suggested to cause cell death and dysfunction in PD, which could occur due to the inherited defective mitochondrial DNA or mutations in mitochondrial

genome caused by possible systemic toxicity. Deregulation of kinase signaling, disruption of signaling mechanisms in dopaminergic neurons and endoplasmic reticulum stress are also key molecular mechanisms in PD (30, 31). PINK1 gene codes for a mitochondrial complex that has been shown to be responsible for autosomal-recessive form of PD, but it is not a major risk factor of sporadic PD. Programmed cell death has been suggested to cause the neuron cell death in PD, but whether it causes cell death due to abnormal pathway or to clear cells injured and damaged by the pathological mechanisms of PD is still not clear (6, 8, 32).

Furthermore, Franco et al. have compiled the aforementioned genes' functionalities in terms of their biochemical, biomolecular, network interactions, and pathway analysis from various *in silico*, *in vitro*, and *in vivo* studies, to formulate putative mechanism: due to genetic mutations observed in the afore mentioned list of genes, involved in protein processing and vesicular trafficking, native alpha synuclein's normal processing is hindered and altered (22).

STRUCTURAL ARCHITECTURE AND DISEASE-CAUSING MUTATIONS IN ALPHA SYNUCLEIN (AS)

Alpha synuclein is a small (14 kDa), intrinsically disordered protein with 140 amino acids that is highly charged and coded by SNCA (synuclein) gene, which is mainly expressed in CNS (33). Alpha synuclein is predominantly found at the presynaptic terminals, where it associates with synaptic vesicles. Alpha synuclein belongs to the synuclein family, which also includes gamma and beta synucleins. One percentage of the total proteins of neuronal cytosol are comprised of AS. AS has an amphipathic N-terminus consisting of 7 imperfect sequence repeats of 11 residues with a possible alpha helix structure that facilitates lipid binding and a potential role in aggregation. Further, the non-amyloid component (NAC) at the C- terminus facilitates calcium binding and inhibits protein aggregation (34).

Moreover, NMR structure of human alpha-synuclein reported by Ulmer et al. indicates that amino acids in-between 3–37 and 45–97 forms alpha-helices (curved), joined by linker (extended) organized in an unprecedented anti-parallel manner. Then, a notably high mobile tail spans in-between 98 and 140 amino acid range. The well-organized orientation of helical connector implicates a demarcated association with lipidic surfaces, indicating a notion that, when adhered to synaptic vesicles with larger diameter, it can function as a modulator between the aforementioned structure and an uninterrupted helical model that was postulated earlier (35).

AS is involved in synucleinopathies like PD (both familial and sporadic), multiple systems atrophy and dementia with LB; however, the physiological function of AS is still unknown. Researchers suggest that AS could potentially play a role in cell function regulation, dopamine release regulation, vesicular trafficking, and oxidative stress. Removal of AS gene in mice has been found to cause the loss of dopaminergic neurons, striatal dopamine reduction, and absence of dopamine-induce

locomotive responses mediated by dopamine transporter (DAT) (4). Missense point mutations of the N-terminal (A53E, A53T, A30P, E46K, H50Q, and G51D) have been strongly correlated with autosomal dominant form of PD, while, the duplication and triplication of AS gene have been shown to be involved in familial PD cases with early onset (36). Mutant AS proteins vary in only a few amino acid residues but result in a significant change in their conformation and the type of aggregates formed (37). There is no explanation for this phenomenon to date. AS inclusions are found to be usually hyperphosphorylated at various sites including Ser129, Ser87, and Tyr125 in LB (38, 39). Mutant AS protein (A30P and A53T disease mutations) involved in familial PD have been shown to be structurally defective for membrane binding, leading to alteration of the protein's binding properties (40). Wild-type AS has been observed to form two different dimers and; the single point mutations (A30P, E46K, and A53T) have been suggested to promote dimerization of AS. Moreover, the structural homogeneity of these dimers has been suggested to lead in different aggregation pathways (41). In a mutation-frequency analysis conducted on Japanese patients, the SNCA p.A53V homozygous mutation was found to cause distinct phenotype of progressive Parkinsonianism and cognitive decline similar to SNCA missense mutation. Particularly, the two newly discovered mutants of AS viz., A18T and A29S were found to aggregate faster than wild-type AS with greater propensity for aggregation. Furthermore, the A18T mutant was found to have faster aggregation kinetics compared to A29S and hence, it makes the protein more sensitive to aggregation by modifying its native conformation (42). Of note, the broken helix structure of AS, which consists of two antiparallel membrane bound helices connected by a non-helical linker causes the protein to interact with synaptic vesicles docked at the plasma membrane. Phosphorylation of tyrosine at position 39 (Y39 phosphorylation) in AS, *in vitro*, was found to free the protein from the membrane surface of vesicles by decreasing the binding of helix-2 in the broken helix state. In addition, this effect was found to be similar to the effect of G51D mutation (43). The peptide (1a) consisting of residues 36–55 of AS was found to form a beta hairpin structure that subsequently assembled into a triangular trimer. Full length AS has been suggested to form such an assembly with evidences from molecular modeling. Also, this 1a peptide was able to bind anionic lipid bilayers membranes and nucleate the oligomerization of AS (44). Recently in a study of 426 Italian PD patient, the 263 bp allelic variant of Rep-1 (D4S3481 microsatellite), present upstream of the SNCA gene translation start site, was found to raise the risk of hallucinations and dementia in patients carrying this variant compared to the non-carriers (45). Chinese PD patients have been found to have lower resting-state brain activity in the lingual gyrus and left caudate, when the amplitude of low-frequency fluctuation (ALFF) values of the brain was compared between the PD patients and healthy controls. Furthermore, the participants carrying the rs894278 single nucleotide polymorphism in SNCA gene, also called as the G allele, were found to have lower ALFF values in the right fusiform compared to non-carriers of the G allele. This study suggests that PD may also alter the brain connections (46).

Apart from aforementioned mutations, certain genetic polymorphism in SCNA can be accredited toward enfeebling sporadic PD. Reports have shown that SNPs: rs7684318, rs894278, and rs2572324 have been known to enhance susceptibility toward sporadic PD. Further, missense mutations A29S and A18T were found in patients with sporadic PD (47). Also, a study conducted in Japanese population states that SNPs rs2736990 and rs356220 were considerably associated with sporadic PD risk (48).

Due to various genetic mutations and other obscure biomolecular circumstances, AS pathogenically misfolds and transmutes into amyloid fibrils that are rich in beta-sheets. Solid state NMR structure of pathogenic AS fibrils was presented by Tuttle et al. it was reported that AS fibril had more than 200 distinctive long-range distance restraints that delineates a consensus structure possessing characteristics such as hydrophobic-core residues and in-register β -sheets, and diverse residual framework such as a glutamine ladder, intermolecular salt bridge, small residues mediated close backbone interactions, and various other steric zippers that stabilizes the orthogonal Greek-key topology (49). However, it should be noted that amyloid fibrils are usually polymorphic: capable of existing in multiple viable forms (50). Recently, Guerrero-Ferreira et al. had reported two new polymorphic structures of AS fibrils, which evince distinct morphological features (51).

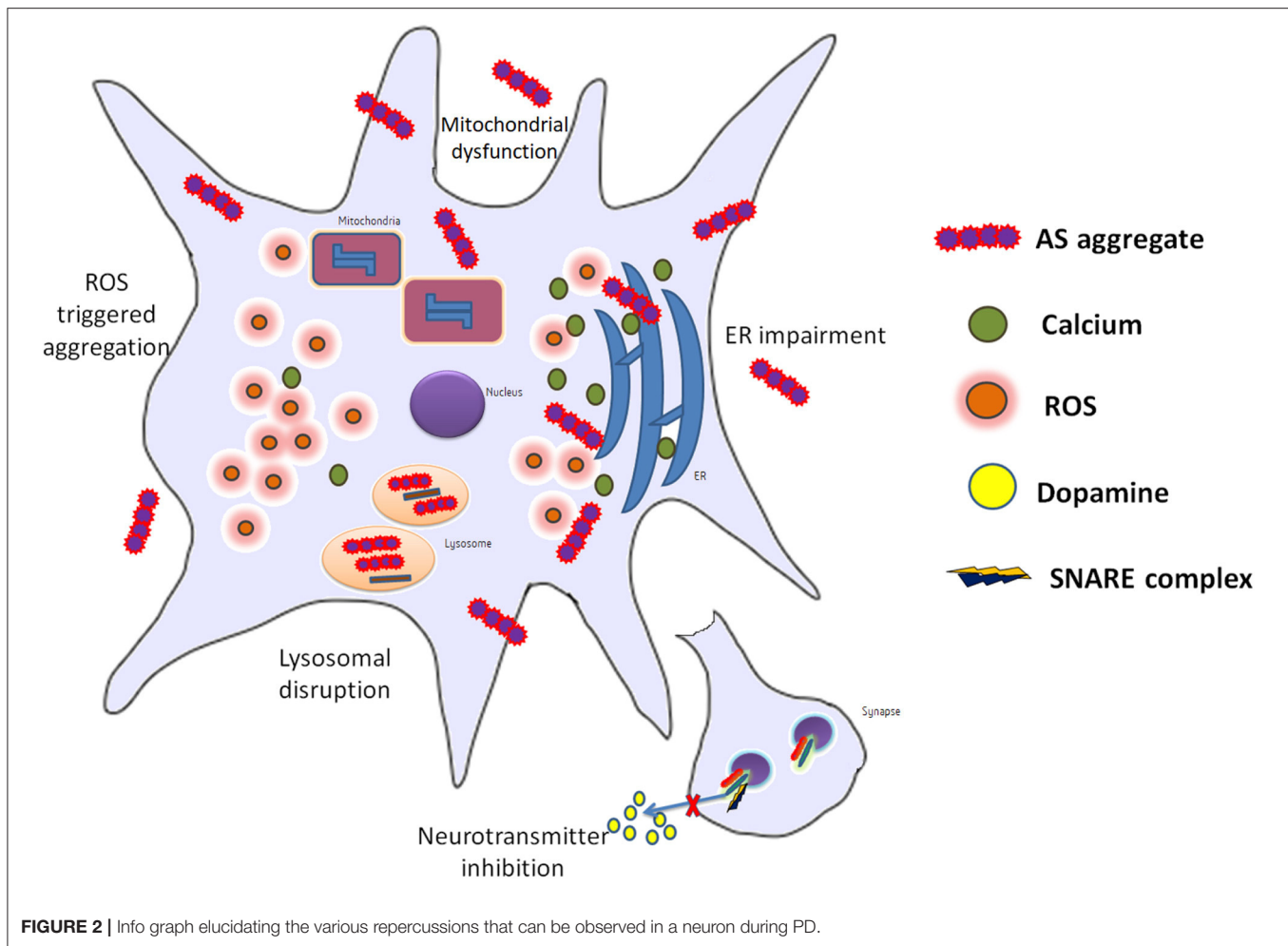
NEUROTOXIC EFFECT OF AS IN PD

AS is associated with calcium homeostasis and its overexpression may disrupt calcium homeostasis making dopaminergic neurons vulnerable to damage (52). Besides, AS has been suggested to have good affinity to phospholipid membranes, especially the synaptic vesicles with a preference for membranes with high curvature and specific membrane microdomains (53). AS has been found to regulate synaptic vesicle distribution and presynaptic terminal size. In specific, the N-terminal alpha helix helps in its lipid-binding capability (54). Increase in AS levels has been suggested to decrease dopamine as well as glutamine neurotransmission by interacting with the SNARE protein complex, modulating endoplasmic reticulum-golgi vesicular trafficking, inhibiting vesicular priming, and reducing synaptic contact (38). Overexpressed monomeric wild-type AS has been shown to inhibit vesicle endocytosis and impair neurotransmission. AS has been suggested to cause PD by disrupting the synthesis, storage, recycling, reuptake, and efflux of dopamine (55). Increased levels of AS has also been found to reduce active tyrosine hydroxylase (TH) enzyme that is involved in the production of dopamine by stabilizing TH in its inactive form. AS overexpression was found to attenuate vesicular monoamine transporter 2 (VMAT2) activity. Thus, the dopamine is stored in synaptic vesicles via VMAT2 after its production to reduce the oxidative damage from its metabolites. Increased cytosol concentration of dopamine due to the reduction of VMAT2 activity by AS has been proposed as a possible neurotoxic pathway in PD (56). Dopamine transporter (DAT) has been associated with

dopamine trafficking, but whether it increases or decreases DAT levels is still a debate because of evidence supporting both sides (57). DAT knock-out mice showed high extracellular dopamine levels and low intraneuronal dopamine concentration; thus, DAT is important for neurotransmission and its activity (especially decreased dopamine uptake as well as increased dopamine clearance and efflux) upon disruption by AS can cause PD. Most of these proposed functions of AS rely on its membrane binding capacity (38, 39, 58–61). *In vivo*, AS is distributed as unstructured, cytosolic, soluble, and partially membrane-bound state. Of note, the equilibrium between the structured and unstructured states or the balance of the order and disorder of AS conformations at the surface of membranes has been suggested to influence the biological functions of the protein and also lead to the aggregation of the protein due to detachment from the membrane (62). Membrane binding by AS requires a conformational transition into a highly helical state and this is promoted by the amphipathic N-terminal. Moreover, the C-terminal of AS was found to be highly disordered and the NAC domain has been found to be highly structured in beta-sheet conformation in the amyloid state of AS (59). AS oligomers were found to stabilize and enhance pre-existing defects in supported lipid bilayers (SLBs) of 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine/1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-L-serine (POPC/POPS). Particularly, the exposed lipid acyl chains at the edges of membrane defects were suggested to promote the membrane-oligomer interactions resulting in the development of fractal domains lacking lipids. The growth of membrane damage pattern did not depend on the lipid-oligomer interaction suggesting an oligomer-dependent, diffusion limited extraction mechanism for the enhancement of membrane damage by AS oligomers (63).

AS may not form pores in membranes or induce damage by itself but enlarges previous membrane defects instead. The synphilin-1 proteins have been found to be accumulated abnormally in AS inclusions of synucleopathies (64). AS and neuroinflammation mediated by the inflammatory response through microglial activation have been suggested to potentiate each other. Misfolded AS may activate microglia by increased expression of TNF- α , IL-1 β , IL-6, iNOS, and COX-2. Following stimulation by AS, receptors (TLRs, CD36, and FC γ R) as well as signaling molecules (galectin-3, MMPs, and PHOX) have been proposed to join the microglial response to activate signaling pathways (NF- κ B and MAPKs). These pathways further contribute to PD progression (65) (Figure 2).

Aberrant AS is also suspected to affect protein degradation systems. It was reported that malformed AS, under certain circumstances can trigger proteasomal dysfunction *in vitro* and *in vivo*. Further, mutant AS could also affect the lysosomes by impairing CMA, which in turn could induce macroautophagy that has a notable influence on neuronal survival. Also, AS is reckoned to influence dynamics of cytoskeletal system. Studies have shown that AS could interact and impact actin polymerization to affect cellular trafficking; substantial deviations on actin were reported, between native and A30P AS mutant. These findings indicate that apart from aberrant AS's neurotoxic



role, it can also venture into aforementioned pathological digressions to propel PD disorder (34).

MECHANISM OF AS AGGREGATION

AS aggregation was found to be triggered by Cu(II) in conditions relevant to PD. Two independent Cu(II) binding sites with varying affinities in the N-terminus of AS were identified and the complex formation between the high affinity site of AS as well as Cu(II) was found to be critical to the metal-mediated fibrillation process. However, this complex formation is also hypothesized to make the protein vulnerable to oxidative damage. Presence of a truncated C-terminal was found to accelerate the process due to reduction of the non-aggregation activity of the C-terminal (66). The formation of N-terminal acetylated AS- Cu(I) complex at the N-terminal site I has been found to stabilize local conformations with alpha-helical secondary structure and restricted motility. Besides, the formation of this complex with stabilized helically folded structure may occur *in vivo* and also, affect the membrane binding and aggregation capability of the N-terminal acetylated

AS (67). AS proteins lacking residues 109–140 in the C-terminal were found to form amyloid fibrils with strongly twisted beta-sheets, an increased beta-sheet distance, higher solvent exposure compared to monomeric wild-type AS, and incompatibility with the wild-type monomeric AS (68). Acetylation of the N-terminus of naturally occurring AS was found to perturb its ability to bind Cu²⁺ and the presence of the H50Q missense mutation along with the N-terminal acetylation prevented copper binding (69). Over expression of AS in frozen-hydrated primary midbrain neurons was found to increase intracellular Mn levels with reduction in levels of Ca, Zn, K, P, and S. Cu/Mn and Fe/Mn levels were found to have a strong correlation with AS overexpression. Mn release from the cells was also found to be reduced (70). Clioquinol was found to reduce the interaction between iron and A53T mutant AS in transgenic mice with improved phenotype (71). These studies suggest the involvement of transition metals in AS aggregation mediated PD pathology and suggest the regulation of these metal levels as a potential therapeutic strategy.

Tubulin polymerization promoting protein (TPPP/p25) has been found to co-localize with AS and induce its

aggregation (especially oligomers and protofilaments) by forming a complex even though they are produced by different cells—oligodendrocytes and neurons, respectively. TPPP/p25 has been shown to be present in LB. Both of these are Neomorphic Moonlighting Proteins. TPPP/p25 is also a chameleon protein with high conformational plasticity and it is involved in physiological functions such as regulation of microtubules. Hence, targeting it with therapeutic drugs is difficult. A study on the chameleon TPPP/p25 protein- AS complex using NMR, CD spectroscopy, ELISA, and PCR followed by computational study using charge-hydropathy plot (CH-plot) and cumulative-distribution function plot (CDF-plot) has proposed that the interface between the two proteins in the pathological complex is a potential target for therapeutic drugs (72, 73). TPPP/p25 α has also been shown to promote release of AS through exophagy and impairment of autophagosome formation/trafficking and lysosomal dysfunction preventing the degradation of AS (74). Cross-seeding between wild-type, A30P and A53T AS variants that vary in only one or two amino acid residues but result in different fibril morphologies were studied. Wherein, the study suggested that similarity of conformations of the seeds and monomers is essential for seed elongation (37). Further, the morphology of wild-type AS fibrils has been suggested to be determined by competitive growth between different polymorphs during fibrillation and slow maturation or annealing process of fibrils (75). AS was found to bind strongly with lipid membranes of large unilamellar vesicles with anionic or zwitterionic headgroups *in vitro* and this binding is suggested to reduce the protein-protein interactions of AS that lead to fibrillation. Recent updates in research pertaining to PD and AS is delineated in **Table 1**.

OVERVIEW OF THERAPEUTICS TARGETING AS

Presently, there are no effective therapeutics contrived for PD, but medication/surgery alleviates the symptoms and ameliorates motor impairments. A prudent way to effectively alleviate PD would be to target one of its crucial causatives, AS. Recently, various therapeutic strategies have been formulated, to encumber AS's toxic effect. One such strategy would be to control transmission by blocking AS receptors. LAG3-directed antibodies were reported to substantially regulate aberrant AS induced toxicity. Concurrently, silencing AS expression in mouse and rat brain models through shRNA and siRNA was also reported. Further, an oligomer regulator: Anle138b [3-(1,3-benzodioxol-5-yl)-5-(3-bromophenyl)-1H-pyrazole] was able to hinder the synthesis and accumulation of AS oligomers. Additionally, numerous small molecule-based inhibitors have been elucidated to impede AS aggregation. One such inhibitor: Methylthioninium was found to effectively control AS fibrillar inclusions *in vitro* and *in vivo*. Besides, numerous phytochemical and plant extracts were also found to effectively modulate AS aggregation in PD models. Lately, the use of phytochemicals to target AS aggregated has gained a lot of attention and focus in the research community (96, 97).

COMPUTATIONAL STUDIES ON AS

Using Monte Carlo (MC) computational methods (Replica Exchange MC and Canonical Protein MC, Molecular Dynamics (MD) based on CHARMM force field and AMBER force field were used in various studies to disintegrate the mutational effect on AS protein. On the other hand, virtual screening, docking using AUTODOCK (using Lamarckian Genetic Algorithm) and Interaction Potentials for assessment of protein druggability (MD- based DruGUI) were used to analyze the effect of drug molecules and other lead compounds on compact structures of AS in aggregates. Moreover, the regions of the protein that exhibit extended variable beta-sheet structure with a potential role in aggregation and the protein druggability for potential drugs inhibiting aggregation has been studied extensively (98). Notably, 332 genes that affect AS toxicity were identified using genome-wide screens of yeast. Later, these genes were used to map the human counterparts by using a newly developed computational method topology called Transpose Net that integrates Steiner prize-collecting approach with homology assignment through sequence, structure and interaction. Gene interaction profiles of ATP13A2/PARK9 and VPS35/PARK17 and network relationships between the genes LRRK2/PARK8, ATXN2, RAB7L1/PARK16, and EIF4G1/PARK18 were identified and confirmed in induced pluripotent stem cell (iPSC) derived neurons following the computational mapping (99).

Computational study using molecular dynamics simulations and modeling has been used to reveal varying aggregation effects in distinct protein conformational disorders (100–105). Thus, MD studies have revealed that the NAC domain of AS monomers strong interacts with the amyloid beta monomers. In specific, the cross-seeded NAC-amyloid beta oligomers were found to show polymorphism with the NAC oligomers preferring to form double layer conformations with amyloid beta over single layer conformations. Further, the self-assembled NAC oligomers with three beta strands connected by two turns were found to affect the secondary structure of self-assembled amyloid beta oligomers. Distinctly, the inner core distance values of NAC oligomers remained unchanged in cross-seeded NAC-amyloid beta oligomers but the inner core distance values in single layer amyloid beta conformations were decreased to form a compact and stable cross beta structure by strong hydrophobic interactions with the inner core domain. N-terminal of AS has been found to play an important role in the self-assembly of AS (1–140) into fibrils of cross-beta structure using models constructed by G-key 3D structure of self-assembled AS, NMR fibril structures, molecular dynamics simulations and GBMV calculations followed by analysis.

Recently, AS has been found to move into the nucleus by binding to Retinoic acid in a calreticulin dependent manner leading to increase toxicity and the AS in nucleus has been suggested to particularly increase the expression of PD-associated genes (ATP13A2, PINK1) in SH-SY5Y cells (106). N-terminal mutants of AS (D2A, D2P) were found to alter the N-terminal acetylation, protein level, stability, risk of neuron death and toxicity of AS in neuron as well as HEK293 models, showcasing the importance of the N-terminal amino acids and

TABLE 1 | Table delineating the recent research updates pertaining to AS and PD.

S.no	Year of publishing	Author(s)	Title	Key research findings	References
1	2020	Mellier et al.	The process of Lewy body formation, rather than simply α -synuclein fibrillization, is one of the major drivers of neurodegeneration	<ul style="list-style-type: none"> Proposed seeding based model for AS fibrillation. Elucidated the interaction between AS aggregates and cellular organelles like mitochondria ER etc. 	(76)
2	2020	Ray et al.	α -Synuclein aggregation nucleates through liquid-liquid phase separation	Provides detailed characterization of AS's phase-separation behavior and its pathogenic transformation into aggregates.	(77)
3	2020	Perren et al.	The structural differences between patient-derived α -synuclein strains dictate characteristics of Parkinson's disease, multiple system atrophy, and dementia with Lewy bodies	Characterizes structural variations of pathogenic AD in various synucleinopathies.	(78)
4	2020	Stephens et al.	Extent of N-terminus exposure of monomeric alpha-synuclein determines its aggregation propensity	Identifies structural orientations and environmental conditions that furthers monomeric AS's aggregation.	(79)
5	2020	Zhao et al.	Parkinson's disease-related phosphorylation at Tyr39 rearranges α -synuclein amyloid fibril structure revealed by cryo-EM	Identifies PD's pathological outcome that reorients AS's 3D structure.	(80)
6	2020	Arlehamn et al.	α -Synuclein-specific T cell reactivity is associated with preclinical and early Parkinson's disease	Examines the association between α -syn-specific T cell responses and PD.	(81)
7	2020	Niu et al.	A longitudinal study on α -synuclein in plasma neuronal exosomes as a biomarker for Parkinson's disease development and progression	Findings indicate that AS found in plasma neuronal exosomes could server as biomarker in early PD detection and also as prognostic marker for the PD's progression.	(82)
8	2019	Bhattacharjee et al.	Mass Spectrometric Analysis of Lewy Body-Enriched α -Synuclein in Parkinson's Disease	Identified 20 AS variant species in PD patients.	(83)
9	2019	Levine et al.	α -Synuclein O-GlcNAcylation alters aggregation and toxicity, revealing certain residues as potential inhibitors of Parkinson's disease	Advocates O-GlcNAcylation as a potential therapeutic tool for regulating PD.	(84)
10	2019	Tian et al.	Erythrocytic α -Synuclein as a potential biomarker for Parkinson's disease	Finding indicate that AS levels are altered in PD patients' erythrocytes and peripheral erythrocytic could be a potential biomarker.	(85)
11	2019	Elfarrash et al.	Organotypic slice culture model demonstrates inter-neuronal spreading of alpha-synuclein aggregates	Characterizes how AS aggregates are formed and spreads across neurons.	(86)
12	2019	Landeck et al.	Toxic effects of human and rodent variants of alpha-synuclein <i>in vivo</i>	Quantifies the neurodegenerative effects of rodent and human AS variants.	(87)
13	2019	Brys et al.	Randomized phase I clinical trial of anti- α -synuclein antibody BIIB054	Delineates the phase I clinical trial outcomes of BIIB054 antibody against AS, which indicates that BIIB054 pose tolerability, safety, and pharmacokinetic propensities in PD patients.	(88)
14	2019	Kang et al.	Comparative study of cerebrospinal fluid α -synuclein seeding aggregation assays for diagnosis of Parkinson's disease	Formulates high accuracy AS seeding aggregation assay that can be used for PD diagnosis.	(89)
15	2018	Pihlstrom et al.	A comprehensive analysis of SNCA-related genetic risk in sporadic Parkinson's disease	Delineates AS encoding SCNA gene polymorphisms that are associated with sporadic PD.	(90)
16	2018	Longhena et al.	Synapsin III is a key component of α -synuclein fibrils in Lewy bodies of PD brains	Identifies Synapsin III as an integral part of AS fibrils in PD patients' brain.	(91)
17	2018	Faustini et al.	Synapsin III deficiency hampers α -synuclein aggregation, striatal synaptic damage, and nigral cell loss in an AAV-based mouse model of Parkinson's disease	Constitutes SynapsinIII as a crucial factor for AS aggregation and could be a potential target in mitigating PD.	(92)
18	2018	Papagiannakis et al.	Alpha-synuclein dimerization in erythrocytes of patients with genetic and non-genetic forms of Parkinson's Disease	Increased AS dimers were observed in PD patients, which could provide insights into PD prognosis and can be used as biomarker.	(93)

(Continued)

TABLE 1 | Continued

S.no	Year of publishing	Author(s)	Title	Key research findings	References
19	2018	Jankovic et al.	Safety and Tolerability of Multiple Ascending Doses of PRX002/RG7935, an Anti- α -Synuclein Monoclonal Antibody, in Patients With Parkinson's Disease: A Randomized Clinical Trial	Multiple doses of PRX002 antibodies was found to be safe and tolerable even resulted in robust interaction with peripheral AS.	(94)
20	2018	Zhang et al.	A Comprehensive Analysis of the Association Between SNCA Polymorphisms and the Risk of Parkinson's Disease	Delineates SNPs in SCNA gene that pose risk of developing PD rs11931074 and rs2736990 increased the risk in East Asian group rs181489, rs356219, rs356165, rs2737029, and rs11931074 increased the risk for European group.	(95)

N-terminal acetylation in AS stability, toxicity, and aggregation (36). Overexpression of mutant A53T AS in HEK293 cells and rat dopaminergic neurons, was found to inhibit the 26S ubiquitin-proteasomal system and accumulation of AS phosphorylated at Ser129 was observed when the ubiquitin proteasomal system was inhibited (107).

INTERPLAY OF PROTEINS AND GENES IN PD PATHOGENESIS

Synaptic vesicle endocytosis, mitochondrial dynamics, and Lysosomal/Proteasomal activity are three critical processes that become dysfunctional during PD pathogenesis. The dysfunction of these processes could be because neurons are vulnerable to oxidative stress as well as accumulation of damaged organelles or misfolded proteins. The over expression of mutant proteins overloads the defense mechanisms of the dopaminergic neurons leading to disease pathogenesis. PINK1 mediates mitophagy either independently or through the recruitment of Parkin to the damaged/depolarized mitochondrial membrane. The PINK1/Parkin pathway is also important for mitochondria fusion and fission dynamics. Hence, mutations/deficiency of these proteins can lead to accumulation of defective mitochondria in the cell and lead to oxidative stress. Oxidized proteins are transported from the mitochondria to lysosomes through mitochondrial vesicles (MDVs) during oxidative stress due to reactive oxygen species and this process has also been suggested to be affected by dysfunction of the PINK-1/Parkin pathway (Figure 3). Similarly, DJ-1 acts as a sensor of oxidative stress in the neurons and it has been suggested to move into damaged mitochondria. DJ-1 also protects cells from apoptosis due to stress. Mutations leading to loss of function of DJ-1 have been suggested to lead to reduced lysosomal activity and autophagy. Furthermore, increased expression of AS due to mutation, gene duplication or triplication has been suggested to cause mitochondrial fragmentation and oxidative stress. Glucocerebrosidase coded by GBA has also been suggested to degrade AS in lysosomes and reduce AS toxicity. It is also responsible for the conversion of glucosylceramide into membrane constituents. Mutations of GBA have been shown to lead to dysfunction of lysosomal autophagy, dysfunctional

ubiquitin proteasome system, mitochondrial fragmentation, oxidative stress, and lead to defective lipid metabolism. Similarly, mutant LRRK2 under oxidative stress have been shown to cause apoptosis, defective vesicular trafficking and synthesis of defective/misfolded proteins (108, 109). Recently, mutated PINK1 expression in patient-derived neurons was found to lead to over-expression of LRRK2 at the mRNA and protein level (110). This suggests that PINK1 modulates LRRK2 level in neurons. DJ-1 has been suggested to regulate the level of PINK-1 and AS in neurons (111).

Parkin has been suggested to regulate the endolysosomal system by ubiquitination of Rab7. It has also been shown to interact with endophilin A (which is involved in autophagy and endocytosis) as well as other proteins involved in neuron survival. LRRK2 is also involved in endophilin A phosphorylation. Mutant AS has been suggested to prevent their own degradation in lysosomes which could be through Rab1a inhibition. Reduced activity of GCase1 due to GBA mutations has been shown to lead to accumulation of glucosylceramides and glycosphingolipids which may lead to greater production of Lewy Bodies and accelerate AS aggregation. Mutations in GBA are also suggested to cause the GCase protein to be stuck in the endoplasmic reticulum leading to ER stress, AS accumulation and interference with lysosomes. Misfolded AS aggregate and influence neurotransmission, synaptic vesicle exocytosis, recycling as well as endocytosis in the substantia nigra region. Furthermore, Aggregated AS has been suggested to target the retromer pathway of endolysosomal trafficking. The AS fibrils/filaments have been suggested to spread to surrounding regions through synapses in a prion-like manner through endocytosis and this has been suggested to evade lysosomal/proteasomal degradation through some way. Moreover, mutant/pathogenic LRRK2 with increased kinase-activity has been speculated to lead to exocytosis of misfolded AS fibrils along with increased endocytosis of such fibrils, which may help these fibrils to spread from cell-to-cell. AS has been suggested to promote the movement of pro-inflammatory monocytes into the substantia nigra region and also lead to the presentation of MHCII leading to inflammation and degeneration of neurons in this region. Although the activity of PINK-1/Parkin can help alleviate this process, mutated PINK-1/Parkin could not function this way (111, 112).

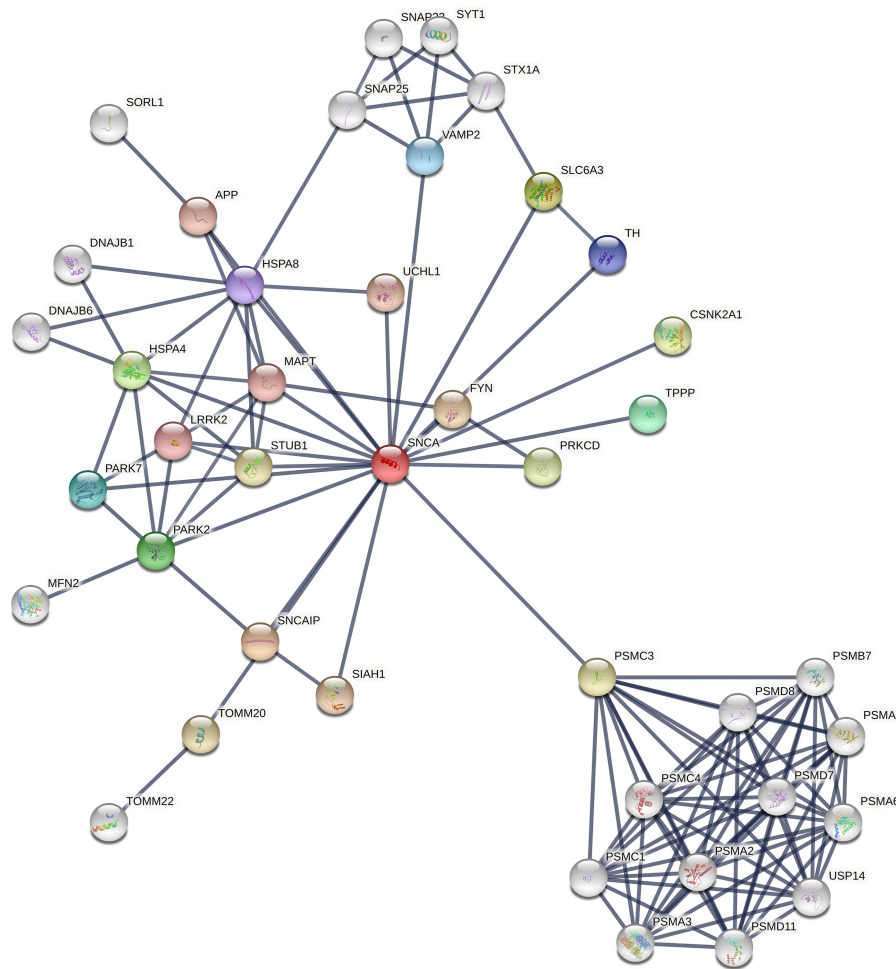


FIGURE 3 | Snapshot of interaction network of the different Proteins involved in PD pathogenesis from STRING Database. The thickness of lines indicates the confidence score for the interaction between two proteins.

ROLE OF PERSONALIZED MEDICINE IN MITIGATING PD PATHOLOGY

Till now, PD's heterogenic pathology has been well-delineated. Further, it paves way for the formulation of personalized therapeutics, wherein it is surmised that pathophysiology and genetic might be crucial. Due to this heterogeneity, an imperative need for developing individualized pathology mitigating therapies have been imposed upon the medical society. Recent medical advancement has made us witness, treatments effectuated upon PD patients with specific genetic aberration.

Glucocerebrosidase is a common risk factor associated with PD. Mutations in GBA gene, which encodes glucocerebrosidase, are found predominantly among people from Netherlands and Ashkenazi Jews. Though GBA's role in PD remains obscure, targeted therapeutics can be steered toward restoration of enzyme function and regulation of glycosphingolipid turnover. Accordingly, gene therapy through adeno-associated virus expression of glucocerebrosidase has shown positive findings in GBA pre-clinical studies. And reduced levels of alpha-synuclein

aggregates were also reported in the same. Further, Ambroxol is currently being investigated in a PD clinical trial.

Like GBA, genetic mutations in LRRK2 are also prevalent in certain ethnicities and also considered to be a credible risk factor for autosomal dominant PD. Enhancing LRRK2 kinase seems to have a beneficial effect. At present, various structurally distinct LRRK2 inhibitors formulated by Merck, Pfizer, Genentech and GSK are in the pipeline.

Of note, the personalized medicine trials are more laborious and complicated than the typical clinical trial and can face various issues including the need for specific genotype patients in large number. Nonetheless, with the SCNA gene and its polymorphisms being studied and characterized extensively, it holds a great potential in the field of personalized medicine, to effectively mitigate toxic AS aggregates (113–115).

CONCLUSION

Though AS's precise neuronal role remains obscure, it still persists as one of the crucial causalities behind PD. Its aberrant

conformational change (due to genetic mutations and other unknown pathophysiological circumstances) into toxic amyloid fibrils disrupts cellular homeostasis and effectuates neuronal deterioration. The present review covers a broad range of information on the disease-causing gene AS related to PD disorder. By contrast with the previous century the current era of medical ailments and research studies had witnessed astonishing progress in understanding PD, particularly with respects to the clinical studies, genetics, pathophysiology and molecular mechanism of PD. Though prodigious, the intricacy in understanding the disease pathology of PD is considered to be a promising approach toward the significant therapeutic targets. Besides, personalized medicine for the treatment of the PD considering the specificity of the defective genes will be the forthcoming approach over the next decades in the field of medical research world. Moreover, the directed delivery of therapies toward the central nervous system crossing the blood brain barrier will embolden credible hope that aid in improvising the new treatment for the devastating PD disorder.

REFERENCES

- Goedert M. Alpha-synuclein and neurodegenerative diseases. *Nat Rev Neurosci.* (2001) 2:492–501. doi: 10.1038/35081564
- Lee VM-Y, Trojanowski JQ. Mechanisms of parkinson's disease linked to pathological α -Synuclein: new targets for drug discovery. *Neuron.* (2006) 52:33–8. doi: 10.1016/j.neuron.2006.09.026
- Spillantini MG, Crowther RA, Jakes R, Hasegawa M, Goedert M. α -Synuclein in filamentous inclusions of Lewy bodies from Parkinson's disease and dementia with Lewy bodies. *Proc Natl Acad Sci USA.* (1998) 95:6469–73.
- Lotharius J, Brundin P. Pathogenesis of Parkinson's disease: dopamine, vesicles and alpha-synuclein. *Nat Rev Neurosci.* (2002) 3:932–42. doi: 10.1038/nrn983
- Liu S, Ninan I, Antonova I, Battaglia F, Trinchese F, Narasanna A, et al. alpha-Synuclein produces a long-lasting increase in neurotransmitter release. *EMBO J.* (2004) 23:4506–16. doi: 10.1038/sj.emboj.7600451
- Davie CA. A review of Parkinson's disease. *Br Med Bull.* (2008) 86:109–27. doi: 10.1093/bmb/ldn013
- Dickson DW. Neuropathology of Parkinson disease. *Parkinsonism Relat Disord.* (2018) 46:S30–3. doi: 10.1016/j.parkreldis.2017.07.033
- Dauer W, Przedborski S. Parkinson's disease: mechanisms and models. *Neuron.* (2003) 39:889–909. doi: 10.1016/s0896-6273(03)00568-3
- Jankovic J. Parkinson's disease: clinical features and diagnosis. *J Neurol Neurosurg Psychiatry.* (2008) 79:368–76. doi: 10.1136/jnnp.2007.131045
- Tseng M-T, Lin C-H. Pain in early-stage Parkinson's disease: implications from clinical features to pathophysiology mechanisms. *J Formosan Med Assoc.* (2017) 116:571–81. doi: 10.1016/j.jfma.2017.04.024
- Giasson BI, Murray IV, Trojanowski JQ, Lee VM, A. hydrophobic stretch of 12 amino acid residues in the middle of alpha-synuclein is essential for filament assembly. *J Biol Chem.* (2001) 276:2380–6. doi: 10.1074/jbc.M008919200
- Hoyer W, Cherny D, Subramaniam V, Jovin TM. Impact of the acidic C-terminal region comprising amino acids 109–140 on alpha-synuclein aggregation in vitro. *Biochemistry.* (2004) 43:16233–42. doi: 10.1021/bi048453u
- Jao CC, Hegde BG, Chen J, Haworth IS, Langen R. Structure of membrane-bound α -synuclein from site-directed spin labeling and computational refinement. *Proc Natl Acad Sci USA.* (2008) 105:19666–71. doi: 10.1073/pnas.0807826105

AUTHOR CONTRIBUTIONS

ES involved in conceptualization and writing and editing the final manuscript. GC and PC involved in sub-section of computational studies on alpha-synuclein. KA, AV, and RK involved in english correction and edited the manuscript. RR and PS involved in the design the review, supervision of the research work, and edited the final manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

ES, GC, PC, and RR thank the management of Vellore Institute of Technology (Deemed to be University) for providing the lab space. ES and KA thank the management of Saveetha School of Engineering, SIMATS. RK and PS thank the management of AIMST University for the support.

- Periquet M, Fulga T, Myllykangas L, Schlossmacher MG, Feany MB. Aggregated alpha-synuclein mediates dopaminergic neurotoxicity in vivo. *J Neurosci.* (2007) 27:3338–46. doi: 10.1523/JNEUROSCI.0285-07.2007
- Serpell LC, Berriman J, Jakes R, Goedert M, Crowther RA. Fiber diffraction of synthetic alpha-synuclein filaments shows amyloid-like cross-beta conformation. *Proc Natl Acad Sci USA.* (2000) 97:4897–902. doi: 10.1073/pnas.97.9.4897
- Vilar M, Chou H-T, Lührs T, Maji SK, Riek-Loher D, Verel R, et al. The fold of α -synuclein fibrils. *Proc Natl Acad Sci USA.* (2008) 105:8637–42. doi: 10.1073/pnas.0712179105
- Shastri BS. Neurodegenerative disorders of protein aggregation. *Neurochem Int.* (2003) 43:1–7. doi: 10.1016/s0197-0186(02)00196-1
- Ross CA, Poirier MA. Protein aggregation and neurodegenerative disease. *Nat Med.* (2004) 10 (Suppl.):S10–7. doi: 10.1038/nm1066
- Lázaro DF, Bellucci A, Brundin P, Outeiro TF. Editorial: protein misfolding and spreading pathology in neurodegenerative diseases. *Front Mol Neurosci.* (2020) 12:312. doi: 10.3389/fnmol.2019.00312
- Wang W, Nema S, Teagarden D. Protein aggregation—Pathways and influencing factors. *Int J Pharm.* (2010) 390:89–99. doi: 10.1016/j.ijpharm.2010.02.025
- Hunn BHM, Cragg SJ, Bolam JP, Spillantini M-G, Wade-Martins R. Impaired intracellular trafficking defines early Parkinson's disease. *Trends Neurosci.* (2015) 38:178–88. doi: 10.1016/j.tins.2014.12.009
- Franco R, Rivas-Santisteban R, Navarro G, Pinna A, Reyes-Resina I. Genes implicated in familial Parkinson's disease provide a dual picture of nigral dopaminergic neurodegeneration with mitochondria taking center stage. *Int J Mol Sci.* (2021) 22:4643. doi: 10.3390/ijms22094643
- Cooper JE, Machiela E, Dues DJ, Spielbauer KK, Senchuk MM, Van Raamsdonk JM. Activation of the mitochondrial unfolded protein response promotes longevity and dopamine neuron survival in Parkinson's disease models. *Sci Rep.* (2017) 7:16441. doi: 10.1038/s41598-017-16637-2
- Pickrell AM, Youle RJ. The roles of PINK1, parkin and mitochondrial fidelity in Parkinson's disease. *Neuron.* (2015) 85:257–73. doi: 10.1016/j.neuron.2014.12.007
- Kluss JH, Mamais A, Cookson MR. LRRK2 links genetic and sporadic Parkinson's disease. *Biochem Soc Trans.* (2019) 47:651–61. doi: 10.1042/BST20180462
- Flagmeier P, Meisl G, Vendruscolo M, Knowles TPJ, Dobson CM, Buell AK, et al. Mutations associated with familial Parkinson's disease alter the initiation and amplification steps of α -synuclein aggregation. *Proc Natl Acad Sci USA.* (2016) 113:10328–33. doi: 10.1073/pnas.1604645113

27. Li J-Q, Tan L, Yu J-T. The role of the LRRK2 gene in Parkinsonism. *Mol Neurodegener.* (2014) 9:47. doi: 10.1186/1750-1326-9-47
28. Bouhouche A, Tibar H, Ben El Haj R, El Bayad K, Razine R, Tazrout S, et al. LRRK2 G2019S mutation: prevalence and clinical features in moroccans with Parkinson's disease. *Parkinsons Dis.* (2017) 2017:2412486. doi: 10.1155/2017/2412486
29. Subramaniam SR, Chesselet M-F. Mitochondrial dysfunction and oxidative stress in Parkinson's disease. *Prog Neurobiol.* (2013) 0:17–32. doi: 10.1016/j.pneurobio.2013.04.004
30. Chen C, Turnbull DM, Reeve AK. Mitochondrial dysfunction in Parkinson's disease—Cause or consequence? *Biology.* (2019) 8:38. doi: 10.3390/biology8020038
31. Moon HE, Paek SH. Mitochondrial dysfunction in Parkinson's disease. *Exp Neurobiol.* (2015) 24:103–16. doi: 10.5607/en.2015.24.2.103
32. Goswami P, Joshi N, Singh S. Neurodegenerative signaling factors and mechanisms in Parkinson's pathology. *Toxicol Vitro.* (2017) 43:104–12. doi: 10.1016/j.tiv.2017.06.008
33. Uversky VNA. protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J Biomol Struct Dyn.* (2003) 21:211–34. doi: 10.1080/07391102.2003.10506918
34. Stefanis L. α -Synuclein in Parkinson's disease. *Cold Spring Harb Perspect Med.* (2012) 2:a009399. doi: 10.1101/cshperspect.a009399
35. Ulmer TS, Bax A, Cole NB, Nussbaum RL. Structure and dynamics of micelle-bound human alpha-synuclein. *J Biol Chem.* (2005) 280:9595–603. doi: 10.1074/jbc.M411805200
36. Vinuela-Gavilanes R, Iñigo-Marco I, Larrea L, Lasa M, Carte B, Santamaría E, et al. N-terminal acetylation mutants affect alpha-synuclein stability, protein levels and neuronal toxicity. *Neurobiol Dis.* (2020) 137:104781. doi: 10.1016/j.nbd.2020.104781
37. Sidhu A, Segers-Nolten I, Subramaniam V. Conformational compatibility is essential for heterologous aggregation of α -Synuclein. *ACS Chem Neurosci.* (2016) 7:719–27. doi: 10.1021/acschemneuro.5b00322
38. Butler B, Sambo D, Khoshbouei H. Alpha-synuclein modulates dopamine neurotransmission. *J Chem Neuroanat.* (2017) 83–4:41–9. doi: 10.1016/j.jchemneu.2016.06.001
39. Bendor J, Logan T, Edwards RH. The function of α -Synuclein. *Neuron.* (2013) 79:1044–66. doi: 10.1016/j.neuron.2013.09.004
40. Robotta M, Cattani J, Martins JC, Subramaniam V, Drescher M. Alpha-synuclein disease mutations are structurally defective and locally affect membrane binding. *J Am Chem Soc.* (2017) 139:4254–7. doi: 10.1021/jacs.6b05335
41. Lv Z, Krasnoslobodtsev AV, Zhang Y, Ysselstein D, Rochet J-C, Blanchard SC, et al. Direct detection of α -Synuclein dimerization dynamics: single-molecule fluorescence analysis. *Biophys J.* (2015) 108:2038–47. doi: 10.1016/j.bpj.2015.03.010
42. Kumar S, Jangir DK, Kumar R, Kumari M, Bhavesh NS, Maiti TK. Role of sporadic Parkinson disease associated mutations A18T and A29S in enhanced α -Synuclein fibrillation and cytotoxicity. *ACS Chem Neurosci.* (2017) 9:230–40. doi: 10.1021/acschemneuro.6b00430
43. Dikiy I, Fauvet B, Jovičić A, Mahul-Mellier A-L, Desobry C, El-Turk F, et al. Semisynthetic and *in Vitro* phosphorylation of Alpha-Synuclein at Y39 promotes functional partly helical membrane-bound states resembling those induced by PD mutations. *ACS Chem Biol.* (2016) 11:2428–37. doi: 10.1021/acschembio.6b00539
44. Salveson PJ, Spencer RK, Nowick JS. X-ray Crystallographic structure of oligomers formed by a toxic β -hairpin derived from α -Synuclein: trimers and higher-order oligomers. *J Am Chem Soc.* (2016) 138:4458–67. doi: 10.1021/jacs.5b13261
45. Corrado L, De Marchi F, Tunesi S, Oggioni GD, Carecchio M, Magistrelli L, et al. The length of SNCA Rep1 microsatellite may influence cognitive evolution in Parkinson's disease. *Front Neurol.* (2018) 9:213. doi: 10.3389/fneur.2018.00213
46. Zhang K, Tang Y, Meng L, Zhu L, Zhou X, Zhao Y, et al. The effects of SNCA rs894278 on resting-state brain activity in Parkinson's disease. *Front Neurosci.* (2019) 13:47. doi: 10.3389/fnins.2019.00047
47. Campêlo CL, das C, Silva RH. Genetic variants in SNCA and the risk of sporadic Parkinson's disease and clinical outcomes: a review. *Parkinsons Dis.* (2017) 2017:4318416. doi: 10.1155/2017/4318416
48. Miyake Y, Tanaka K, Fukushima W, Kiyohara C, Sasaki S, Tsuboi Y, et al. SNCA polymorphisms, smoking, and sporadic Parkinson's disease in Japanese. *Parkinsonism Relat Disord.* (2012) 18:557–61. doi: 10.1016/j.parkreldis.2012.02.016
49. Tittle MD, Comellas G, Nieuwkoop AJ, Covell DJ, Berthold DA, Kloepper KD, et al. Solid-state NMR structure of a pathogenic fibril of full-length human α -synuclein. *Nat Struct Mol Biol.* (2016) 23:409–15. doi: 10.1038/nsmb.3194
50. Li B, Ge P, Murray KA, Sheth P, Zhang M, Nair G, et al. Cryo-EM of full-length α -synuclein reveals fibril polymorphs with a common structural kernel. *Nat Commun.* (2018) 9:3609. doi: 10.1038/s41467-018-05971-2
51. Guerrero-Ferreira R, Taylor NM, Arteni A-A, Kumari P, Mona D, Ringler P, et al. Two new polymorphic structures of human full-length alpha-synuclein fibrils solved by cryo-electron microscopy. *Elife.* (2019) 8:e48907. doi: 10.7554/eLife.48907
52. Cali T, Ottolini D, Negro A, Brini M. α -Synuclein controls mitochondrial calcium homeostasis by enhancing endoplasmic reticulum-mitochondria interactions. *J Biol Chem.* (2012) 287:17914–29. doi: 10.1074/jbc.M111.302794
53. Vicario M, Cieri D, Brini M, Cali T. The close encounter between Alpha-synuclein and mitochondria. *Front Neurosci.* (2018) 12:388. doi: 10.3389/fnins.2018.00388
54. Perez RG, Waymire JC, Lin E, Liu JJ, Guo F, Zigmond MJ, et al. Role for α -Synuclein in the regulation of dopamine biosynthesis. *J Neurosci.* (2002) 22:3090–9. doi: 10.1523/JNEUROSCI.22-08-03090.2002
55. Cook C, Stetler C, Petrucelli L. Disruption of protein quality control in Parkinson's disease. *Cold Spring Harb Perspect Med.* (2012) 2:a009423. doi: 10.1101/cshperspect.a009423
56. Alter SP, Lenzi GM, Bernstein AI, Miller GW. Vesicular integrity in Parkinson's disease. *Curr Neurol Neurosci Rep.* (2013) 13:362. doi: 10.1007/s11910-013-0362-3
57. Chen R, Furman CA, Gnegy ME. Dopamine transporter trafficking: rapid response on demand. *Future Neurol.* (2010) 5:123. doi: 10.2217/fnl.09.76
58. Snead D, Eliez D. Alpha-synuclein function and dysfunction on cellular membranes. *Exp Neurobiol.* (2014) 23:292–313. doi: 10.5607/en.2014.23.4.292
59. Fusco G, Sanz-Hernandez M, De Simone A. Order and disorder in the physiological membrane binding of α -synuclein. *Curr Opin Struct Biol.* (2018) 48:49–57. doi: 10.1016/j.sbi.2017.09.004
60. Vargas KJ, Schrod N, Davis T, Fernandez-Busnadiego R, Taguchi YV, Laugks U, et al. Synucleins have multiple effects on presynaptic architecture. *Cell Rep.* (2017) 18:161–73. doi: 10.1016/j.celrep.2016.12.023
61. Eguchi K, Taoufiq Z, Thorn-Seshold O, Trauner D, Hasegawa M, Takahashi T. Wild-type monomeric α -Synuclein can impair vesicle endocytosis and synaptic fidelity via tubulin polymerization at the calyx of held. *J Neurosci.* (2017) 37:6043–52. doi: 10.1523/JNEUROSCI.0179-17.2017
62. Fakhree MAA, Nolten IS, Blum C, Claessens MMAE. Different conformational subensembles of the intrinsically disordered protein α -Synuclein in cells. *J Phys Chem Lett.* (2018) 9:1249–53. doi: 10.1021/acs.jpclett.8b00092
63. Chaudhary H, Iyer A, Subramaniam V, Claessens MMAE. α -Synuclein oligomers stabilize pre-existing defects in supported bilayers and propagate membrane damage in a fractal-like pattern. *Langmuir.* (2016) 32:11827–36. doi: 10.1021/acs.langmuir.6b02572
64. Wakabayashi K, Takahashi H. α -Synuclein, synphilin-1 and inclusion body formation in α -synucleinopathies. *Int Congress Ser.* (2003) 1251:149–56. doi: 10.1016/S0531-5131(03)00115-8
65. Zhang Q-S, Heng Y, Yuan Y-H, Chen N-H. Pathological α -synuclein exacerbates the progression of Parkinson's disease through microglial activation. *Toxicol Lett.* (2017) 265:30–7. doi: 10.1016/j.toxlet.2016.11.002
66. Binolfi A, Rodriguez EE, Valensin D, D'Amelio N, Ippoliti E, Obal G, et al. Bioinorganic chemistry of Parkinson's disease: structural determinants for the copper-mediated amyloid formation of Alpha-Synuclein. *Inorg Chem.* (2010) 49:10668–79. doi: 10.1021/ic1016752
67. Miotto MC, Valiente-Gabioud AA, Rossetti G, Zweckstetter M, Carloni P, Selenko P, et al. Copper binding to the N-terminally acetylated, naturally occurring form of Alpha-Synuclein induces local helical folding. *J Am Chem Soc.* (2015) 137:6444–7. doi: 10.1021/jacs.5b01911

68. Iyer A, Roeters SJ, Kogan V, Woutersen S, Claessens MMAE, Subramaniam V. C-Terminal truncated α -Synuclein fibrils contain strongly twisted β -sheets. *J Am Chem Soc.* (2017) 139:15392–400. doi: 10.1021/jacs.7b07403
69. Mason RJ, Paskins AR, Dalton CE, Smith DP. Copper binding and subsequent aggregation of α -Synuclein are modulated by N-terminal acetylation and ablated by the H50Q missense mutation. *Biochemistry.* (2016) 55:4737–41. doi: 10.1021/acs.biochem.6b00708
70. Dučić T, Carboni E, Lai B, Chen S, Michalke B, Lázaro DF, et al. Alpha-Synuclein regulates neuronal levels of manganese and calcium. *ACS Chem Neurosci.* (2015) 6:1769–79. doi: 10.1021/acschemneuro.5b00093
71. Finkelstein DJ, Hare DJ, Billings JL, Sedjahtera A, Nurjono M, Arthofer E, et al. Clioquinol improves cognitive, motor function, and microanatomy of the Alpha-Synuclein hA53T transgenic mice. *ACS Chem Neurosci.* (2016) 7:119–29. doi: 10.1021/acschemneuro.5b00253
72. Szénási T, Oláh J, Szabó A, Szunyogh S, Láng A, Perczel A, et al. Challenging drug target for Parkinson's disease: Pathological complex of the chameleon TPPP/p25 and alpha-synuclein proteins. *Biochim Biophys Acta.* (2017) 1863:310–23. doi: 10.1016/j.bbdis.2016.09.017
73. Oláh J, Bertrand P, Ovádi J. Role of the microtubule-associated TPPP/p25 in Parkinson's and related diseases and its therapeutic potential. *Expert Rev Proteomics.* (2017) 14:301–9. doi: 10.1080/14789450.2017.1304216
74. Ejlerskov P, Rasmussen I, Nielsen TT, Bergström A-L, Tohyama Y, Jensen PH, et al. Tubulin Polymerization-promoting Protein (TPPP/p25 α) promotes unconventional secretion of α -Synuclein through exophagy by impairing autophagosome-lysosome fusion. *J Biol Chem.* (2013) 288:17313–35. doi: 10.1074/jbc.M112.401174
75. Sidhu A, Segers-Nolten I, Raussens V, Claessens MMAE, Subramaniam V. Distinct mechanisms determine α -Synuclein fibril morphology during growth and maturation. *ACS Chem Neurosci.* (2017) 8:538–47. doi: 10.1021/acschemneuro.6b00287
76. Mahul-Mellier A-L, Bartscher J, Maharjan N, Weerens L, Croisier M, Kuttler F, et al. The process of Lewy body formation, rather than simply α -synuclein fibrillization, is one of the major drivers of neurodegeneration. *Proc Natl Acad Sci USA.* (2020) 117:4971–82. doi: 10.1073/pnas.1913904117
77. Ray S, Singh N, Kumar R, Patel K, Pandey S, Datta D, et al. α -Synuclein aggregation nucleates through liquid-liquid phase separation. *Nat Chem.* (2020) 12:705–16. doi: 10.1038/s41557-020-0465-9
78. Van der Perren A, Gelders G, Fenyi A, Bousset L, Brito F, Peelaerts W, et al. The structural differences between patient-derived α -synuclein strains dictate characteristics of Parkinson's disease, multiple system atrophy and dementia with Lewy bodies. *Acta Neuropathol.* (2020) 139:977–1000. doi: 10.1007/s00401-020-02157-3
79. Stephens AD, Zacharopoulou M, Moons R, Fusco G, Seetaloo N, Chiki A, et al. Extent of N-terminus exposure of monomeric alpha-synuclein determines its aggregation propensity. *Nat Commun.* (2020) 11:2820. doi: 10.1038/s41467-020-16564-3
80. Zhao K, Lim Y-J, Liu Z, Long H, Sun Y, Hu J-J, et al. Parkinson's disease-related phosphorylation at Tyr39 rearranges α -synuclein amyloid fibril structure revealed by cryo-EM. *Proc Natl Acad Sci USA.* (2020) 117:20305–15. doi: 10.1073/pnas.1922741117
81. Lindestam Arlehamn CS, Dhanwani R, Pham J, Kuan R, Frazier A, Rezende Dutra J, et al. α -Synuclein-specific T cell reactivity is associated with preclinical and early Parkinson's disease. *Nat Commun.* (2020) 11:1875. doi: 10.1038/s41467-020-15626-w
82. Niu M, Li Y, Li G, Zhou L, Luo N, Yao M, et al. longitudinal study on α -synuclein in plasma neuronal exosomes as a biomarker for Parkinson's disease development and progression. *Eur J Neurol.* (2020) 27:967–74. doi: 10.1111/ene.14208
83. Bhattacharjee P, Öhrfelt A, Lashley T, Blennow K, Brinkmalm A, Zetterberg H. Mass spectrometric analysis of lewy body-enriched α -Synuclein in Parkinson's disease. *J Proteome Res.* (2019) 18:2109–20. doi: 10.1021/acs.jproteome.8b00982
84. Levine PM, Galesic A, Balana AT, Mahul-Mellier A-L, Navarro MX, De Leon CA, et al. α -Synuclein O-GlcNAcylation alters aggregation and toxicity, revealing certain residues as potential inhibitors of Parkinson's disease. *Proc Natl Acad Sci USA.* (2019) 116:1511–9. doi: 10.1073/pnas.1808845116
85. Tian C, Liu G, Gao L, Soltys D, Pan C, Stewart T, et al. Erythrocytic α -Synuclein as a potential biomarker for Parkinson's disease. *Transl Neurodegener.* (2019) 8:15. doi: 10.1186/s40035-019-0155-y
86. Elfarrash S, Jensen NM, Ferreira N, Betzer C, Thevathasan JV, Diekmann R, et al. Organotypic slice culture model demonstrates inter-neuronal spreading of alpha-synuclein aggregates. *Acta Neuropathol Commun.* (2019) 7:213. doi: 10.1186/s40478-019-0865-5
87. Landeck N, Buck K, Kirik D. Toxic effects of human and rodent variants of alpha-synuclein *in vivo*. *Eur J Neurosci.* (2017) 45:536–47. doi: 10.1111/ejn.13493
88. Brys M, Fanning L, Hung S, Ellenbogen A, Penner N, Yang M, et al. Randomized phase I clinical trial of anti- α -synuclein antibody BIIB054. *Mov Disord.* (2019) 34:1154–63. doi: 10.1002/mds.27738
89. Kang UJ, Boehme AK, Fairfoul G, Shahnawaz M, Ma TC, Hutten SJ, et al. Comparative study of cerebrospinal fluid α -synuclein seeding aggregation assays for diagnosis of Parkinson's disease. *Mov Disord.* (2019) 34:536–44. doi: 10.1002/mds.27646
90. Pihlström L, Blauwendraat C, Cappelletti C, Berge-Seidl V, Langmyhr M, Henriksen SP, et al. A comprehensive analysis of SNCA-related genetic risk in sporadic Parkinson disease. *Ann Neurol.* (2018) 84:117–29. doi: 10.1002/ana.25274
91. Longhena F, Faustini G, Varanita T, Zaltieri M, Porrini V, Tessari I, et al. Synapsin III is a key component of α -synuclein fibrils in Lewy bodies of PD brains. *Brain Pathol.* (2018) 28:875–88. doi: 10.1111/bpa.12587
92. Faustini G, Longhena F, Varanita T, Bubacco L, Pizzi M, Missale C, et al. Synapsin III deficiency hampers α -synuclein aggregation, striatal synaptic damage and nigral cell loss in an AAV-based mouse model of Parkinson's disease. *Acta Neuropathol.* (2018) 136:621–39. doi: 10.1007/s00401-018-1892-1
93. Papagiannakis N, Koros C, Stamelou M, Simitsi A-M, Maniati M, Antonelou R, et al. Alpha-synuclein dimerization in erythrocytes of patients with genetic and non-genetic forms of Parkinson's Disease. *Neurosci Lett.* (2018) 672:145–9. doi: 10.1016/j.neulet.2017.11.012
94. Jankovic J, Goodman I, Safirstein B, Marmon TK, Schenk DB, Koller M, et al. Safety and tolerability of multiple ascending doses of PRX002/RG7935, an anti- α -Synuclein monoclonal antibody, in patients with Parkinson disease: a randomized clinical trial. *JAMA Neurol.* (2018) 75:1206–14. doi: 10.1001/jamaneurol.2018.1487
95. Zhang Y, Shu L, Sun Q, Pan H, Guo J, Tang B, et al. Comprehensive analysis of the association between SNCA polymorphisms and the risk of Parkinson's disease. *Front Mol Neurosci.* (2018) 11:391. doi: 10.3389/fnmol.2018.00391
96. Javed H, Nagoor Meeran MF, Azimullah S, Adem A, Sadek B, Ojha SK. Plant extracts and phytochemicals targeting α -Synuclein aggregation in Parkinson's disease models. *Front Pharmacol.* (2019) 9:1555. doi: 10.3389/fphar.2018.01555
97. Fields CR, Bengoa-Vergniory N, Wade-Martins R. Targeting Alpha-Synuclein as a therapy for Parkinson's disease. *Front Mol Neurosci.* (2019) 12:299. doi: 10.3389/fnmol.2019.00299
98. Healey MA. *Computational Study of Alpha-synuclein Structure and Druggability*. University of Alberta (2016).
99. Khurana V, Peng J, Chung CY, Auluck PK, Fanning S, Tardiff DE, et al. Genome-scale networks link neurodegenerative disease genes to α -Synuclein through specific molecular pathways. *Cell Syst.* (2017) 4:157–70.e14. doi: 10.1016/j.cels.2016.12.011
100. Srinivasan E, Rajasekaran R. Computational investigation of curcumin, a natural polyphenol that inhibits the destabilization and the aggregation of human SOD1 mutant (Ala4Val). *RSC Adv.* (2016) 6:102744–53. doi: 10.1039/C6RA21927F
101. Srinivasan E, Rajasekaran R. Exploring the cause of aggregation and reduced Zn binding affinity by G85R mutation in SOD1 rendering amyotrophic lateral sclerosis. *Proteins.* (2017) 85:1276–86. doi: 10.1002/prot.25288
102. Srinivasan E, Rajasekaran R. Cysteine to serine conversion at 111th position renders the disaggregation and retains the stabilization of detrimental SOD1 A4V mutant against amyotrophic lateral sclerosis in human—A discrete molecular dynamics study. *Cell Biochem Biophys.* (2018) 76:231–41. doi: 10.1007/s12013-017-0830-5
103. Srinivasan E, Rajasekaran R. Effect of β -cyclodextrin-EGCG complexation against aggregated α -synuclein through density functional

- theory and discrete molecular dynamics. *Chem Phys Lett.* (2019) 717:38–46. doi: 10.1016/j.cplett.2018.12.042
104. Srinivasan E, Rajasekaran R. Molecular binding response of naringin and naringenin to H46R mutant SOD1 protein in combating protein aggregation using density functional theory and discrete molecular dynamics. *Prog Biophys Mol Biol.* (2019) 145:40–51. doi: 10.1016/j.pbiomolbio.2018.12.003
 105. Srinivasan E, Ravikumar S, Venkataramanan S, Purohit R, Rajasekaran R. Molecular mechanics and quantum chemical calculations unveil the combating effect of baicalein on human islet amyloid polypeptide aggregates. *Mol Simul.* (2019) 45:1538–48. doi: 10.1080/08927022.2019.1660778
 106. Davidi D, Schechter M, Elhadi SA, Matatov A, Nathanson L, Sharon R. α -Synuclein translocates to the nucleus to activate retinoic-acid-dependent gene transcription. *iScience.* (2020) 23:100910. doi: 10.1016/j.isci.2020.100910
 107. McKinnon C, De Snoo ML, Gondard E, Neudorfer C, Chau H, Ngana SG, et al. Early-onset impairment of the ubiquitin-proteasome system in dopaminergic neurons caused by α -synuclein. *Acta Neuropathol Commun.* (2020) 8:17. doi: 10.1186/s40478-020-0894-0
 108. Larsen SB, Hanss Z, Krüger R. The genetic architecture of mitochondrial dysfunction in Parkinson's disease. *Cell Tissue Res.* (2018) 373:21–37. doi: 10.1007/s00441-017-2768-8
 109. van der Vlag M, Havekes R, Heckman PRA. The contribution of Parkin, PINK1 and DJ-1 genes to selective neuronal degeneration in Parkinson's disease. *Eur J Neurosci.* (2020) 52:3256–68. doi: 10.1111/ejn.14689
 110. Azkona G, López de Maturana R, del Rio P, Sousa A, Vazquez N, Zubiarrain A, et al. LRRK2 Expression Is Deregulated in Fibroblasts and Neurons from Parkinson Patients with Mutations in PINK1. *Mol Neurobiol.* (2018) 55:506–16. doi: 10.1007/s12035-016-0303-7
 111. Zeng X-S, Geng W-S, Jia J-J, Chen L, Zhang P-P. Cellular and molecular basis of neurodegeneration in Parkinson disease. *Front Aging Neurosci.* (2018) 10:109. doi: 10.3389/fnagi.2018.00109
 112. Vidyadhara DJ, Lee JE, Chandra SS. Role of the endolysosomal system in Parkinson's disease. *J Neurochem.* (2019) 150:487–506. doi: 10.1111/jnc.14820
 113. Bu L-L, Yang K, Xiong W-X, Liu F-T, Anderson B, Wang Y, et al. Toward precision medicine in Parkinson's disease. *Ann Transl Med.* (2016) 4:26. doi: 10.3978/j.issn.2305-5839.2016.01.21
 114. Schneider SA, Alcalay RN. Precision medicine in Parkinson's disease: emerging treatments for genetic Parkinson's disease. *J Neurol.* (2020) 267:860–9. doi: 10.1007/s00415-020-09705-7
 115. Ryden LE, Lewis SJG. Parkinson's Disease in the era of personalised medicine: one size does not fit all. *Drugs Aging.* (2019) 36:103–13. doi: 10.1007/s40266-018-0624-5

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

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

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



Identification of Cross-Pathway Connections via Protein-Protein Interactions Linked to Altered States of Metabolic Enzymes in Cervical Cancer

Krishna Kumar*, Sarpita Bose and Saikat Chakrabarti*

Structural Biology and Bioinformatics Division, Council of Scientific & Industrial Research (CSIR)-Indian Institute of Chemical Biology, Kolkata, India

OPEN ACCESS

Edited by:

Thirumal Kumar D,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Jiansong Fang,
Guangzhou University of Chinese
Medicine, China
Vera Luiza Capelozzi,
University of São Paulo, Brazil

*Correspondence:

Krishna Kumar
krishna.kumarbt2009@gmail.com
Saikat Chakrabarti
saikat@iicb.res.in

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 05 July 2021

Accepted: 29 September 2021

Published: 01 November 2021

Citation:

Kumar K, Bose S and Chakrabarti S
(2021) Identification of Cross-Pathway
Connections via Protein-Protein
Interactions Linked to Altered States
of Metabolic Enzymes in Cervical
Cancer. *Front. Med.* 8:736495.
doi: 10.3389/fmed.2021.736495

Metabolic reprogramming is one of the emerging hallmarks of cancer cells. Various factors, such as signaling proteins (S), miRNA, and transcription factors (TFs), may play important roles in altering the metabolic status in cancer cells by interacting with metabolic enzymes either directly or via protein-protein interactions (PPIs). Therefore, it is important to understand the coordination among these cellular pathways, which may provide better insight into the molecular mechanism behind metabolic adaptations in cancer cells. In this study, we have designed a cervical cancer-specific supra-interaction network where signaling pathway proteins, TFs, and microRNAs (miRs) are connected to metabolic enzymes via PPIs to investigate novel molecular targets and connections/links/paths regulating the metabolic enzymes. Using publicly available omics data and PPIs, we have developed a Hidden Markov Model (HMM)-based mathematical model yielding 94, 236, and 27 probable links/paths connecting signaling pathway proteins, TFs, and miRNAs to metabolic enzymes, respectively, out of which 83 paths connect to six common metabolic enzymes (RRM2, NDUFA11, ENO2, EZH2, AKR1C2, and TYMS). Signaling proteins (e.g., PPARG, BAD, GNB5, CHECK1, PAK2, PLK1, BRCA1, MAML3, and SPP1), TFs (e.g., KAT2B, ING1, MED1, ZEB1, AR, NCOA2, EGR1, TWIST1, E2F1, ID4, RBL1, ESR1, and HSF2), and miR (e.g., mir-147a, mir-593-5p, mir-138-5p, mir-16-5p, and mir-15b-5p) were found to regulate two key metabolic enzymes, EZH2 and AKR1C2, with altered metabolites (L-lysine and tetrahydrodeoxycorticosterone, THDOC) status in cervical cancer. We believe, the biology-based approach of our system will pave the way for future studies, which could be aimed toward identifying novel signaling, transcriptional, and post-transcriptional regulators of metabolic alterations in cervical cancer.

Keywords: metabolic reprogramming, cervical cancer, mathematical modeling, systems biology, signaling pathway proteins, transcription factor, microRNA, metabolic enzymes

INTRODUCTION

Cervical cancer is the fourth most frequently occurring cancer and the fourth leading cause of death in women worldwide with an estimate of 5,70,000 cases and 3,11,000 deaths in 2018. Approximately 80–85% of the deaths from cervical cancer occur in lower and middle-income countries compared to high-income countries (1, 2). Squamous cell carcinoma (SCC) and adenocarcinomas are the two main types of cervical cancer. Above 90% of patients with cervical cancer belong to SCC (3). The persistent infection with human papillomavirus (HPV), a particularly high-risk type of HPV (mainly HPV16 and HPV18 type), is considered the primary cause of cervical cancer (4–6). Only HPV16 and HPV18 types are responsible for almost 70% of cases of cervical cancer globally (7). While infection by high-risk HPV is necessary for developing cervical cancer, it alone may not be sufficient. Various studies suggest that the pathogenesis of cervical cancer depends on various other factors acting in concert with disease-associated HPV types (8–10). Therefore, it is important to understand the molecular mechanism behind the development of cervical cancer.

Metabolic reprogramming is considered one of the emerging hallmarks of cancer cells, and it is essential for cancer cell growth and proliferation to evolve into a more aggressive malignant state (11, 12). Understanding the coordination among various biological pathways, such as gene-regulatory, signaling, and metabolic pathways is important and may provide clues into the molecular mechanism of metabolic adaptation in cancer and associated cells. To understand that, one needs to investigate the molecular mechanism by which the impact of signaling, transcriptional, and post-transcriptional aberration is transgressed to metabolic reprogramming. Various studies demonstrated that the metabolic status in cancer cells is regulated by oncogenic changes in signaling pathways (13–15), transcription factors (TFs) (16–18), and miRNAs (19–21). However, these studies are focused either on a single molecule or pathways and may not capture the complex interconnectivity among various biological processes.

To overcome the complexity of interconnected biological pathways, biological approaches to efficient systems need to be developed. Computational and/or mathematical model-based system biology approaches provide an effective way to discover new drug targets for cancer therapy (22, 23). Mathematical model-based system biology approaches are successful for signaling and metabolic network analyses (24–30). Mathematical models for signaling pathways have been developed based on logical models (27–30), kinetic models (31, 32), Petri nets (33), decision tree (34), ordinary differential equations (35), and linear programming (LP)-based model (22, 36). Previously, we also have established a Hidden Markov Model (HMM)-based mathematical model to analyze the signaling-metabolic (S-M) interconnecting networks (37).

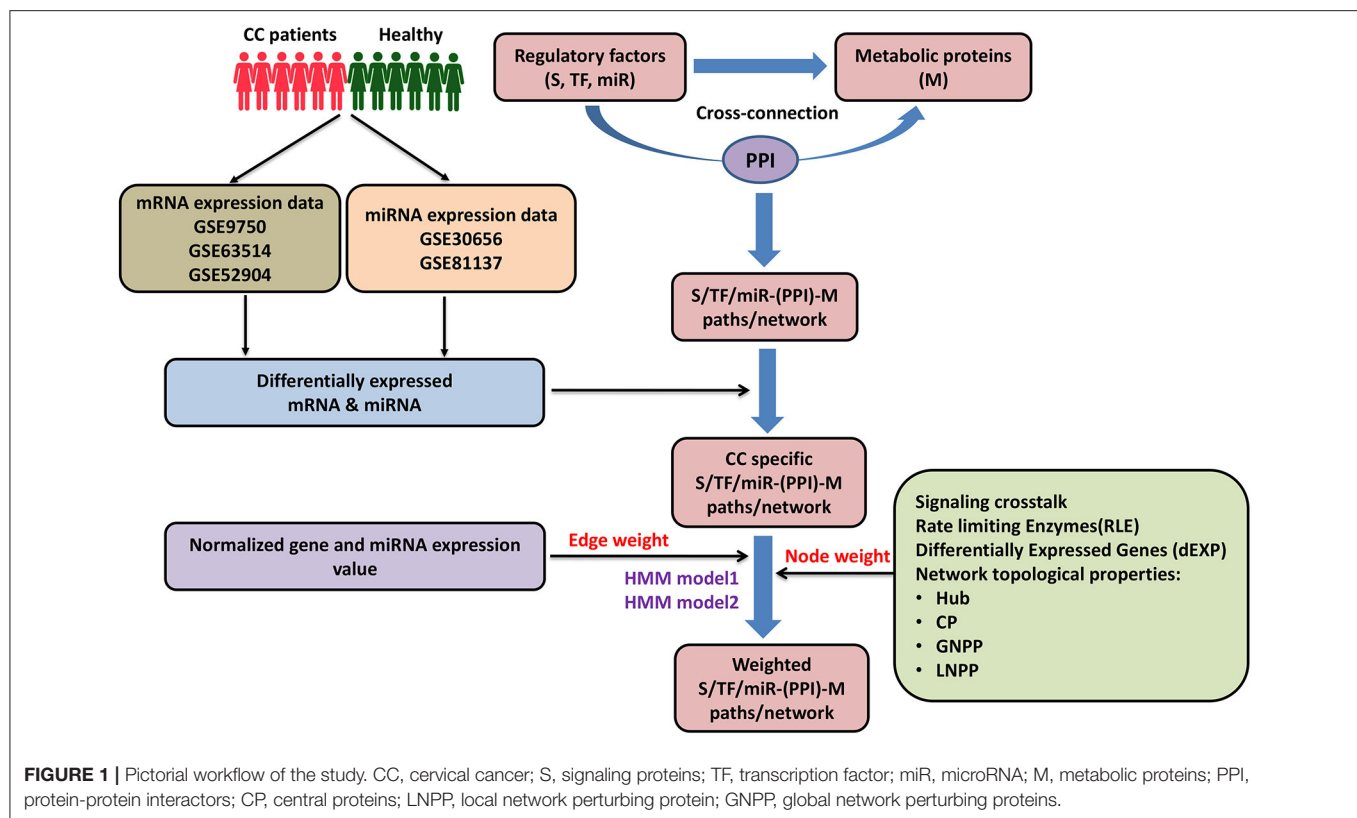
In the present study, we have designed a cervical cancer-specific supra-interaction network model incorporating transcriptome data onto a protein-protein interaction (PPIs) network to investigate novel molecular targets and connections regulating the status of metabolic enzymes. We have developed

a biology framework of a comprehensive system where signaling (S) pathway proteins, miRNA, and TF-based gene-regulatory modules are connected to metabolic (M) pathway proteins through protein-protein interactors (PPIs; **Figure 1**). Initially, network topologically IINs, such as a hub, central node (CN), local network perturbing nodes (LNPNs), and global network perturbing nodes (GNPN), were identified in different [S-M, TF-metabolic (TF-M), and miRNA-metabolic] modules of cervical cancer-specific networks using graph theory approach previously reported by our laboratory (38). These IINs may serve as potential diagnostic and/or prognostic biomarkers in cervical cancer. Furthermore, signaling pathway proteins, TFs, and microRNA (miR) to metabolic enzymes interconnecting paths/links [S-PPI-M, TF-target genes (TG)-PPI-M, and miR-TG-PPI-M] were identified in cervical cancer. Publicly available transcriptomic data derived from cervical cancer patients were incorporated into the HMM-based mathematical modeling set-up to weigh and rank the interconnecting link/paths in each module. Additional confidence values based on biological and network topological properties (hub, CN, GNPN, and LNPN) were assigned to each gene/protein/miRNA in the paths/links identified after model implementation to extract out high confident connections/links specific to cervical cancer. *In silico* validation of these selected genes/proteins/miRNAs and paths has been performed through perturbation analysis, demonstrating the importance of certain genes/proteins/miRNAs forming critical inter-pathway connections. PPI links connecting to key metabolic enzymes, such as RRM2, NDUFA11, ENO2, EZH2, AKR1C2, and TYMS, are identified from signaling proteins (e.g., PPARG, BAD, GNB5, CHECK1, PAK2, PLK1, BRCA1, MAML3, and SPP1), TFs (e.g., KAT2B, ING1, MED1, ZEB1, AR, NCOA2, EGR1, TWIST1, E2F1, ID4, RBL1, ESR1, and HSF2), and miR (e.g., mir-147a, mir-593-5p, mir-138-5p, mir-16-5p, and mir-15b-5p) in cervical cancer scenario. Out of the six metabolic enzymes that are commonly linked by 83 paths/links, EZH2 and AKR1C2 were mapped with deregulated metabolite status. Further, comparative analysis of the identified genes/proteins/miRNAs and the associated molecular pairs and paths in different modules were performed using transcriptomics data obtained from cervical, breast, and ovarian cancer patients. This study led to novel inter-bio-molecular links between signaling, gene-regulatory components, and metabolic enzymes paving the probable way(s) to identify drug targets to inhibit cervical cancer progression in a more specific manner.

MATERIALS AND METHODS

Messenger Ribonucleic Acid (mRNA) and Micro Ribonucleic Acid (miRNA) Expression Datasets

mRNA and miRNA expression datasets of cervical, breast cancer, and ovarian cancer patients were extracted from the Gene Expression Omnibus (GEO) database (39) to study the gene and miRNA expression profiles. For the mRNA expression datasets, similar microarray (Affymetrix microarrays) platforms were used for each cancer type to minimize the undesirable variations that



occurred due to different microarray platforms. Further, the cancer samples in each dataset were considered in a 1:1 ratio with normal samples to avoid sample heterogeneity (as shown in Table 1).

Differential Expression Analysis

The differential expression analysis of each dataset was performed separately using the GEO2R web tool (40–42) available at the GEO database. Genes having $\log_2FC \geq +1.5$ and $\log_2FC \leq -1.5$ were considered as upregulated and downregulated genes, respectively. The genes with \log_2FC values between -1.5 and $+1.5$ were considered only as neutrally expressed genes (EG). Benjamini and Hochberg's method (43) was used to control the false discovery rate. Genes with an adjusted $p \leq 0.05$ were considered significant. For ovarian cancer datasets, the \log_2FC and p -value threshold of ± 2 and ≤ 0.01 , respectively, were considered for the selection of upregulated, downregulated, and EG. For differential miRNA expression adjusted $p \leq 0.05$ and $\log_2FC \geq +1.0$ and $\log_2FC \leq -1.0$ were considered as thresholds for the identification of upregulated and downregulated miRNAs, respectively.

Construction of Human Protein-Protein Interaction Network

The HPPIN was constructed by extracting experimentally verified (confidence score ≥ 0.7) HPPIs available in the STRING v11.0 (44) database. The resulting network (proteins as nodes and

edges demark interaction) consisted of 5,048 proteins and 18,044 interactions, respectively.

Construction of Cancer-Specific Protein-Protein Interaction Network

Differentially expressed genes (dEXP) and EG from each cervical cancer dataset (GSE9750, GSE63514, and GSE52904; Table 1) were mapped onto the HPPIN to construct a cervical CC-PPIN. The interactions were considered up to the second level (i.e., interactors of interactors). In the first level, interactions mediated by only deregulated genes were considered where their interactors could be either deregulated or neutrally expressed. The resulting network consisted of 2,240 proteins interconnected via 5,452 edges. Similarly, breast and ovarian CC-PPINs were constructed using the corresponding transcriptomic datasets (Table 1).

Construction of Cancer-Specific Transcriptional Regulatory Network

The transcriptional regulatory network was constructed by collating deregulated TF-TG interaction and associated protein-protein interactions. Experimentally verified strong evidenced TF-TG interactions were retrieved from Human Transcriptional Regulation Interactions database (HTRIdb) (45) and Transcriptional Regulatory Relationships Unraveled by Sentence-based Text mining (TRUSST) (46) databases and were merged generating 22,480 interactions among 697 TFs and 12,407 TG. The combined deregulated genes (differentially

TABLE 1 | Differential expression analysis of mRNA and miRNA.

mRNA dataset		Origin	Normal samples	Patient samples	Up-Regulated	Down-Regulated	Neutrally expressed
Cervical cancer	GSE9750	USA	19	19	330	634	4,338
	GSE63514	USA	19	19	1,019	676	4,390
	GSE52904	Mexico	17	17	177	322	6,228
	Total	–	45	45	1,250	1,331	10,053
Breast cancer	GSE29044	Saudi Arabia	27	27	356	729	9,175
	GSE42568	Ireland	17	17	1,270	1,343	7,636
	GSE103512	USA	10	10	57	101	8,016
	Total	–	54	54	1,385	1,676	1,2717
Ovarian cancer	GSE38666	USA	12	12	2,733	954	4,847
	GSE54388	USA	6	6	232	527	4,087
	GSE66957	USA	12	12	2,201	872	9,028
	Total	–	30	30	4,481	1,901	11,560
miRNA dataset		Origin	Normal samples	Patient samples	Up-regulated	Down-regulated	Neutrally expressed
Cervical cancer	GSE30656	Netherland	10	10	4	12	NA
	GSE81137	India	3	3	15	20	NA
	Total	–	13	13	19	31	NA
Breast cancer	GSE45498	Switzerland	59	59	66	13	NA
	GSE97811	Japan	16	16	39	15	NA
	GSE143564	China	3	3	141	20	NA
	Total	–	78	78	197	40	NA
Ovarian cancer	GSE23383	USA	3	3	53	15	NA
	GSE47841	Norway	9	9	52	105	NA
	GSE119056	China	3	3	55	147	NA
	Total	–	15	15	137	245	NA

expressed and neutrally expressed) of all the three datasets (GSE9750, GSE63514, and GSE52904) were mapped on to TF-TG interactions to filter the cervical cancer-specific TF-TG interactions. In this study, only those interactions were considered where both TFs and their TG were deregulated. The deregulated TF-TG were further mapped to experimentally verify HPPIs up to the second level, and only those interactions were considered where the interacting partners were either deregulated or expressed. Finally, the filtered TF-TG and PPI interactions were merged to form the TF-TG-PPI network. The resulting network consisted of 2,894 nodes interconnected via 5,694 edges. Breast and ovarian cancer-specific TF-TG-PPI networks were also constructed with the corresponding mRNA transcriptomics data (Table 1) using the same protocol.

Construction of Cancer-Specific Post-transcriptional Regulatory Network

The post-transcriptional regulatory network was constructed by collating miRNA-TG interactions and protein-protein interactions. The experimentally verified miRNA-TG interactions were retrieved from mirTarbase (47) and Tarbase (48) databases and merged, which consisted of 8,407 miRNA-TG interactions forming among 743 miRNA and 2,891 TG.

Subsequently, the same procedure described for the construction of the TF-TG-PPIN network was followed for the construction of the miRNA-TG-PPIN network. The miRNA-TG-PPIN using cervical cancer-specific miRNA transcriptomics data consisted of 1,017 edges connecting 1,718 nodes. Likewise, breast and ovarian TF-TG-PPINs consisting of 2,668, 2,657 nodes and 6,197, 5,702 interactions, respectively, were also constructed.

Characterization of Cancer-Specific Networks

The cancer-specific networks, PPIN, TF-TG-PPIN, and miRNA-TG-PPIN, described above were compared with the respective random networks of the same number of interactions. Ten random networks were generated by the NetworkX program (49) against each cervical cancer-specific network described above, and the degree distribution of each network was compared with the respective random networks. The degree distribution was calculated using the following formula:

$$P(k) = n_k/N$$

Where the degree distribution of network $P(k)$ signifies the fraction of the node with degree k . For the network with a node size of N , n_k nodes will have the degree k .

Identification of IINs

Topologically IINs (genes/proteins/miRNA) of the constructed cancer-specific networks described above were identified by utilizing procedures based on graph theory methods described earlier in Bhattacharyya and Chakrabarti (38). Identification of important interacting genes/proteins/miRNAs in the network is based on some independent network properties, such as hub (highly connected nodes in the network), CNs of the network, GNPN, and LNPN. The nodes (genes/proteins/miRNAs) that were identified as topologically important in at least two categories (Hub, CN, LNPN, and GNPN) were considered as IINs.

Over-representation Analysis

Kyoto encyclopedia of genes and genomes (KEGG) pathway-based ORA was performed with deregulated genes extracted from the mRNA expression datasets used and IINs (except miRNAs) identified in each of the regulatory networks described above using “protein-coding gene set” as the reference gene set in WebGestalt (50) web tool. The top 20 pathway categories were ranked based on significant false detection rate (FDR) calculated using Benjamini and Hochberg procedure (43) and enrichment ratio.

Additionally, Gene Ontology (GO)-based molecular functions and online mendelian inheritance in man (OMIM)-based disease pathway over-representation analyses were also performed for the deregulated genes/proteins in cervical cancer.

Construction of S-M Enzyme Cross-Connecting Paths and Network

A signaling-metabolic inter-connection network was constructed using 23 signaling pathway (cancer-specific) genes/proteins and all the metabolic pathways (85 pathways) genes/proteins. Signaling and metabolic gene/protein datasets were created by extracting all the genes/proteins from the KEGG (51, 52) database. All possible unique connections (maximum three proteins involved in between) to a metabolic pathway protein (M) were established through PPIs (up to the second level), considering a signaling pathway protein (S) as a starting point in the HPPIN. Four different types of linking paths were established where signaling proteins were connected to metabolic pathway proteins either directly (S-M) or via one (S-P-M), two (S-P-P-M), or three (S-P-P-P-M) PPIs, respectively. NetworkX program (49) was used to construct all possible signaling to metabolic interconnecting paths. These paths/connections were converted into a network to construct a signaling-metabolic interaction network (SMIN).

Construction of TF to Metabolic Enzyme Cross-Connecting Paths and Network

Connections between TFs and metabolic pathway genes/proteins were established through TF-TG interactions and PPIs of TF-TG (up to the second level) considering TFs as a source. In this study, five different types of the path were established where TFs were connected to metabolic pathway proteins either directly (TF-TG/M), or TFs were connected to their TG, and their TG were connected to metabolic proteins directly (TF-TG-M), or

through one (TF-TG-P-M), two (TF-TG-P-P-M), and three (TF-TG-P-P-P-M) PPIs of TG, respectively. The resulting paths/links were converted into a network to construct a TF-metabolic interaction network (TFMIN).

Construction of miRNA to Metabolic Enzyme Cross-Connecting Paths and Network

MicroRNAs to metabolic pathway proteins interconnecting all possible paths were established using the NetworkX program (49). The miRs to metabolic pathway proteins (M) interconnecting paths were established using miRNA-TG interactions and PPIs (up to the second level) of miRNA TG considering miRNA as the source. The resulting paths were of five types viz; miRNA-TG/M, miR-TG-M, miR-TG/P-P-M, miR-TG/P-P-P-M, and miR-TG/P-P-P-P-M paths. The resulting paths were converted into a network to form a miR-metabolic interconnecting network (miRMIN).

Contextualization of Regulatory Molecules (Signaling Pathway Proteins, TF, and miRNA) to Metabolic Enzyme Cross-Connecting Paths and Network

The deregulated (upregulated and downregulated) and neutrally EG and miRNAs identified from the cervical, breast, and ovarian cancer patients specific transcriptomic datasets were mapped onto all possible paths/connections mentioned above to filter cancer-specific regulatory molecules (signaling pathway proteins, TF, and miRNA) to metabolic enzymes cross-connecting paths. For the signaling to metabolic interconnecting paths/connections/links, the paths having deregulated (upregulated and downregulated) genes/proteins at the terminals and deregulated or EG/proteins in middle were filtered out and converted into a network to form cervical, breast, and ovarian cancer-specific signaling-metabolic interconnecting subnetwork (CC-SMIN, BC-SMIN, and OC-SMIN, respectively).

Similarly, to construct the cancer-specific TF and miRNA to metabolic interconnecting sub-networks (CC/BC/OC-TFMIN and CC/BC/OC-miRMIN, respectively), the paths having deregulated genes/miRNA at the terminal and their target and deregulated or EG/proteins in the middle were considered. The respective resulting paths were converted into a network to form cancer-specific (CC/BC/OC-TFMIN and CC/BC/OC-miRMIN) sub-networks.

Calculation of Edge Weight

The edge weights in each sub-networks (CC/BC/OC-SMIN, CC/BC/OC-TFMIN, and CC/BC/OC-miRMIN) were defined in terms of local entropy using an in-house program (37). The probability of interactions of a gene/protein with its interactors in the sub-network was determined by using the principle of mass action to define the local entropy of a gene/protein. The calculation of the interaction probabilities is based on the assumption that two proteins known to interact will have a higher probability of interaction when they are highly expressed. The normalized expression values of sub-network genes in the

samples of cancer patients used in this study were utilized to calculate the interaction probabilities.

Calculation of Node Weight and Effect-On-Node

To incorporate the importance and impact of the interactors of a particular node in the sub-networks, the node-weight (W_i) of every node i was defined based on its biological properties [signaling cross-talk (SC) protein and rate-limiting enzyme (RLE)], differentially expressed gene (DEXP), and network topological properties [hub, CP, local network perturbing protein (LNPP), and global network perturbing proteins (GNPP)] using the following formula.

$$W_i = \begin{cases} 1; & \text{if } (DEXP = \frac{1}{7}, RLE = \frac{1}{7}, SC = \frac{1}{7}, HUB = \frac{1}{7}, \\ & CP = \frac{1}{7}, GNPP = \frac{1}{7}, LNPP = \frac{1}{7}) \\ 0; & \text{else} \end{cases}$$

Effect of interactors on a node in the sub-network was defined as effect-on-node (*effs*) depending on the node weight of its interactors up to the second level.

$$effs = \sum_j^k \left(\sum_i^n \frac{w_i}{n_i} + \frac{w_j}{n_j} \right)$$

Where k is the degree of node s , n_i is the degree of node i , n_j is the degree of node j , and w_i and w_j are the weights of nodes i and j .

Identification of Significant Regulatory Molecules (S/TF/miR) to a Metabolic Enzyme (M) Interconnecting Pairs and Path

To understand the information flow starting from a regulatory molecule (S/TF/miRNA) to metabolic enzyme (M), cancer-specific cross-connecting paths mentioned above were scored by implementing an HMM-based mathematical model established in our laboratory earlier (37). In this study, two separate models were used to identify the significant S/TF/miR–M pairs (the source; signaling pathway protein/TFs/miR; and destination: metabolic pathway protein) and S/TF/miR–M interconnecting paths. Model 1 was applied to identify the S/TF/miR–M pairs. Model 2 was applied to identify the S/TF/miR–M interconnecting paths between the S/TF/miR–M pairs selected after Model 1. Edge weight and node weight of genes/proteins/miRNAs involved in the S/TF/miR–M path were used to calculate the path scores. The path score of each S/TF/miR to M linking path calculated by Model 1 and Model 2 was converted into a statistical z-score to identify paths deviating from the mean. A z-score (Z) ≥ 1 filter was applied to select the significant S/TF/miR–M pairs. Paths having path score $\geq 80\%$ of the highest path score for every S/TF/miR–M pair were considered as significant S/TF/miR–M interconnecting paths from Model 2.

All the networks were visualized and represented by using Cytoscape (53). The signaling, TE, and miR to metabolic pathway connections were represented as the Circos plot (54).

In-silico Perturbation Analysis

In-silico perturbation analysis was performed for each signaling pathway protein, TF, and miRNA-based gene regulatory module to identify the paths that change significantly upon removal of a node (protein/TF/miRNA). To identify the key nodes in the final paths/networks of every module, each of the nodes present in the paths/network having z-score ≥ 1 was removed individually from the HPPIN, and the path score was recalculated for the resulting paths/network by using the HMM Models 1 and 2. The perturbation score was calculated by using the average path score before and after perturbation as below.

$$Perturbation_{score} = Path_{score'} - Path_{score}$$

Where, $Path_{score}$ and $Path_{score'}$ are average path scores before and after perturbation, respectively.

The difference of average path score (before vs. after perturbation) for each perturbed node was converted into a z-score (Z), and the nodes for which z-scores deviated from the mean as $-1 \geq Z \geq 1$ were selected as effective or key nodes in significant paths/network.

Metabolomics Data Collection and Integration Into Cancer-Specific Paths

The deregulated metabolites in cervical, breast, and ovarian cancer patient were extracted from literature (55–57). The cervical cancer metabolomic dataset consisted of 55 downregulated and seven upregulated metabolites. Twenty-one downregulated and 41 upregulated metabolites were found in breast cancer whereas the ovarian cancer metabolomic dataset consisted of 46 downregulated and 116 upregulated metabolites. The metabolic genes corresponding to these metabolites were obtained from the Human metabolome database (HMDB) (58). The deregulated metabolites were mapped to the paths obtained after model implementation.

Survival Analysis

Kaplan-Meier (KM) plotter software (59, 60) was used to perform the overall survival (OS) analysis of the constituent genes/miRNAs of the identified cross-pathway paths/links. We used a KM plotter using survival and expression data of 307 cervical cancer patients obtained from the TCGA dataset (project ID: TCGA-CESC; phs000178). To estimate the survival prognostic value of a specific gene/miRNA, the patient samples were divided into high- and low-expression cohorts according to the median expression of the given gene/miRNA, and KM plots were created. Additionally, the hazard ratio (HR) and the log-rank p -value were calculated. The survival estimate of a gene/miRNA with a p -value < 0.05 was considered to be statistically significant.

Drug/Chemotherapy Response Analysis

The receiver operator characteristic (ROC) plotter (61) was used to predict the utility of the genes as predictive biomarkers with respect to drug/chemotherapy response. ROC plotter is capable to link gene expression and response to therapy using transcriptome-level data of 3,104 breast cancer patients.

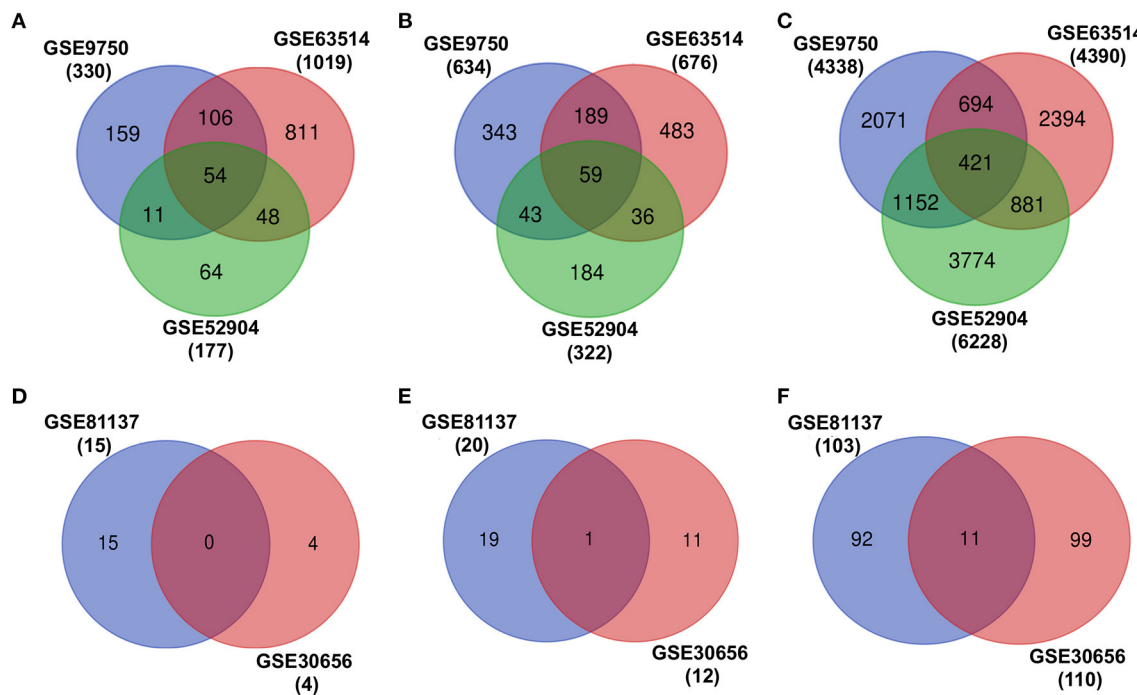


FIGURE 2 | Differential expression analysis of transcriptomic datasets associated with cervical cancer. (A–C) represent the overlap of upregulated, downregulated, and EG, respectively. (D–F) show the overlap of upregulated, downregulated, and expressed miRNAs, respectively.

RESULTS

mRNA and miRNA Expressions in Cervical Cancer Patients

The individual differential expression analysis of three different cervical cancer mRNA expression datasets (GSE9752, GSE63514, and GSE52904) leads to the identification of several upregulated, downregulated, and neutrally EG in cervical cancer (Figure 2, Table 1). However, a comparison of the gene expression patterns across datasets showed that only 54 upregulated, 59 downregulated, and 421 neutrally EG were found to be overlapped (Figures 2A–C). The differential expression analysis of miRNA resulted in four upregulated, 12 downregulated, and 110 neutrally expressed miRNAs in GSE30656 and 15 upregulated, 20 downregulated, and 103 neutrally expressed miRNAs in the GSE81137 dataset (Figures 2D–F).

Over-representation analysis-based enrichment for cellular pathways, molecular functions, and biological processes was performed using a merged list of deregulated (upregulated and downregulated) genes of all the three mRNA datasets. Supplementary Figures 1A,B show the top 20 most enriched pathway filters based on FDR for proteins encoded by upregulated and downregulated genes, respectively. Most highly enriched pathways were found to be DNA replication ($p = 4.42E^{-13}$) and arachidonic acid metabolism ($p = 2.76E^{-08}$) for the proteins encoded by upregulated and downregulated genes, respectively. GO-based molecular function and biological process ORA was also performed using deregulated genes in cervical cancer. The most significantly enriched molecular

functions were found to be collagen binding ($p = 4.36E^{-07}$) and oxidoreductase activity ($p = 4.35E^{-04}$) for proteins encoded by upregulated and downregulated genes, respectively. However, microtubule cytoskeleton organization involved in mitosis ($p = 2.13E^{-14}$) and peptide cross-linking ($p = 8.81E^{-11}$) were highly enriched biological processes, for upregulated gene-encoded proteins and downregulated gene-encoded proteins, respectively (Supplementary Figure 1).

Construction and Characterization of Cervical Cancer-Specific Networks

The cervical CC-PPIN, TF-TG-PPIN, and miR-TG-PPIN were constructed by mapping differentially expressed genes, miRNA, and neutrally EG from each cervical cancer dataset (see Methods; Supplementary Figures 2A, 3A, 4A, respectively). The resulting CC-PPIN, TF-TG-PPIN, and miR-TG-PPIN were validated by comparing them with corresponding 10 random networks of the same size. The degree distributions of CC-PPIN, TF-TG-PPIN, and miR-TG-PPIN networks followed the Power law and were considered to have scale-free organization. However, the degree distribution of 10 corresponding random networks showed binomial distribution (Supplementary Figures 2B, 3B, 4B).

IINs in Cervical Cancer-Specific Networks

Network topologically important nodes, such as hubs (highly connected nodes in the network), CNs, LNPns, and GNPns in a scale-free network could play important roles in maintaining the network integrity and function. Various topologically important interacting genes/proteins/miRNAs

(hubs, CN, GNPN, and LNPN) in cervical cancer-specific networks (CC-PPIN, TF-TG-PPIN, and miR-TG-PPIN) were identified by implementing a graph theory-based method described earlier by Bhattacharyya and Chakrabarti (38). A total of 165, 167, 96, and 67 nodes/proteins in CC-PPIN (**Supplementary Figure 2C**, **Supplementary Table 1**), 62, 36, 60, and 80 nodes/genes/proteins in TF-TG-PPIN (**Supplementary Figure 3C**, **Supplementary Table 2**), and 45, 45, 30, and 24 nodes/genes/proteins/miRNAs in miR-TG-PPIN (**Supplementary Figure 4C**, **Supplementary Table 3**) were identified as hubs, CNs, GNPNs, and LNPNS, respectively. The nodes/genes/proteins/miRNAs common in at least any two of the four categories (hubs, CN, GNPN, and LNPN) were considered as IINs in cervical cancer networks. When the IINs from each network were compared, 30 IINs were found to be common in CC-PPIN and TF-TG-PPIN, 38 IINs were common in CC-PPIN, and miR-TG-PPIN, and 17 IINs were shared by TF-TG-PPIN and miR-TG-PPIN. However, 17 IINs were found to be common in all three cervical cancer networks (**Supplementary Figure 5A**, **Supplementary Table 4**).

Over-representation analysis-based pathway enrichment was performed using IINs in the above described cervical cancer-specific regulatory networks. Top three enriched pathways were found to be DNA replication ($p = 1.30E^{-10}$), basal TFs ($p = 3.07E^{-10}$), and cell cycle ($p = 0.00$) for IINs in CC-PPIN (**Supplementary Figure 2D**). DNA replication ($p = 1.21E^{-10}$), mismatch repair ($p = 1.29E^{-04}$), and cell cycle ($p = 1.44E^{-15}$) were the top three enriched pathways for IINs in TF-TG-PPIN (**Supplementary Figure 3D**), respectively. However, for IINs (except miRNA) of miR-TG-PPIN, top three enriched pathways were found to be DNA replication ($p = 8.88E^{-16}$), cell cycle ($p = 0.00$), and mismatch repair ($p = 5.23E^{-06}$; **Supplementary Figure 4D**). When the top 20 enriched pathways for IINs in each network were compared, 8 enriched pathways were found to be common (**Supplementary Figure 5B**). The common pathways were cell cycle, DNA replication, nucleotide excision repair, mismatch repair, prostate cancer, herpes simplex infection, oocyte meiosis, and viral carcinogenesis pathways.

S-M Enzyme Cross-Connecting Paths and Network in Cervical Cancer

Experimentally supported HPPIs (experimental score ≥ 0.7) were utilized to establish all possible S-M cross-connecting paths, where signaling pathway proteins were connected to metabolic enzymes either directly (S-M paths) or via one (S-P-M paths), two (S-P-P-M paths), and three (S-P-P-P-M paths) PPIs in between them. The resulting S-M cross-connecting paths consisted of 210 direct (S-M) connections, 2,669 via one PPI (S-P-M), 40,266 via two PPIs (S-P-P-M), and 7,35,395 via three PPIs (S-P-P-P-M) interconnections. These interconnections/paths were formed between 210, 1,697, 7,965, and 28,920 S-M pathway protein pairs, respectively. These S-M paths were converted into a network to form a SMIN, which consisted of 11,442 interactions formed among 2,603 genes/proteins.

To understand the flow of information from a signaling protein to metabolic enzymes probably leading to metabolic

adaptations in the case of cervical cancer cells, we mapped differentially expressed (up and downregulated) and neutrally expressed EG onto the abovementioned S-M interconnecting paths. The S-M paths having deregulated genes at the terminal and deregulated or EG in middle were extracted and considered for further analysis. The resulting paths/interconnections consisted of four (S-M), 47 (S-P-M), 639 (S-P-P-M), and 10,311 (S-P-P-P-M) paths formed between 4, 39, 180, and 631 S-M pairs, respectively. The filtered paths were converted into a network to construct the cervical cancer-specific SMIN (CC-SMIN) network. The CC-SMIN consisted of 1,425 interactions forming among 439 genes/proteins.

To identify the potential disease-specific paths/pairs, each node and each edge of the CC-PPIN network were weighted based on their biological properties, differential expression status, and network topological properties (hubs, CN, GNPN, and LNPN) of CC-PPIN. Local signaling entropy (S_i) was integrated to understand the system-level network property. The significance of each node (gene/protein) in the cancer-specific network was estimated in the form of effect-on-node ($effs$) based on SC, RLE, dEXP (differentially expressed genes), hub, CN, GNPN, and LNPN, respectively. To identify the probable and significant paths of information flow from signaling pathway to metabolic pathway in cervical cancer cells, local signaling entropy (S_i) and effect-on-node ($effs$) properties were incorporated as node weights. The edge weight of every two interacting nodes of CC-PPIN was defined as the probability of interaction using their normalized expression value in cervical cancer patient samples. Node weight and edge weight were integrated into HMM-based mathematical models (Models 1 and 2) to identify S-M linking pairs and paths. Model 1 was applied to identify the S-M pairs (the source signaling pathway protein and destination metabolic pathway protein). Model 2 was applied to identify the S-M interconnecting paths between the S-M pairs selected after Model 1. The path score of each S-M linking path calculated by Models 1 and 2 was converted into a statistical z-score to identify paths deviating from the mean. A z-score ≥ 1 filter was applied to select the significant S-M pairs. Using these filtering criteria, we identified 81 S-M pairs and 94 S-PPI-M paths in cervical cancer. The selected paths were converted into a network to form a significant CC-PPIN network which consisted of 152 interactions forming among 135 genes/proteins (**Table 2**, **Figure 3A**).

Mapping pathway information to the terminal nodes (source signaling protein and destination metabolic enzyme) showed that the Ras signaling pathway had maximum connections (50) to all the six metabolic pathways followed by cell cycle (44), Map-kinase (44), epidermal growth factor receptor (EGFR) signaling pathway (42), and p53 signaling pathway (34), respectively. However, among metabolic pathways, nucleotide metabolism had maximum connections (87) with signaling pathways, followed by amino acid (63), energy (58), xenobiotics (44), carbohydrate (42), and lipid metabolism (11), respectively. 1:1 interconnections between signaling and metabolic pathways showed that the cell cycle had maximum connections with nucleotide metabolism (18 connections; **Figure 3B**, **Supplementary Table 5**).

TABLE 2 | Signaling to metabolic pathways interconnecting paths and pairs.

Connection types	Pairs and paths							
	Signaling-Metabolic interaction network (SMIN)		Cervical cancer (CC)		Breast cancer (BC)		Ovarian cancer (OC)	
			Significant CC-SMIN		Significant BC-SMIN		Significant OC-SMIN	
	Pairs	Paths	Model 1 pair selection	Model 2 paths selection	Model 1 pair selection	Model 2 paths selection	Model 1 pair selection	Model 2 paths selection
S-M	210	210	$Z \geq 1$ pairs	$Z \geq 1$ or score $\geq 80\%$ paths	$Z \geq 1$ pairs	$Z \geq 1$ or score $\geq 80\%$ paths	$Z \geq 1$ pairs	$Z \geq 1$ or score $\geq 80\%$ paths
S-P-M	1,697	2,669	1	1	1	1	2	2
S-P-P-M	7,964	40,266	15	14	9	12	7	9
S-P-P-P-M	28,920	735,395	79	78	57	79	39	48
Total	29,179	778,540	81	94	223	187	193	191
					232	279	218	250

S-M, signaling-metabolic; S-P-M, single protein-protein interaction; S-P-P-M, two protein-protein interactions; S-P-P-P-M, three protein-protein interactions.

Mapping the deregulated metabolites in cervical cancer to significant S-M paths yielded 12 S-PPI-M interconnections/paths where four metabolites [L-lysine, oxoglutaric acid, tetrahydrodeoxycorticosterone (THDOC), and pyruvic acid] were regulated by eight signaling pathway proteins (BAD, CHEK1, GNB5, MAML3, MAP3K1, PAK2, PPARD, and SPP1). The metabolic enzymes connecting these four metabolites were enhancers of zeste homolog 2 (EZH2), procollagen lysine hydroxylase and glycosyltransferase LH3 (PLOD3), aldo-keto reductase family 1 member C2 (AKR1C2), and 3-mercaptopyruvate sulfurtransferase (MPST). L-lysine is the substrate of both EZH2 and PLOD3. Whereas, THDOC and pyruvic acid are the products of metabolic enzymes AKR1C2 and MPST, respectively. Hence, these paths/connections showed the correlated status of the metabolic enzymes and the corresponding metabolites (Figure 3C).

TF to Metabolic Enzyme Cross-Connecting Paths and Network in Cervical Cancer

All possible paths/links connecting TF to the metabolic enzyme (M) were established using TF-TG interaction and HPPIN (refer to Methods). The resulting paths/links consisted of 930 TF-TG/M paths, 4,276 TF-TG/P-M, 114,844 TF-TG/P-P-M, 9,384,069 TF-TG/P-P-P-M, and 188,563,171 TF-TG/P-P-P-P-M paths forming between 930, 2,299, 9,121, 33,676 and 90,477 TF-M pairs, respectively. These paths were converted into a network to form a TFMIN.

All possible TF-TG-PPI-M paths were filtered by mapping deregulated and neutrally EG to establish the context-specific (cervical cancer-specific) paths and network. The resulting paths consisted of 89 TF-TG/M, 20 TF-TG/P-M, 110 TF-TG/P-P-M, 1,262 TF-TG/P-P-P-M, and 21,016 TF-TG/P-P-P-P-M paths formed between 89, 17, 53, 162 and 508 TF-M pairs, respectively. The filtered paths were converted into a network to construct the cervical cancer-specific TFMIN (CC-TFMIN), which consisted of 815 nodes and 2,364 edges formed among them.

Each node and each edge in the CC-TFMIN were weighted to identify the potential significant paths in the cervical cancer-specific network. Node weights and edge weights were incorporated into HMM-based mathematical models (Models 1 and 2) to identify TF-M pairs and TF-PPI-M paths forming between them (refer to Methods). Model 1 resulted in the identification of 172 significant ($z \geq 1$) TF-M pairs and Model 2 resulted in 236 significant TF-PPI-M paths connecting the TF-M pairs obtained after Model 1. The significant TF-PPI-M paths in cervical cancer were converted into a network to form significant CC-TFMIN that consisted of 179 interactions among 141 genes/proteins (Figure 4A, Table 3).

The metabolic pathway information was mapped onto the terminal metabolic enzymes of the significant TF-PPI-M paths to identify metabolic pathways that are highly connected to specific TFs. Nucleotide metabolism yielded maximum connections followed by energy metabolism, amino acid metabolism, carbohydrate metabolism, lipid metabolism, and xenobiotics biodegradation and metabolism (Figure 4B, Supplementary Table 6).

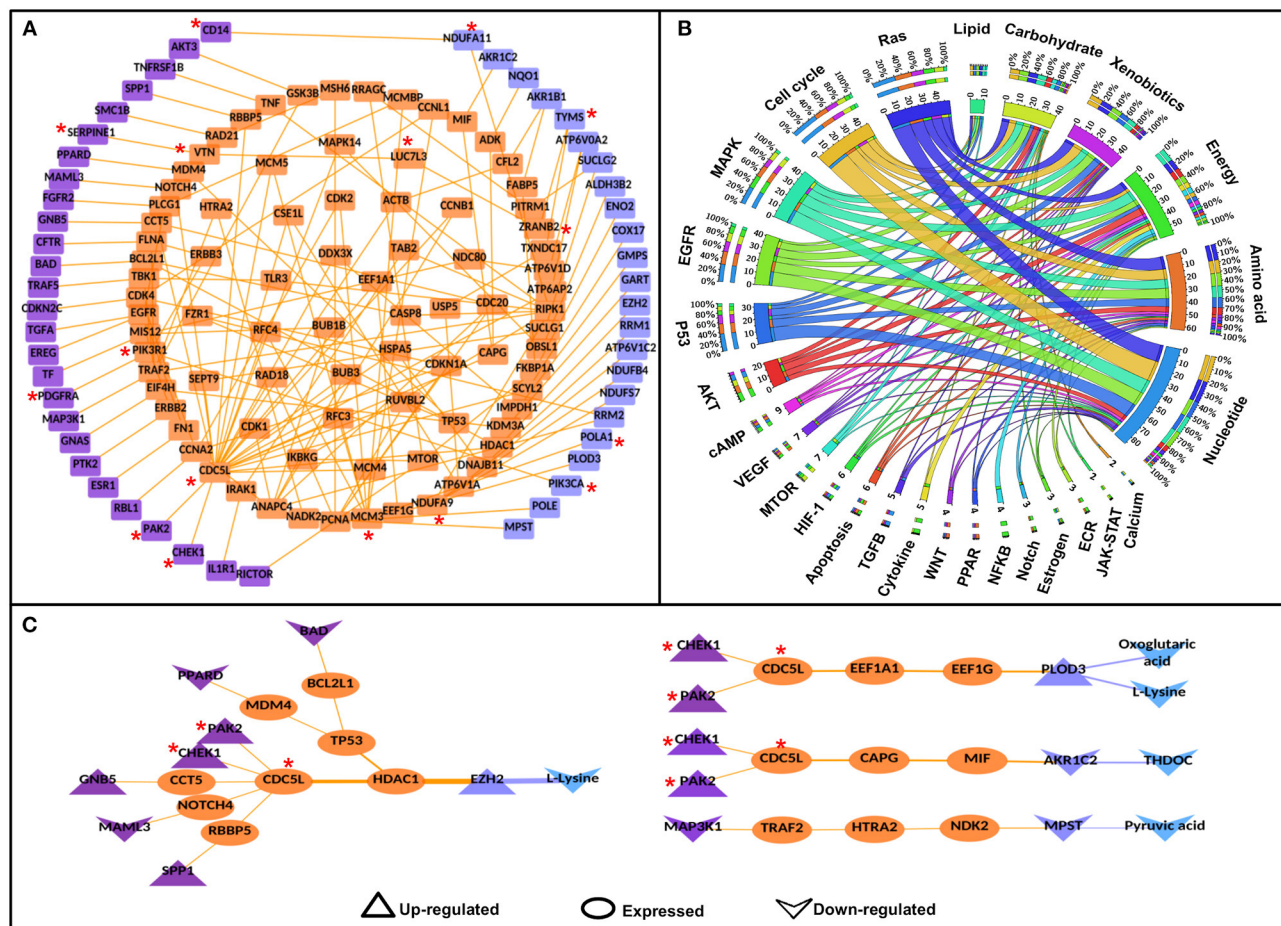


FIGURE 3 | Significant signaling-metabolic pathway interconnecting paths/network in cervical cancer. (A) represents a significant signaling metabolic interaction network, (B) shows the signaling metabolic pathways interconnectivity, and (C) shows the significant signaling to metabolic paths regulating metabolites in cervical cancer. Terminal signaling pathway proteins and metabolic enzymes are colored in purple and blue. Protein-protein interactors are colored in orange. Protein-protein interactions are represented by orange edges. Nodes with an asterisk (*) are key or effector nodes in the significant paths/network.

Mapping of deregulated metabolites onto the terminal metabolic enzymes resulted in 28 paths where four metabolites were deregulated in cervical cancer. The deregulated metabolites were L-lysine, oxoglutaric acid, pyruvic acid, and THDOC. L-lysine was found to be linked with AR, ESR1, ZEB1, NCOA2, HSF2, RBL1, ID4, E2F1, EGR1, MED1, TWIST1, KAT2B, ING1, and PGR TFs via 19 different paths. Oxoglutaric acid was linked with PGR, EGR1, and ESR1 via three paths. MED1 was found to be regulating (probably) pyruvic acid whereas THDOC was found to be linked with AR, TWIST1, and ING1 in four different paths (Figure 4C).

miR to Metabolic Enzyme Cross-Connecting Paths and Network in Cervical Cancer

Similar to SMIN and TFMIN, all possible paths/links were established considering miRNA as a source node and metabolic enzymes as a destination by collating miRNA-TG interactions

and HPPIN. The resulted paths/links consisted of 577 direct (miR-TG/M) paths, 1,145 via their TG (miR-TG/P-M), 26,330 via one PPI (miR-TG/P-P-M), 826,207 via two PPI (miR-TG/P-P-P-M), and 33,934,931 via three PPI (miR-TG/P-P-P-P-M) paths formed by 577, 1,128, 9,904, 44,271, and 111,730 miR-M pairs, respectively. The deregulated miRNA, genes, and neutrally EG were mapped onto all possible miR-PPI-M paths to filter the cervical cancer-specific miRNA to metabolic enzymes interconnections. The filtered paths/links consisted of 13 miR-TG/M, 1 miR-TG/P-M, 7 miR-TG/P-P-M, 95 miR-TG/P-P-P-M, and 1,851 miR-TG/P-P-P-P-M paths formed between 13, 1, 7, 38 and 149 miR-M pairs, respectively. The resulted paths were converted into a network to form a cervical cancer-specific miR-metabolic enzyme interaction network (CC-miRMIN) which consisted of 309 nodes and 952 interactions among them (Table 4).

The nodes and edges in the CC-miRMIN were weighted based on their biological properties, differential expression status, and network topological properties in cervical cancer-specific

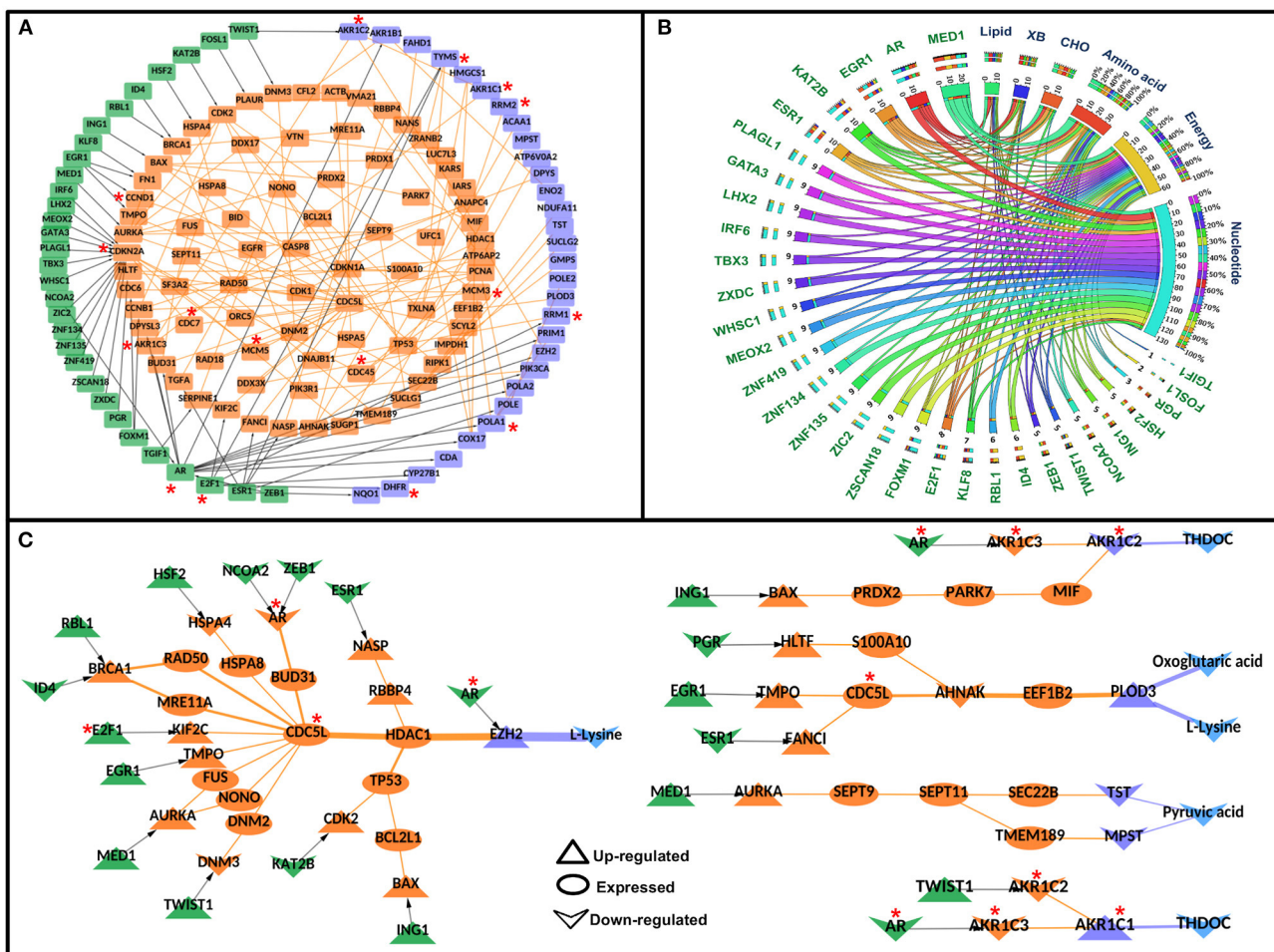


FIGURE 4 | Significant Transcription factor-metabolic pathway interconnecting paths/network in cervical cancer. **(A)** represents a significant transcription factor metabolic interaction network, **(B)** shows the transcription factor to metabolic pathways interconnectivity, and **(C)** shows the significant transcription factor to metabolic paths regulating metabolites in cervical cancer. Terminal transcription factors and metabolic enzymes are colored green and blue. Protein-protein interactors are colored in orange. Gene regulatory edges are represented as black arrows and protein-protein interactions are represented by orange edges. Nodes with an asterisk (*) are key or effector nodes in the significant paths/network.

miR-TG-PPIN. After the incorporation of HMM Model 1, 22 significant ($Z \geq 1$) miR-M pairs were identified. HMM, Model 2, resulted in the identification of a total of 27 miR-PPI-M paths, where 22 significant miR-M pairs obtained after Model 1 were connected via their TG and PPIs. The resulting miR-PPI-M paths were converted into a network to form significant CC-miRMIN that consisted of 59 nodes and 67 interactions among them (Figure 5A).

Mapping metabolic pathway information to the terminal metabolic enzymes in significant miR-PPI-M paths showed that amino acid metabolism was highly regulated by miRNAs, followed by nucleotide metabolism, xenobiotics biodegradation and metabolism, energy metabolism, carbohydrate metabolism, and lipid metabolism (Figure 5B).

After mapping deregulated metabolites to the terminal metabolic enzymes, 13 miR-PPI-M paths were found to regulate the metabolites. L-lysine was found to be linked/regulated by

miR-138-5p, miR-223-3p, miR-203a-3p, miR-593-5p, miR-15b-5p, miR-16-5p, and miR-147a via 11 miR-PPI-M paths. THDOC was found to be linked with miR-593-5p and miR-193b-3p via two different miR-PPI-M paths (Figure 5C).

***In-silico* Perturbation of Nodes in the Final Weighted Paths/Network**

In-silico perturbation analysis was performed to identify the paths that change significantly upon removal of a node (protein/TF/miRNA). To identify the key nodes in the final paths/networks of every module, each of the nodes present in the paths/network having $Z \geq 1$ was removed individually from the HPPIN, and the path score was recalculated for the resulting paths/network by using HMM Models 1 and 2. Accordingly, new significant pairs and paths were identified based on Models 1 and 2, respectively. The difference of average path score (before vs. after perturbation) for each perturbed node was converted into a

TABLE 3 | Transcription factor to metabolic pathways interconnecting paths and pairs.

Connection types	Pairs and paths							
	TF-Metabolic interaction network (TFMIN)		Cervical cancer (CC)		Breast cancer (BC)		Ovarian cancer (OC)	
			Significant CC-TFMIN		Significant BC-TFMIN		Significant OC-TFMIN	
	Pairs	Paths	Model 1 pair selection $Z \geq 1$ Pairs	Model 2 paths selection $Z \geq 1$ or score $\geq 80\%$ paths	Model 1 pair selection $Z \geq 1$ pairs	Model 2 paths selection $Z \geq 1$ or score $\geq 80\%$ paths	Model 1 pair selection $Z \geq 1$ pairs	Model 2 paths selection $Z \geq 1$ or score $\geq 80\%$ paths
TF-TG/M	930	930	6	14	2	2	2	3
TF-TG/P-M	2,299	4,276	5	6	3	4	5	12
TF-TG/P-P-M	9,121	114,844	10	9	11	18	4	17
TF-TG/P-P-P-M	33,676	9,384,069	53	66	46	87	116	155
TF-TG/P-P-P-P-M	90,477	188,563,171	159	141	251	200	562	482
Total	91,790	193,067,290	172	236	261	311	577	669

TF-TG/M, TFs connected to metabolic pathway proteins; TF-TG-M, TFs connected to TG and TG connected to metabolic proteins directly; TF-TG-P-M, TFs connected to TG and TG connected to metabolic proteins directly through one PPI of TG; TF-TG-P-P-M, TFs connected to TG and TG connected to metabolic proteins directly two PPIs of TG; TF-TG-P-P-P-M, TFs connected to TG and TG connected to metabolic proteins directly three PPIs of TG.

TABLE 4 | microRNA to metabolic pathways interconnecting paths and pairs.

Connection types	Pairs and paths							
	miRNA-Metabolic interaction network (miRMIN)		Cervical cancer (CC)		Breast cancer (BC)		Ovarian cancer (OC)	
			Significant CC-miRMIN		Significant BC-miRMIN		Significant OC-miRMIN	
	Pairs	Paths	Model 1 pair selection $Z \geq 1$ pairs	Model 2 paths selection $Z \geq 1$ or score $\geq 80\%$ paths	Model 1 pair selection $Z \geq 1$ pairs	Model 2 paths selection $Z \geq 1$ or score $\geq 80\%$ paths	Model 1 pair selection $Z \geq 1$ pairs	Model 2 paths selection $Z \geq 1$ or score $\geq 80\%$ paths
miR-TG/M	577	577	1	2	9	9	0	0
miR-TG/P-M	1,128	1,145	1	1	4	6	2	2
miR-TG/P-P-M	9,904	26,330	2	2	14	34	5	8
miR-TG/P-P-P-M	44,271	826,207	6	7	42	73	21	38
miR-TG/P-P-P-P-M	111,730	33,934,931	20	15	299	291	139	118
Total	112,745	34,789,190	22	27	325	413	150	166

z-score and the nodes for which z-scores deviated from the mean as $-1 \geq Z \geq 1$ were selected as effective or key nodes in significant paths/networks. Sixteen nodes/proteins (CDC5L, PAK2, CHECK1, NDUFA9, MCM, POLA1, PIK3CA, PIK3R1, PDGFRA, LUC7L3, SERPINE1, VTN, ZRANB2, TYMS, CD14, and NDUFA11) in the signaling module (**Figure 2A**), 16 nodes/proteins (CDKN2A, POLA1, CDC45, CCND1, MCM5, CDC7, RRM2, MCM3, DHFR, AKR1C3, AKR1C2, AKR1C1, RRM1, TYMS, AR, and E2F1) in TF-based gene regulatory module (**Figure 4A**), and nine nodes/miRNA/proteins (CDK2, MCM3, mir-147a, AKR1C2, CDC5L, PLK1, mir-593-5p, TYMS, and mir-196a-5p) in miRNA-based gene regulatory module (**Figure 5A**) were identified as an effector or key nodes in the significant paths/networks.

S, TF, and miR Cross-Talks in Cervical Cancer

Comparing cervical cancer-specific significant S-PPI-M, TF-PPI-M, and miR-PPI-M paths or links discussed above resulted in the identification of 83 paths/links where six metabolic enzymes (RRM2, AKR1C2, ENO2, TYMS, EZH2, and NDUFA11) were probably regulated by signaling pathway proteins (BAD, PPARG, GNB5, TF, PAK2, RBL1, CDK2NC, TRAF5, CFTR, AKT3, MAP3K1, IL1R1, RICTOR, TNFRSF1B, CHEK1, BRCA1, MAML3, SPP1, PLK1, ATP6V1C2, and SERPINE1), TFs (TGIF1, FOSL1, E2F1, TWIST1, ING1, HSF2, ESR1, RBL1, ID4, EGR1, NCOA2, ZEB1, AR, MED1, KAT2B, FOXM1, and KLF8), and miRs (miR-593-5p, miR-15b-5p, miR-106b-5p, miR-147a, miR-494-3p, miR-138-1-3p, miR-196a-5p, miR-138-5p, miR-16-5p, and miR-223-3p) (**Figure 6**). Out of six metabolic enzymes, AKR1C2 and EZH2 were mapped to the deregulated metabolites THDOC and L-lysine, respectively (**Figures 6B,D**).

Survival Analysis of the Genes/miRNAs of Identified Paths/Links in Cervical Cancer

Potential prognostic values of the genes and miRNAs of the signaling, transcriptional and post-transcriptional cross-connecting paths/links to metabolic enzymes in cervical cancer patients were explored by evaluating the correlation and OS. A total of 53 genes and 10 miRNAs were found to be significantly associated with the OS in the log-rank test with a $p < 0.05$ (**Supplementary Table 12**). Mapping these genes and miRNAs onto the corresponding cross-connecting paths/links yielded 16, 34, 20, and 9 paths/links to have 1–25, 26–50, 51–75, and 76–100% of their component as a prognostic marker in cervical cancer patients (**Figure 7A**). Almost all the final selected paths/links (79 out of 83) possess at least one node (gene/miRNA), whose expression is significantly associated with cervical cancer patients' survival. In total, 38% (30 out of 79) of the selected paths/links have more than 50% nodes to be significant ($p < 0.05$) prognosis marker (**Figure 7A**). Further, we checked the status of these prognostic markers in different types of paths/links. Most of the two-component (2C) paths were found to have 100% of their component as a prognostic marker. Three-component (3C) paths were found to have 51–75% of their component nodes as a prognostic marker. Similarly, significantly

higher numbers of longer or higher component paths (e.g., 4, 5, and 6C, respectively) also possess more than 25% of their nodes as a prognostic marker (**Figure 7B**).

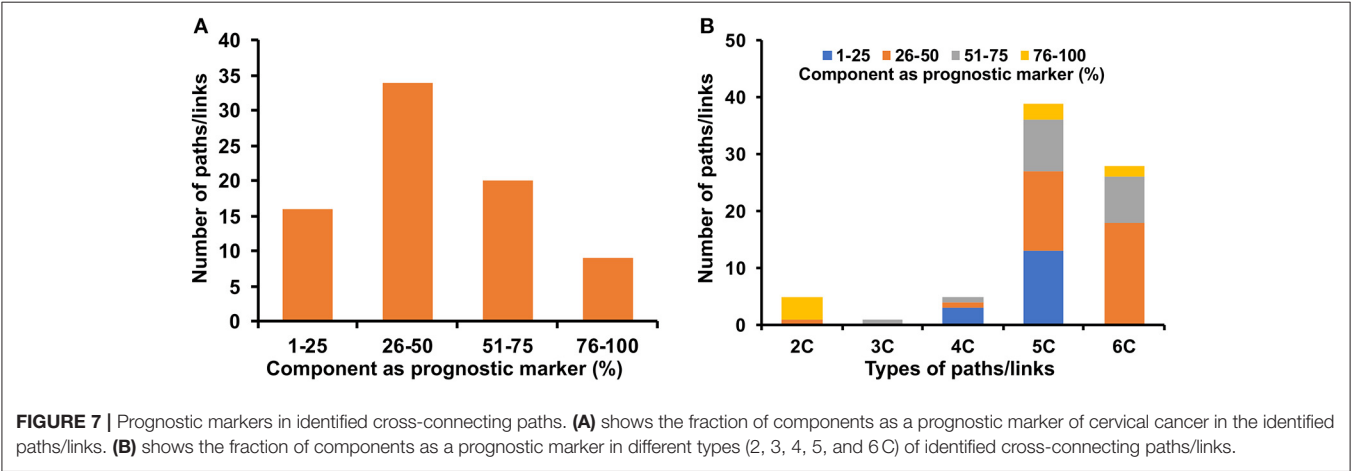
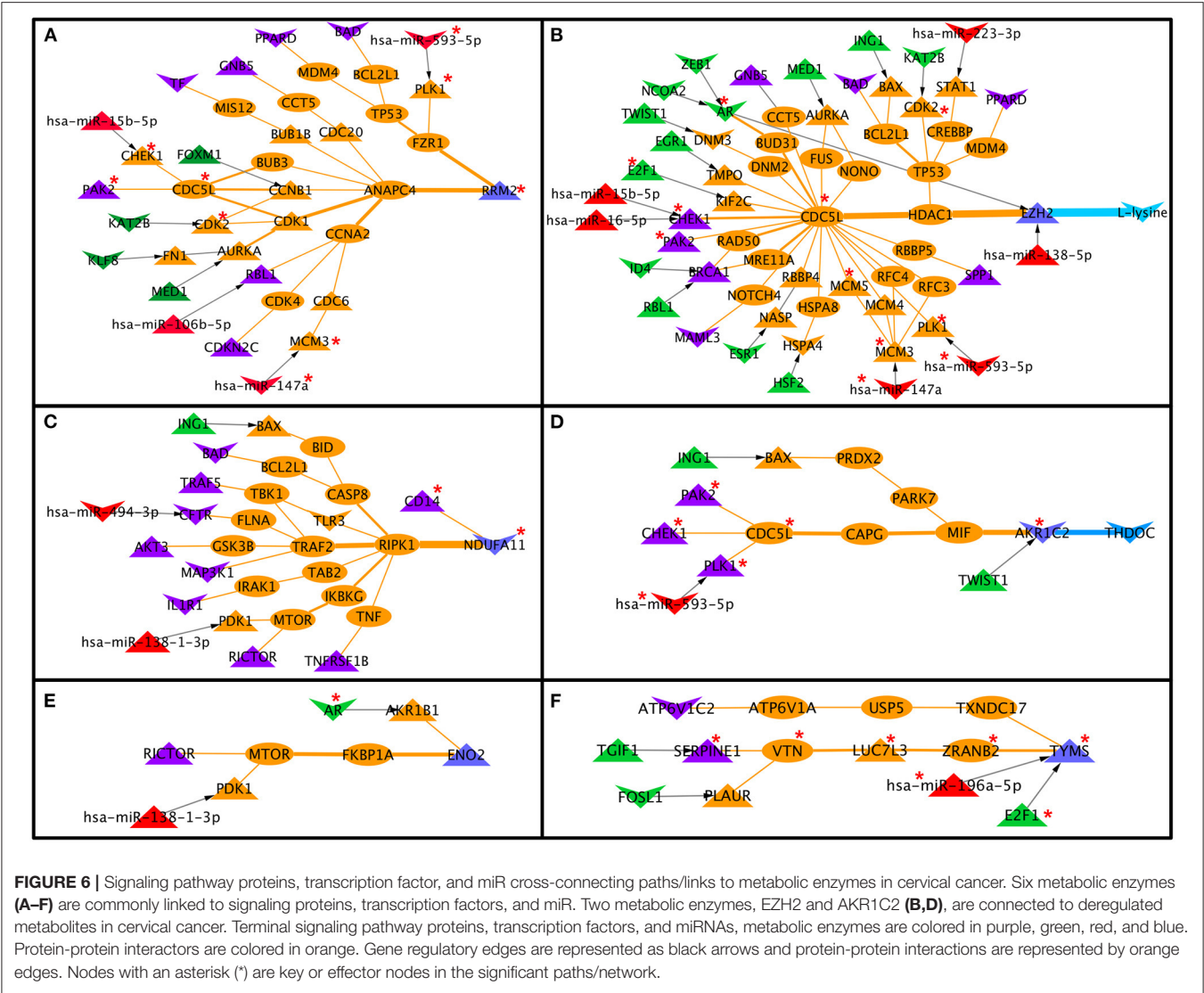
S, TF, and miR Cross-Talks in Breast and Ovarian Cancers

To investigate whether the cross-pathway links are specific to cervical cancer, we identified such paths from two other female-specific cancers, such as breast and ovarian cancers. As mentioned in the Methods, paths originating from S, TF, and miR connecting metabolic enzymes (M) were identified in breast and ovarian cancers using the cancer-specific transcriptomics data mapping followed by implementation of HMM-based mathematical models. A total of 2,79, 311 and 413 S-PPI-M, TF-PPI-M, and miR-PPI-M paths were identified in breast cancer connecting 232, 261 and 325 S-M, TF-M, and miR-M pairs, respectively. Similarly, 250, 669 and 166 S-PPI-M, TF-PPI-M, and miR-PPI-M paths were identified in ovarian cancer connecting 218, 577 and 150 S-M, TF-M, and miR-M pairs, respectively (**Tables 1–3**). Mapping of deregulated metabolites resulted in 69 paths in breast cancer and 481 paths in ovarian cancer.

The signaling (S-M), transcriptional (TF-M), and post-transcriptional (miR-M) regulatory links identified from cervical, breast, and ovarian cancer networks were compared to estimate the common and specific regulators and regulatory links (**Figure 8**). Interestingly, a very little overlap of regulatory paths and pairs was observed among the three types of cancers (**Figures 8A–F**). In total, 32% of terminal signaling proteins and 43% of terminal metabolic enzymes forming CC-specific S-M enzymes paths were found to be common with that extracted from breast and ovarian cancer networks (**Figures 8G,J**). Similarly, 46% of terminal TFs and 30% of terminal metabolic enzymes forming CC-specific TF-metabolic enzyme paths were found to be common with those extracted from breast and ovarian cancer networks (**Figures 8H,K**). Three metabolic enzymes (EZH2, ENO2, and RRM2) were found to be commonly regulated by S-M, TF-M, and miR in all three cancers (**Figure 8**).

The metabolic enzymes EZH2 and MIF connected to deregulated metabolites L-lysine and citric acid were commonly regulated in breast cancer. However, 17 metabolic enzymes (EZH2, MTHFD1, ALDH3B2, ATP6V1B1, TCIRG1, AGMAT, SETDB1, PFKL, TKT, INPPL1, MTHFD2, CPS1, MARS, PFKP, AASDHPPT, ATP6V0D2, and PLOD3) connected to nine deregulated metabolites (L-lysine, adenosine monophosphate, fructose 6-phosphate, homovanillic acid, methylimidazole acetic acid, N-acetylglutamic acid, phenylacetic acid, phosphate, and Urea) were found to be commonly regulated in ovarian cancer.

The metabolic enzyme EZH2 was connected to deregulated metabolites L-lysine in cervical, breast, and ovarian cancer (**Figure 6B**, **Supplementary Figure 6**). Twenty-five out of the 27 paths/links connected to metabolic enzyme EZH2 in the breast cancer network (**Supplementary Figure 6A**) possesses at least one gene, whose expression is significantly associated with drug/chemotherapy response. Seventeen out of the



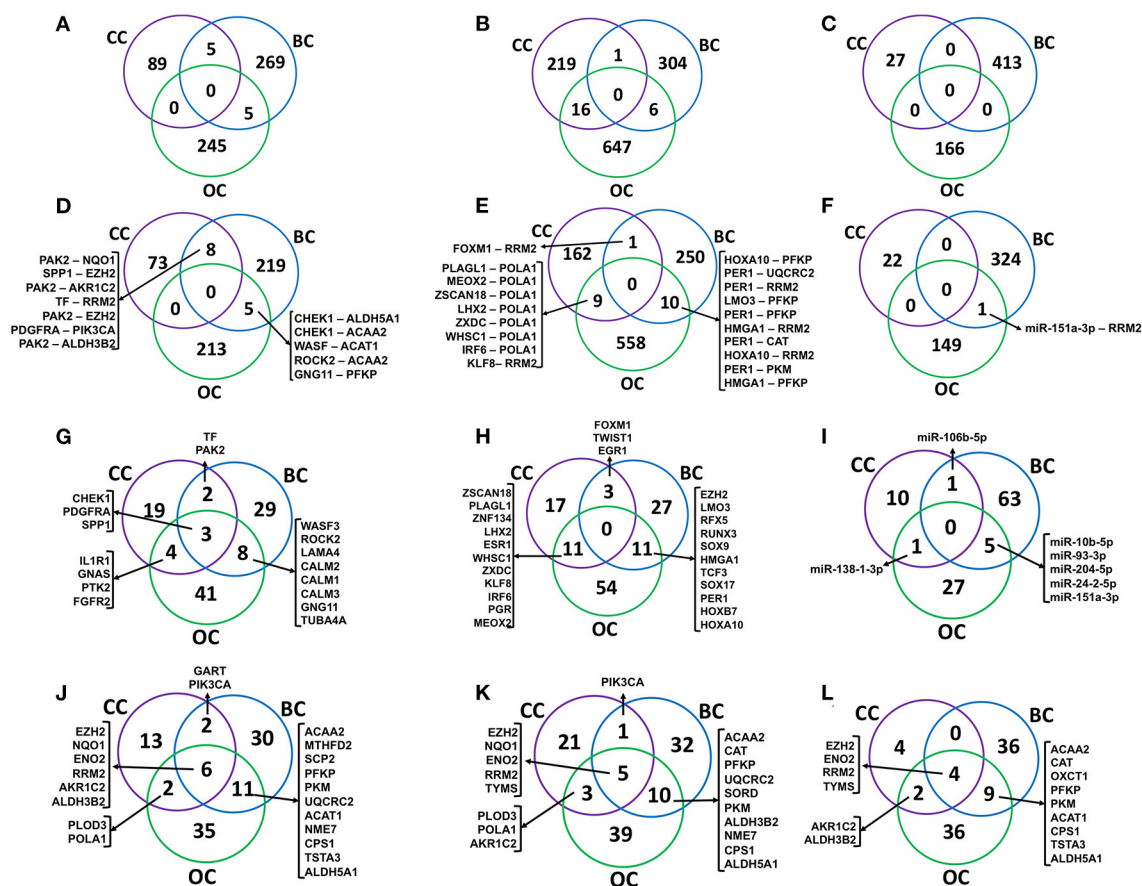


FIGURE 8 | Comparison of regulatory molecules and links in female-specific cancer. (A–C) show the overlap of significant signaling proteins (S), transcription factor (TF), and miR to metabolic enzyme connecting paths, respectively, identified from cervical cancer (CC), breast cancer (BC), and ovarian cancer (OC) specific networks. (D–F) show the overlap of S-M, TF-M, and miR-M pairs, respectively. (G–I) represent the overlap of terminal signaling protein in S-M paths, a transcription factor in TF-M paths, and miR-M paths, respectively. (J–L) represent an overlap of terminal metabolic enzymes. S-M, signaling-metabolic; TF-M, TF-metabolic; miR-M, microRNA-metabolic.

set-up to weigh and rank the interconnecting link/paths in addition to biological and network topological properties to extract out high confidence inter-pathway connections that are perhaps responsible for facilitating metabolic adaptation in cervical cancer.

In our previous study, we implemented the underlined mathematical model-based approach for the development, test, and validation of S-M interconnecting links using glioblastoma multiform (GBM) cell line-derived transcriptomics and proteomics data (37). Further, *in-vitro* perturbation of genes/proteins involved in forming a high-score interconnection between S-M pathway proteins showed a significant change in the expression of proteins involved in the metabolic pathway. This validated our model for discovering hitherto unknown connections/involvement between signaling and metabolic genes/proteins. As a natural follow-up study, here we have significantly upgraded the previous model with two entirely new types of connectivity paths linking TF and miRNA-based regulatory mechanisms to altered states of metabolic enzymes. We have used large-scale patient-derived cervical cancer data

and implemented additional network topology-based weights to signify the identified cross-pathway links. Further, we have utilized differential metabolite data to extract out paths that correlate with the altered status of the metabolic enzymes that were proposed to be regulated via signaling and regulatory factors. Comparison of the significant paths originated with S, TFs, and miRNAs yielded 88 commonly linked paths connecting six common metabolic enzymes (e.g., RRM2, AKR1C2, ENO2, TYMS, EZH2, and NDUFA11).

The ribonucleotide reductase subunit M2 (RRM2) was found to be significantly upregulated in cervical cancer tissue and is linked to promoting the progression of cervical cancer (65). RRM2 is likely to become a novel potential diagnostic and prognostic biomarker of cervical cancer.

Aldo-keto reductase subfamily 1C2, which plays a major role in regulating the activity of androgens, estrogens, progesterone, and prostaglandins metabolisms, is also implicated with cervical, endometrial, and bladder cancers (66, 67). Overexpression of AKR1C2 is found to be a mildly favorable prognostic marker (Supplementary Table 12) but

lower expression of NADH:ubiquinone oxidoreductase subunit A11 (NDUFA11) is prognostically unfavorable in cervical cancer (**Supplementary Table 12**). Enolase 2 (ENO2) and thymidylate synthetase (TYMS) are found to be upregulated in cervical cancer transcription datasets (68, 69). Overexpression of enhancer zeste homolog 2 (EZH2) has been linked with proliferation, progression, and prognosis of cervical cancer (70). However, our survival analysis using data of cervical cancer patients from TCGA suggested much lower survival with lower expression of EZH2 (**Supplementary Table 12**).

Several miRNAs have been identified whose roles have been implicated in cervical cancer progression. Most of the miRNAs except one (miR-593-5p) forming the metabolic pathway PPI links were previously reported to be dysregulated in cervical cancers (71–75). However, the diverse mechanisms by which these miRNAs could regulate cervical cancer progression were not well-known especially their roles in regulating the metabolic adaptation in cervical cancer cells. Our study provides novel avenues to study the impact of these important miRNAs in the regulation of metabolic reprogramming in cervical cancer. miR-593 plays important role in the regulation of lung, breast, and gastric cancer proliferation (76–79). Higher expression of miR-593 is found to be unfavorable for the survival of cervical cancer patients (**Supplementary Table 12**). Hence, its role in cervical cancer especially in its metabolic adaptation is worth investigating further.

Transcription factors are key regulators of cancer proliferation and metastasis. Roles of several transcription factors, such as SOX2, E2F4, E2F1, POU5F1, SMAD3, SMAD2, VDR, ERG, TP53, EWS, c-fos, fra-1, OCT4, KLF4, C-MYC, and NANOG, were established in cervical cancer (80, 81). Our study highlights the probable roles of important transcription factors in regulating the metabolic status of cervical cancer cells via modulating the metabolic enzymes. Thirty-one TFs were found to be connected to 30 metabolic enzymes via the TG and their respective PPIs. Fifteen TFs were found to be linked with six metabolic enzymes for which altered metabolite status could be associated (**Figure 4**).

Ras, cell cycle, MAPK, EGFR, and p53 are among the top five most connected signaling pathways to the metabolic enzymes via PPI interconnectivity (**Figure 3**). Among the 27 terminal signaling proteins that form significant connections with metabolic enzymes, 9 (CHEK1, MAP3K1, CDKN2C, PAK2, EGFR, FGFR2, PDGFRA, and PTK2) are found to be kinases. TFs and miRNAs are generally regarded as “undruggable,” hence, these regulatory kinases could be ideal candidates for targets of small molecules inhibitors/drugs to check their roles in altering the functional activities of the connected enzymes.

All cervical cancer-based cross-pathway links are provided in **Supplementary Tables 8–11**. Similarly, an online platform is also created as a separately published work (82) where cervical cancer dataset-specific S-M, TF-metabolic, miRNA-metabolic, and combined paths are made available at <http://www.hpppi.iicb.res.in/APODHIN/home.html>.

Cervical cancer-based PPI links to metabolic enzymes originated from signaling (S), TF, and miR regulatory molecules were also compared to the same identified from breast and ovarian cancer networks. Comparison of the regulators and regulatory links yielded little overlap among the three cancers (**Figure 8**) indicating the existence of cancer-specific regulatory mechanisms for probable metabolic alterations. However, some signaling proteins were found to be common regulatory molecules among the three cancers whereas terminal enzymes, such as EZH2, ENO2, RRM2, and TYMS, were found to be commonly regulated in all the cancers by three different regulatory mechanisms (**Figure 8**).

We understand that our approach is computation heavy and our findings require further *in vitro* and *in vivo* experimental validations. Similarly, for effective stratification of the patients, multiple omics data (e.g., transcriptomics, proteomics, and metabolomics) need to be generated from an individual patient. Nevertheless, we believe that our systems biology-based approach of identifying multi-factor signature links connected to the regulation of status and functionalities of metabolic enzymes paves the way for future studies which could be aimed toward identifying novel regulators of metabolic alterations in cervical cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: Gene expression omnibus (GEO).

AUTHOR CONTRIBUTIONS

SC conceptualized the project. KK and SB performed the analysis. KK and SC analyzed the data and drafted the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

SC acknowledges the Systems Medicine Cluster (SyMeC) Grant (GAP357), Department of Biotechnology (DBT) for funding. KK acknowledges the Department of Biotechnology (DBT) for fellowship (DBT/2016/IICB/491). SB acknowledges the Council of Scientific and Industrial Research (CSIR) for fellowship grant [31/002(1090)/2017-EMR-I].

ACKNOWLEDGMENTS

The authors acknowledge CSIR-Indian Institute of Chemical Biology for infrastructural support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.736495/full#supplementary-material>

REFERENCES

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2020) 68:394–424. doi: 10.3322/caac.21492
- World Health, Organization. *Human Papillomavirus (HPV) and Cervical Cancer.* (2020). Available online at: [https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-\(hpv\)-and-cervical-cancer](https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer) (accessed June 10, 2020).
- Anttila T, Saikku P, Koskela P, Bloigu A, Dillner J, Ikäheimo I, et al. Serotypes of *Chlamydia trachomatis* and risk for development of cervical squamous cell carcinoma. *J Am Med Assoc.* (2001) 285:47–51. doi: 10.1001/jama.285.1.47
- Jee B, Yadav R, Pankaj S, Shahi SK. Immunology of HPV-mediated cervical cancer: current understanding. *Int Rev Immunol.* (2020) 40:359–78. doi: 10.1080/08830185.2020.1811859
- Walboomers JM, Jacobs MV, Manos MM, Bosch FX, Kummer JA, Shah KV, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol.* (1999) 189:12–9. doi: 10.1002/(SICI)1096-9896(199909)189:1<12::AID-PATH431>3.0.CO;2-F
- Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsagué X, Shah KV, et al. Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N Engl J Med.* (2003) 348:518–27. doi: 10.1056/NEJMoa021641
- de Sanjose S, Quint WG, Alemany L, Geraets DT, Klaustermeier JE, Lloveras B, et al. Human papillomavirus genotype attribution in invasive cervical cancer: a retrospective cross-sectional worldwide study. *Lancet Oncol.* (2010) 11:1048–56. doi: 10.1016/S1470-2045(10)70230-8
- Haedicke J, Iftner T. Human papillomaviruses and cancer. *RadiotherOncol.* (2013) 108:397–402. doi: 10.1016/j.radonc.2013.06.004
- Groves IJ, Coleman N. Pathogenesis of human papillomavirus-associated mucosal disease. *J Pathol.* (2015) 235:527–38. doi: 10.1002/path.4496
- Narisawasaito M, Kiyono T. Basic mechanisms of high-risk human papillomavirus-induced carcinogenesis: roles of E6 and E7 proteins. *Cancer Sci.* (2007) 98:1505–11. doi: 10.1111/j.1349-7006.2007.00546.x
- Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* (2011) 144:646–74. doi: 10.1016/j.cell.2011.02.013
- Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* (2000) 100:57–70. doi: 10.1016/S0092-8674(00)81683-9
- Ward PS, Thompson CB. Signaling in control of cell growth and metabolism. *Cold Spring Harbor Perspect Biol.* (2012) 4:7a006783. doi: 10.1101/cshperspect.a006783
- Mo Y, Wang Y, Zhang L, Yang L, Zhou M, Li X, et al. The role of Wnt signaling pathway in tumor metabolic reprogramming. *J Cancer.* (2019) 10:3789–97. doi: 10.7150/jca.31166
- Papa S, Choy PM, Bubici C. The ERK and JNK pathways in the regulation of metabolic reprogramming. *Oncogene.* (2019) 38:2223–40. doi: 10.1038/s41388-018-0582-8
- Martin-Martin N, Carracedo A, Torrano V. Metabolism and transcription in cancer: merging two classic tales. *Front Cell Dev Biol.* (2017) 5:119. doi: 10.3389/fcell.2017.00119
- Dong Y, Tu R, Liu H, Quing G. Regulation of cancer cell metabolism: oncogenic MYC in the driver's seat. *Sig Transduct Target Ther.* (2020) 5:124. doi: 10.1038/s41392-020-00235-2
- Machida K. Pluripotency transcription factors and metabolic reprogramming of mitochondria in tumor-initiating stem-like cells. *Antioxid Redox Signal.* (2018) 28:1080–9. doi: 10.1089/ars.2017.7241
- Rottiers V, Naar AM. MicroRNAs in metabolism and metabolic disorders. *Nat Rev Mol Cell Biol.* (2012) 13:239–50. doi: 10.1038/nrm3313
- Chen B, Li H, Zeng X, Yang P, Liu X, Zhou X, et al. Roles of microRNA on cancer cell metabolism. *J Transl Med.* (2012) 10:228. doi: 10.1186/1479-5876-10-228
- Singh PK, Mehla K, Hollingsworth MA, Johnson KR. Regulation of aerobic glycolysis by microRNAs in cancer. *Mol Cell Pharmacol.* (2011) 3:125–34.
- Ji Z, Su J, Liu C, Wang H, Huang D, Zhou X. Integrating genomics and proteomics data to predict drug effects using binary linear programming. *PLoS ONE.* (2014) 9:e102798. doi: 10.1371/journal.pone.0102798
- Cheng F, Murray JL, Zhao J, Sheng J, Zhou Z, Rubin DH, et al. Systems biology-based investigation of cellular antiviral drug targets identified by gene-trap insertional mutagenesis. *PLOS Comput Biol.* (2016) 12:e1005074. doi: 10.1371/journal.pcbi.1005074
- Puniya BL, Kulshreshtha D, Mittal I, Mobeen A, Ramachandran S. Integration of metabolic modeling with gene co-expression reveals transcriptionally programmed reactions explaining robustness in mycobacterium tuberculosis. *Sci Rep.* (2016) 6:23440. doi: 10.1038/srep24916
- Puniya BL, Allen L, Hochfelder C, Majumder M, Helikar T. Systems perturbation analysis of a large-scale signal transduction model reveals potentially influential candidates for cancer therapeutics. *Front Bioeng Biotechnol.* (2016) 4:10. doi: 10.3389/fbioe.2016.00010
- Samaga R, Klamt S. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. *Cell Commun Signal.* (2013) 11:43. doi: 10.1186/1478-811X-11-43
- Albert R, Thakar J. Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *Wiley Interdiscip Rev Syst Biol Med.* (2014) 6:353–69. doi: 10.1002/wsbm.1273
- Le Novère N. Quantitative and logic modelling of molecular and gene networks. *Nat Rev Genet.* (2015) 16:146–58. doi: 10.1038/nrg3885
- Naldi A, Monteiro PT, Müsel C, Consortium for Logical, Models, Tools., et al. Cooperative development of logical modelling standards and tools with CoLoMoTo. *Bioinformatics.* (2015) 31:1154–9. doi: 10.1093/bioinformatics/btv013
- Saez-Rodriguez J, Alexopoulos LG, Zhang M, Morris MK, Lauffenburger DA, Sorger PK. Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res.* (2011) 71:5400–11. doi: 10.1158/0008-5472.CAN-10-4453
- Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol.* (2002) 20:370–5. doi: 10.1038/nbt0402-370
- Huang Z, Mayr NA, Yuh WT, Lo SS, Montebello JF, Grecula JC, et al. Predicting outcomes in cervical cancer: a kinetic model of tumor regression during radiation therapy. *Cancer Res.* (2010) 70:463–70. doi: 10.1158/0008-5472.CAN-09-2501
- Chaouiya C. Petri net modelling of biological networks. *Brief Bioinform.* (2007) 8:210–9. doi: 10.1093/bib/bbm029
- Hautaniemi S, Kharait S, Iwabu A, Wells A, Lauffenburger DA. Modeling of signal-response cascades using decision tree analysis. *Bioinformatics.* (2005) 21:2027–35. doi: 10.1093/bioinformatics/bti278
- Peng H, Zhao W, Tan H, Ji Z, Li J, Li K, et al. Prediction of treatment efficacy for prostate cancer using a mathematical model. *Sci Rep.* (2016) 6:21599. doi: 10.1038/srep21599
- Orth JD, Thiele I, Palsson BO. What is flux balance analysis? *Nat Biotechnol.* (2010) 28:245–8. doi: 10.1038/nbt.1614
- Bag AK, Mandloi S, Jarmalavicius S, Mondal S, Kumar K, Mandal C, et al. Connecting signaling and metabolic pathways in EGF receptor-mediated oncogenesis of glioblastoma. *PLoS Comput Biol.* (2019) 15:e1007090. doi: 10.1371/journal.pcbi.1007090
- Bhattacharyya M, Chakrabarti S. Identification of important interacting proteins (IIPs) in *Plasmodium falciparum* using large-scale interaction network analysis and in-silico knock-out studies. *Malar J.* (2015) 14:70. doi: 10.1186/s12936-015-0562-1
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* (2002) 30:207–10. doi: 10.1093/nar/30.1.207
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* (2004) 3:3. doi: 10.2202/1544-6115.1027
- Smyth GK. Limma: linear models for microarray data. In: eds Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York, NY: Springer (2005). p. 397–420. doi: 10.1007/0-387-29362-0_23

42. Davis S, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and bioconductor. *Bioinformatics*. (2007) 23:1846–7. doi: 10.1093/bioinformatics/btm254
43. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B*. (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
44. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. (2019) 47:D607–13. doi: 10.1093/nar/gky1131
45. Bovolenta LA, Acencio ML, Lemke N. HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics*. (2012) 13:405. doi: 10.1186/1471-2164-13-405
46. Han H, Cho JW, Lee S, Yun A, Kim H, Bae D, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res*. (2018) 46:D380–6. doi: 10.1093/nar/gkx1013
47. Huang HY, Lin YC, Li J, Huang KY, Shrestha S, Hong HC, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res*. (2020) 48:D148–4. doi: 10.1093/nar/gkz896
48. Karagkouni D, Paraskevopoulou MD, Chatzopoulos S, Vlachos IS, Tastsoglou S, et al. DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interaction. *Nucleic Acids Res*. (2018) 46:D239–45. doi: 10.1093/nar/gkx1141
49. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, and Millman J, editors. *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA (2008). p. 11–15.
50. Liao Y, Wang J, Jaehnig E, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. (2019) 47:W199–205. doi: 10.1093/nar/gkz401
51. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. (2017) 45:D353–61. doi: 10.1093/nar/gkw1092
52. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. (2000) 28:27–30. doi: 10.1093/nar/28.1.27
53. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. (2003) 13:2498–504. doi: 10.1101/gr.1239303
54. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. *Genome Res*. (2009) 19:1639–45. doi: 10.1101/gr.092759.109
55. Yang K, Xia B, Wang W, Cheng J, Yin M, Xie H, et al. A comprehensive analysis of metabolomics and transcriptomics in cervical cancer. *Sci Rep*. (2017) 7:43353. doi: 10.1038/srep43353
56. Park J, Shin Y, Kim TH, Kim DH, Lee A. Plasma metabolites as possible biomarkers for diagnosis of breast cancer. *PLoS ONE*. (2019) 14:e0225129. doi: 10.1371/journal.pone.0225129
57. Turkoglu O, Zeb A, Graham S, Szyperski T, Szender JB, Odunsi K, et al. Metabolomics of biomarker discovery in ovarian cancer: a systematic review of the current literature. *Metabolomics*. (2016) 12:60. doi: 10.1007/s11306-016-0990-0
58. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Freshno R, et al. HMDB 4.0 – the human metabolome database for (2018). *Nucleic Acids Res*. (2018) 46:D608–17. doi: 10.1093/nar/gkx1089
59. Gyorffy B. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Comput Struct Biotechnol J*. (2021) 19:4101–9. doi: 10.1016/j.csbj.2021.07.014
60. Nagy A, Munkacsy G, Gyorffy B. Pancancer survival analysis of cancer hallmark genes. *Sci Rep*. (2021) 11:6047. doi: 10.1038/s41598-021-84787-5
61. Fekete J, Gyorffy B. ROCplot.org: Validating predictive biomarkers of chemotherapy/hormonal therapy/anti-Her2 therapy using transcriptomic data of 3,104 breast cancer patients. *Int J Cancer*. (2019) 145:3140–51. doi: 10.1002/ijc.32369
62. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. *Cell*. (2011) 144:986–98. doi: 10.1016/j.cell.2011.02.016
63. Sevimoglu T, Arga KY. The role of protein interaction networks in systems biomedicine. *Comput Struct Biotechnol J*. (2014) 11:22–7. doi: 10.1016/j.csbj.2014.08.008
64. Wang J, Yi Y, Chen Y, Xiong Y, Zhang W. Potential mechanism of RRM2 for promoting cervical cancer based on weighted gene co-expression network analysis. *Int J Med Sci*. (2020) 17:2362–72. doi: 10.7150/ijms.47356
65. Rižner TL. Enzymes of the AKR1B and AKR1C subfamilies and uterine diseases. *Front Pharmacol*. (2012) 3:34. doi: 10.3389/fphar.2012.00034
66. Tai HL, Lin TS, Huang HH, Lin TY, Chou MC, Chiou SH, et al. Overexpression of aldo-keto reductase 1C2 as a high-risk factor in bladder cancer. *Oncol Rep*. (2007) 17:305–11. doi: 10.3892/or.17.2.305
67. Liu D, Mao Y, Chen C, Zhu F, Lu W, Ma H. Expression patterns and clinical significances of ENO2 in lung cancer: an analysis based on oncomine database. *Ann Transl Med*. (2020) 8:639. doi: 10.21037/atm-20-3354
68. Yang HJ, Xue JM, Li J, Wan LH, Zhu YX. Identification of key genes and pathways of diagnosis and prognosis in cervical cancer by bioinformatics analysis. *Mol Genet Genom Med*. (2020) 8:e1200. doi: 10.1002/mggg.1200
69. Liu Y, Liu T, Bao X, He M, Li L, Yang X. Increased EZH2 expression is associated with proliferation and progression of cervical cancer and indicates a poor prognosis. *Int J Gynecol Pathol*. (2014) 33:218–24. doi: 10.1097/PGP.0b013e31829c6574
70. Sharma G, Dua P, Agarwal SM. A comprehensive review of dysregulated miRNAs involved in cervical cancer. *Curr Genomics*. (2014) 15:310–23. doi: 10.2174/1389202915666140528003249
71. Wang JY, Chen LJ. The role of miRNAs in the invasion and metastasis of cervical cancer. *Biosci Rep*. (2019) 39:BSR20181377. doi: 10.1042/BSR20181377
72. Pardini B, De Maria D, Francavilla A, Di Gaetano C, Ronco G, Naccarati A. MicroRNAs as markers of progression in cervical cancer: a systematic review. *BMC Cancer*. (2018) 18:696. doi: 10.1186/s12885-018-4590-4
73. Ma Z, Cai Y, Zhang L, Tian C, Lyu L. LINC00319 promotes cervical cancer progression via targeting miR-147a/IGF1R pathway. *Cancer Biother Radiopharm*. (2020). doi: 10.1089/cbr.2020.3722. [Epub ahead of print].
74. Wang F, Zhang H, Xu N, Huang N, Tian C, Ye A, et al. A novel hypoxia-induced miR-147a regulates cell proliferation through a positive feedback loop of stabilizing HIF-1 α . *Cancer Biol Ther*. (2016) 17:790–8. doi: 10.1080/15384047.2016.1195040
75. Wei F, Wang M, Li Z, Wang Y, Zhou Y. miR-593 inhibits proliferation and invasion and promotes apoptosis in non-small cell lung cancer cells by targeting SLUG-associated signaling pathways. *Mol Med Rep*. (2019) 20:5172–82. doi: 10.3892/mmr.2019.10776
76. Han W, Wang L, Zhang L, Wang Y, Li Y. Circular RNA circ-RAD23B promotes cell growth and invasion by miR-593-3p/CCND2 and miR-653-5p/TIAM1 pathways in non-small cell lung cancer. *Biochem Biophys Res Commun*. (2019) 510:462–6. doi: 10.1016/j.bbrc.2019.01.131
77. Song L, Xiao Y. Downregulation of hsa_circ_0007534 suppresses breast cancer cell proliferation and invasion by targeting miR-593/MUC19 signal pathway. *Biochem Biophys Res Commun*. (2018) 503:2603–10. doi: 10.1016/j.bbrc.2018.08.007
78. Dong L, Hong H, Chen X, Huang Z, Wu W, Wu F. LINC02163 regulates growth and epithelial-to-mesenchymal transition phenotype via miR-593-3p/FOXK1 axis in gastric cancer cells. *Artif Cells Nanomed Biotechnol*. (2018) 46:607–15. doi: 10.1080/21691401.2018.1464462
79. Yu H, Wei W, Cao W, Zhan Z, Yan L, Wu K, et al. Regulation of cell proliferation and metastasis by microRNA-593-5p in human gastric cancer. *Oncol Targets Ther*. (2018) 11:7429–40. doi: 10.2147/OTT.S178151
80. Sinkala M, Zulu M, Nkhoma P, Kafita D, Zulu E, Tembo R, et al. A systems approach identifies key regulators of hpv-positive cervical cancer. *medRxiv [Preprint]*. (2020). doi: 10.1101/2020.05.12.20099424
81. Ruiz G, Valencia-González HA, Pérez-Montiel D, Muñoz F, Ocádiz-Delgado R, Fernández-Retana J, et al. Genes involved in the transcriptional regulation of pluripotency are expressed in malignant tumors of the uterine cervix and can induce tumorigenic capacity in a nontumorigenic cell line. *Stem Cells Int*. (2019) 2019:7683817. doi: 10.1155/2019/7683817
82. Biswas N, Kumar K, Bose S, Bera R, Chakrabarti S. Analysis of pan-omics data in human interactome network (APODHIN).

Front Genet. (2020) 11:589231. doi: 10.3389/fgene.2020.589231

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

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

this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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



Clinical Significance and Prognostic Value of Human Soluble Resistance-Related Calcium-Binding Protein: A Pan-Cancer Analysis

Jinguo Zhang*, Jian Chen, Benjie Shan, Lin Lin, Jie Dong, Qingqing Sun, Qiong Zhou and Xinghua Han*

Department of Medical Oncology, The First Affiliated Hospital of USTC, Division of Life Science and Medicine, University of Science and Technology of China, Hefei, China

OPEN ACCESS

Edited by:

Thirumal Kumar D,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Simin Li,
Southern Medical University, China
Yang Yu,
First Affiliated Hospital of Harbin
Medical University, China

*Correspondence:

Jinguo Zhang
Zhangjg2019@163.com
orcid.org/0000-0001-9952-9307
Xinghua Han
hxhmail@ustc.edu.cn
orcid.org/0000-0002-0632-6955

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 03 August 2021

Accepted: 12 October 2021

Published: 16 November 2021

Citation:

Zhang J, Chen J, Shan B, Lin L,
Dong J, Sun Q, Zhou Q and Han X
(2021) Clinical Significance and
Prognostic Value of Human Soluble
Resistance-Related Calcium-Binding
Protein: A Pan-Cancer Analysis.
Front. Med. 8:752619.
doi: 10.3389/fmed.2021.752619

The soluble resistance-related calcium-binding protein (sorcina, SRI) serves as the calcium-binding protein for the regulation of calcium homeostasis and multidrug resistance. Although the mounting evidence suggests a crucial role of SRI in the chemotherapeutic resistance of certain types of tumors, insights into pan-cancer analysis of SRI are unavailable. Therefore, this study aimed to probe the multifaceted properties of SRI across the 33 cancer types. The SRI expression was analyzed via The Cancer Genome Atlas (TCGA) and Genotype Tissue-Expression (GTEx) database. The SRI genomic alterations and drug sensitivity analysis were performed based on the cBioPortal and the CellMiner database. Furthermore, the correlations among the SRI expression and survival outcomes, clinical features, stemness, tumor mutation burden (TMB), microsatellite instability (MSI), and immune cells infiltration were analyzed using TCGA data. The differential analysis showed that SRI was upregulated in 25 tumor types compared with the normal tissues. Aberrant expression of SRI was able to predict survival in different cancers. Further, the most frequent alteration of SRI genomic was amplification. Moreover, the aberrant SRI expression was related to stemness score, epithelial-mesenchymal-transition (EMT)-related genes, MSI, TMB, and tumor immune microenvironment in various types of cancer. TIMER database mining further found that the SRI expression was significantly correlated with the infiltration levels of various immune cells in certain types of cancer. Intriguingly, the SRI expression was negatively correlated with drug sensitivity of fluorouracil, paclitaxel, docetaxel, and isotretinoin. Our findings highlight the predictive value of SRI in cancer and provide insights for illustrating the role of SRI in tumorigenesis and drug resistance.

Keywords: pan-cancer, sorcina, prognosis (carcinoma), MSI, TMB

INTRODUCTION

Recent statistics showed that cancer has become a worldwide public health issue with an estimated 1,898,160 new cancer cases and 608,570 cancer deaths in 2021 (1). In recent decades, great advances have been achieved in the diagnostics and treatment of cancer, in particular checkpoint blockade-based immunotherapy (2). Currently, reliable predictive biomarkers and new

immunotherapy targets have attracted considerable attention among scientists. The SRI gene, which encodes the soluble resistance-related calcium-binding protein (sorcin), is located at chromosome 7q21.12 spanning about 21.9 kb of the human genome (3). Sorcin serves as a calcium-binding protein that is a member of the penta-EF hand (PEF) family (4). Sorcin exists in a soluble form in a low cytoplasmic calcium state but translocates to the membrane to exercise its function in a high cytoplasmic calcium state (5). Classically, sorcin holds a crucial role in the regulation of calcium homeostasis through diverse mechanisms. Sorcin not only directly binds to calcium but also interacts with an L-type calcium channel and cardiac ryanodine receptor-2 to modulate calcium balance (6, 7). The aberrant expression of SRI is reported to be associated with the neurodegenerative diseases, hereditary spherocytosis cells, and women with unexplained infertility (8–10). However, the role of SRI in cancer is receiving gradually more attention.

The previous studies proved that SRI acted as a pro-oncogenic and multidrug resistance gene in certain types of cancers (11–15). The resistance to chemotherapeutic agents is recognized as a major hurdle in cancer therapy. In the past years, MDR1 (ABCB1, P-glycoprotein, and P-gp) and MDR-associated protein 1 (MRP1) were the most extensively investigated resistance proteins in cancer (16, 17). However, SRI as a novel resistance gene has begun to attract substantial attention from scientists (18). Of note, SRI and MDR1 co-localize on the same amplicon and often co-amplify in multidrug-resistant tumor cells (19). The overexpression of sorcin contributed to resistance to many chemotherapy agents, and sorcin knockdown has been found to reverse the multidrug resistance in certain types of cancer (20–22). To date, most of the studies on SRI in tumors are limited by a specific cancer type and many studies have focused on *in vitro* cellular level. Therefore, dissecting the role of SRI in pan-cancer is required.

Our previous findings revealed that SRI promoted the paclitaxel resistance and malignant progression in ovarian cancer (23). Thus, here we postulated that SRI might function as a critical oncogenic, resistant effector in pan-cancer, and played crucial roles in cancer immunity. To uncover the role of SRI in pan-cancer, we systematically integrated multiple databases from bioinformatics point of view. In this study, the expression of SRI was comprehensively investigated in normal tissues and their cancer counterparts using Genotype-Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), and Oncomine database. Meanwhile, the prognostic value of SRI to predict the survival outcomes was also evaluated. Then, the potential relationships among the SRI expression and clinical features, cancer stemness score, tumor mutation burden (TMB), microsatellite instability (MSI), and infiltrating immune cells were explored in pan-cancer. In addition, the SRI genomic alternations and effect of SRI on the drug sensitivity were determined using the cBioPortal and the CellMiner database. Further, the gene set enrichment analysis (GSEA) was applied to elucidate the biological function of SRI in cancer. Overall, this study highlights the multifaceted role of SRI in pan-cancer, which provides a rationale for targeting SRI as a novel therapeutic strategy.

METHODS

Data Processing and the SRI Expression Analysis

Publicly available transcriptome data of TCGA pan-cancer and the related clinical features were obtained from the UCSC XENA (<https://xena.ucsc.edu/>). The expression matrices of 31 human normal tissues were downloaded from GTEx web portal (<https://www.gtexportal.org/>). The strawberry Perl script was developed (version 5.30.0.1, <http://strawberryperl.com/>) to extract the SRI expression data in 33 TCGA tumor types and GTEx normal tissues. The mRNA level of SRI in the healthy men and women tissues was visualized with “gganatomogram” R package. The expression data were log2(TPM) transformed excluding missing data and duplicated values. The differences in the SRI expression between the tumor and normal tissues were examined by the Wilcoxon rank-sum test. The differential expression of SRI mRNA was evaluated in various tumor types using the Oncomine database (www.oncomine.org) (24). All the analyses were conducted using the R version 4.0.2 software (<https://www.Rproject.org/>). The overall workflow of our study is presented in Figure 1A.

Correlation of SRI Expression With Survival Prognosis and Clinical Features

The survival and clinical characteristics data were obtained from the UCSC XENA repository. The association between the SRI expression level and survival outcomes, such as overall survival (OS), progression-free interval (PFS), disease-specific survival (DSS), and disease-free survival (DFS), was evaluated using Kaplan–Meier method and Cox proportional hazards model. The “survival,” “survminer,” and “forestplot” packages were employed to draw the Kaplan–Meier and forest plots. The clinical records, such as patients age, tumor stage, and tumor status, were applied to investigate their relationship with the SRI expression. The data representations were performed using “limma” and “ggpubr” R-packages.

Genomic Alterations SRI in Cancers and the Co-expression Gene Analysis

The SRI gene alternations in TCGA pan-cancer datasets across 10,953 patients were analyzed using cBioPortal database (<http://www.cbioportal.org/>) (25). The “Oncoprint” and “Cancer Type Summary” modules were used to investigate the genetic alterations of SRI. The “Mutations” module was applied to obtain the mutated site information of SRI. The effect of SRI alternations on the OS and gene mutation co-occurrence analysis were obtained from the “Comparison/Survival” module. Furthermore, the copy number alterations data, mRNA Expression Z-scores data, and protein level Z-scores data of SRI in ovarian cancer were downloaded from the TCGA PanCancer Atlas dataset. To investigate the co-expression genes of SRI in ovarian cancer, the TCGA-OV RNA-seq data were accessed using the Linkedomics online tool (<http://www.linkedomics.org/login.php/>). The top 50 positive and negative correlated genes were selected to construct a co-expression network. The network was visualized in the STRING database (<https://string-db.org/>).

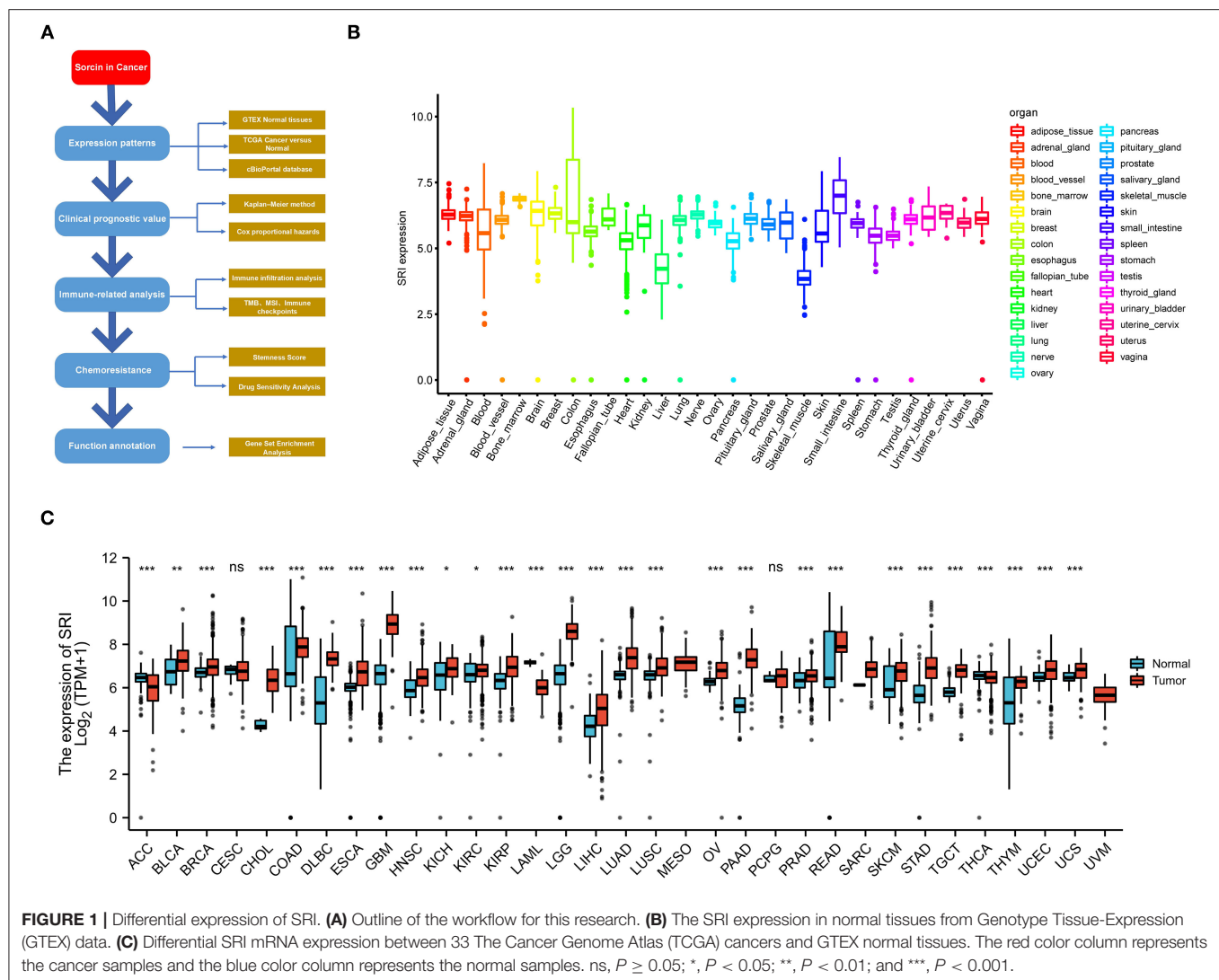


FIGURE 1 | Differential expression of SRI. **(A)** Outline of the workflow for this research. **(B)** The SRI expression in normal tissues from Genotype Tissue-Expression (GTEx) data. **(C)** Differential SRI mRNA expression between 33 The Cancer Genome Atlas (TCGA) cancers and GTEx normal tissues. The red color column represents the cancer samples and the blue color column represents the normal samples. ns, $P \geq 0.05$; *, $P < 0.05$; **, $P < 0.01$; and ***, $P < 0.001$.

Correlation Analysis of SRI Expression With Stemness Score and EMT-Related Genes in Cancers

Cancer stemness was reported to be capable of evaluating by RNA stemness score (RNAss) based on mRNA expression (26). Correlation analysis between SRI expression and RNAss was examined using Spearman rank-based testing. Data were visualized with the R-package “corrplot”. We further determined the correlation between the level of SRI and cancer stem cell marker genes, epithelial-mesenchymal-transition (EMT)-related genes. The heatmaps were generated with the “reshape2” and “RColorBrewer” R-packages.

Relationship Between SRI Expression and TMB, MSI in Pan-Cancer

Recent studies have revealed that TMB and MSI could become independent biomarkers for immune checkpoint inhibitors response (27, 28). The TCGA pan-cancer mutation data were

applied to calculate the TMB scores of each sample. The MSI scores of the TCGA pan-cancer samples were obtained according to a previous study (29). The analyses regarding the correlations between the SRI expression and TMB, MSI were calculated using Spearman’s coefficient. The results were displayed as radar plots using the R-package “fmsb.” The tumors with high microsatellite instability are often characterized by a defective DNA mismatch repair system (MMR) system (30). The association between the SRI expression and MMR-related genes was further explored in cancers. The results were visualized as the heatmaps using the “reshape2” and “RColorBrewer” R-packages.

Association Analysis of the SRI Expression With Tumor Immune Microenvironment in Cancers

The Estimation of Stromal and Immune cells in Malignant Tumors using Expression data (ESTIMATE) algorithm was performed to calculate the ImmuneScore, StromalScore, and ESTIMATEScore using the R package “ESTIMATE” (31).

The results were displayed using the R-package “corrplot” using Spearman’s rank-based testing. Moreover, we analyzed the correlation of SRI expression with the abundance of various immune cell infiltrates in pan-cancer using the TIMER database (32) (<https://cistrome.shinyapps.io/timer/>). In addition, a correlation analysis of SRI and immune checkpoint genes was performed. The results were processed using the “reshape2” and “RColorBrewer” R-packages.

Drug Sensitivity of SRI and GSEA

The NCI-60 compound activity data and the RNA-seq expression profiles were obtained from CellMiner (<https://discover.nci.nih.gov/cellminer/home.do>) (33). The drugs that were considered FDA approved or in the clinical trials were selected for further analysis. Then, the effect of SRI on the drug sensitivity was analyzed using the “impute,” “limma,” “ggplot2,” and “ggpubr” R package. The GSEA analysis was performed to explore the biological functions of SRI in cancer. The gene sets of Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) signature were obtained from GSEA online (<https://www.gsea-msigdb.org/gsea/downloads.jsp>). The GSEA analysis was processed with the R-packages “enrichplot,” “org.Hs.eg.db,” “clusterProfiler,” “DOSE,” and “limma”.

Sorcin Expression in the Ovarian Cancer Sphere Cells

The cell culture and sphere formation assays were performed according to the previously described method (34). The sorcin expression in the ovarian cancer sphere cells was evaluated using a western blot. Anti-SRI antibody (ab71983) was purchased from Abcam (Cambridge, UK).

Statistical Analysis

A log-rank test was applied in the Kaplan–Meier survival curves. Hazard ratio (HR) was determined by a Cox proportional hazard regression model. The correlation analysis was executed by Spearman’s rank test. The statistical analyses were performed using the R version 4.0.2 software (R Foundation for Statistical Computing, Vienna, Austria) or GraphPad Prism 8 (San Diego, CA, USA). Results with $P < 0.05$ were considered statistically significant (* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, and **** $P < 0.0001$).

RESULTS

Expression Levels of SRI in Normal Tissues and Pan-Cancer

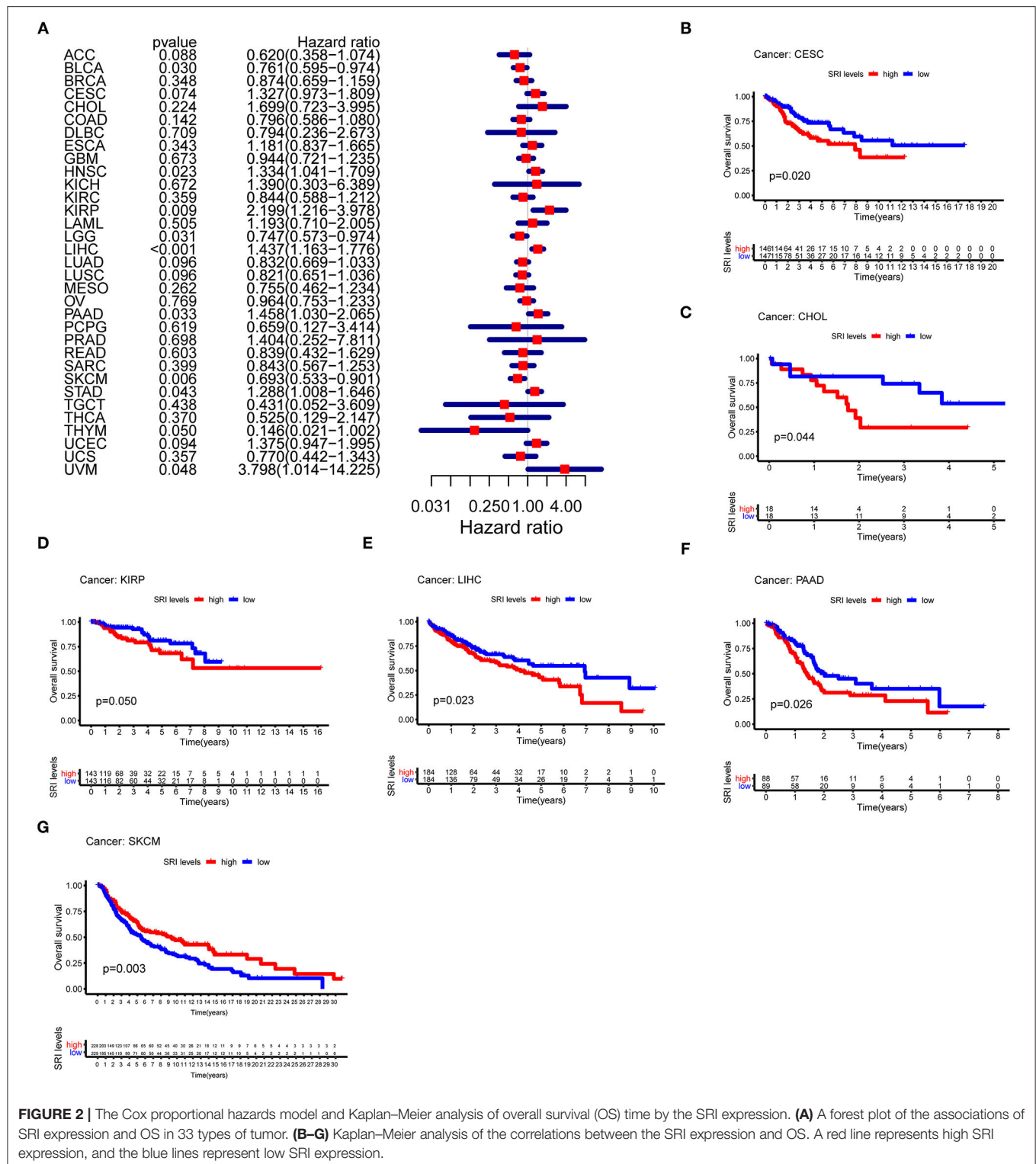
To gain insights into the expression pattern of SRI in the human normal tissues, the SRI expression in tissue physiological state was investigated according to GTEX dataset. SRI was highly expressed in the small intestine, bone marrow, brain, and breast tissues, while the skeletal muscle and liver tissues expressed low levels of SRI (Figure 1B). The SRI expression abundances of various tissues in men and women are displayed in

Supplementary Figures 1A,B. Overall, no gender difference was observed in the mRNA expression levels of SRI (Supplementary Figure 1C).

To further explore the SRI expression in human cancers, the SRI expression in various types of cancers was analyzed using the RNA-seq data of TCGA database (Figure 1C). The aberrant expression of SRI was detected in 28 types of cancer except for those cancers where no normal tissue data were available. The SRI expression was significantly higher in bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), brain lower grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), and uterine carcinosarcoma (UCS). In contrast, the SRI levels were significantly downregulated in adrenocortical carcinoma (ACC), acute myeloid leukemia (LAML), and thyroid carcinoma (THCA). Aside from this, the higher level of SRI was reconfirmed in the brain, bladder, esophageal, gastric, head and neck, kidney, liver, melanoma, myeloma, pancreatic, and prostate cancer compared with the normal tissues using ONCOMINE database (Supplementary Figure 1D).

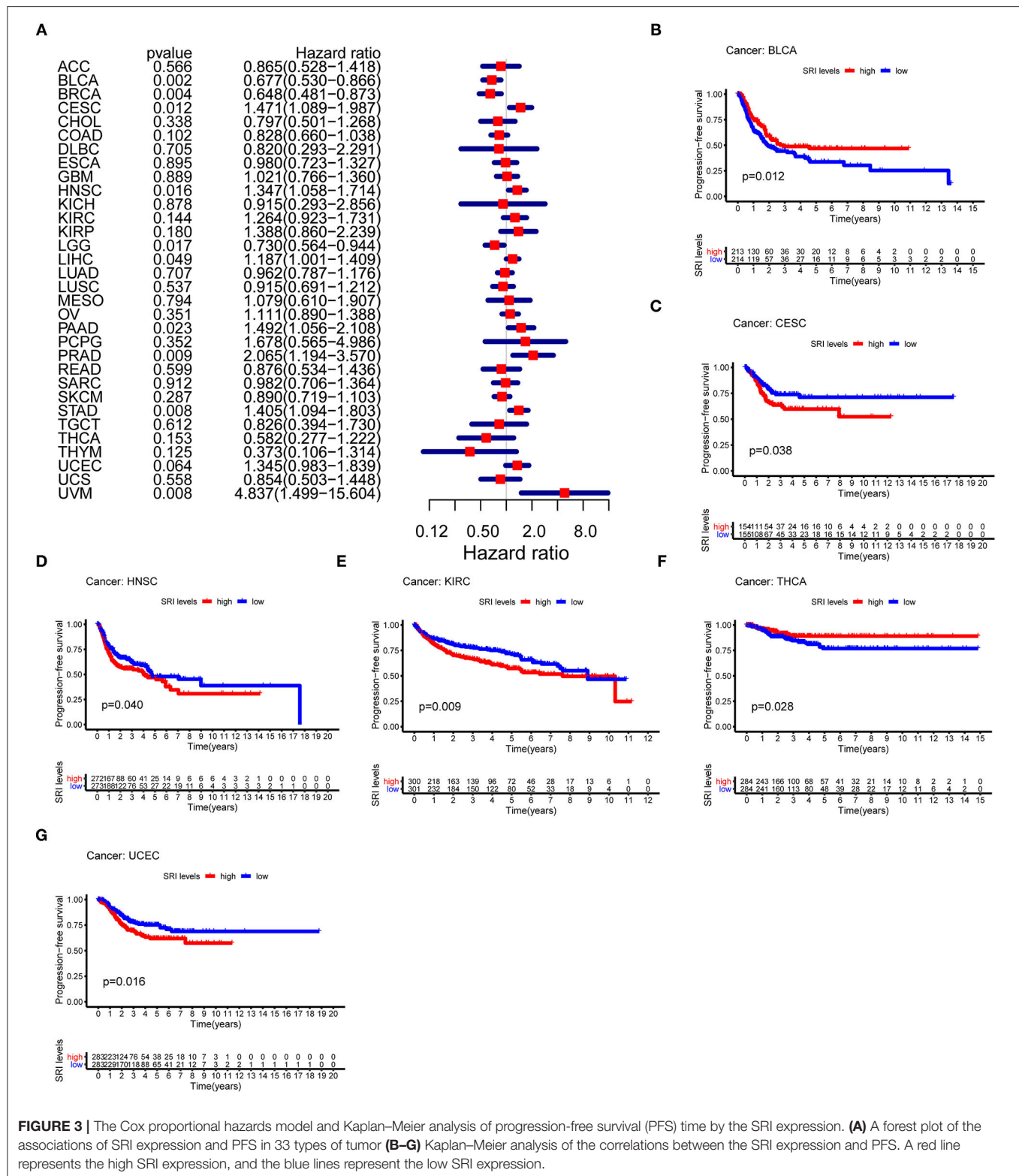
Prognostic Value of SRI in Pan-Cancer

To further explore the prognostic value of SRI in pan-cancer, the association between the SRI expression level and survival of patients was evaluated using the Cox proportional hazards model and the Kaplan–Meier analysis. The results from a Cox proportional hazards regression model revealed that the SRI expression levels were correlated with OS in BLCA ($P = 0.030$), HNSC ($P = 0.023$), KIRP ($P = 0.009$), LIHC ($P < 0.001$), LGG ($P = 0.001$), PAAD ($P = 0.033$), SKCM ($P = 0.006$), STAD ($P = 0.043$), THYM ($P = 0.05$), and uveal melanoma (UVM) ($P = 0.048$) (Figure 2A). SRI served as a high-risk factor for HNSC, KIRP, LIHC, PAAD, STAD, and UVM, while it acted as a low-risk gene for BLCA, LGG, SKCM, and THYM. Furthermore, the Kaplan–Meier survival analysis demonstrated that the high level of SRI predicted poor OS in cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) (Figure 2B, $P = 0.020$), CHOL (Figure 2C, $P = 0.044$), KIRP (Figure 2D, $P = 0.050$), LIHC (Figure 2E, $P = 0.023$), and PAAD (Figure 2F, $P = 0.026$). While low expression of SRI was correlated with the shortened OS in SKCM (Figure 2G, $P = 0.003$). Regarding the PFS, the high SRI level represented an adverse factor in CESC ($P = 0.012$), HNSC ($P = 0.016$), LIHC ($P = 0.049$), PAAD ($P = 0.023$), PRAD ($P = 0.009$), STAD ($P = 0.008$),



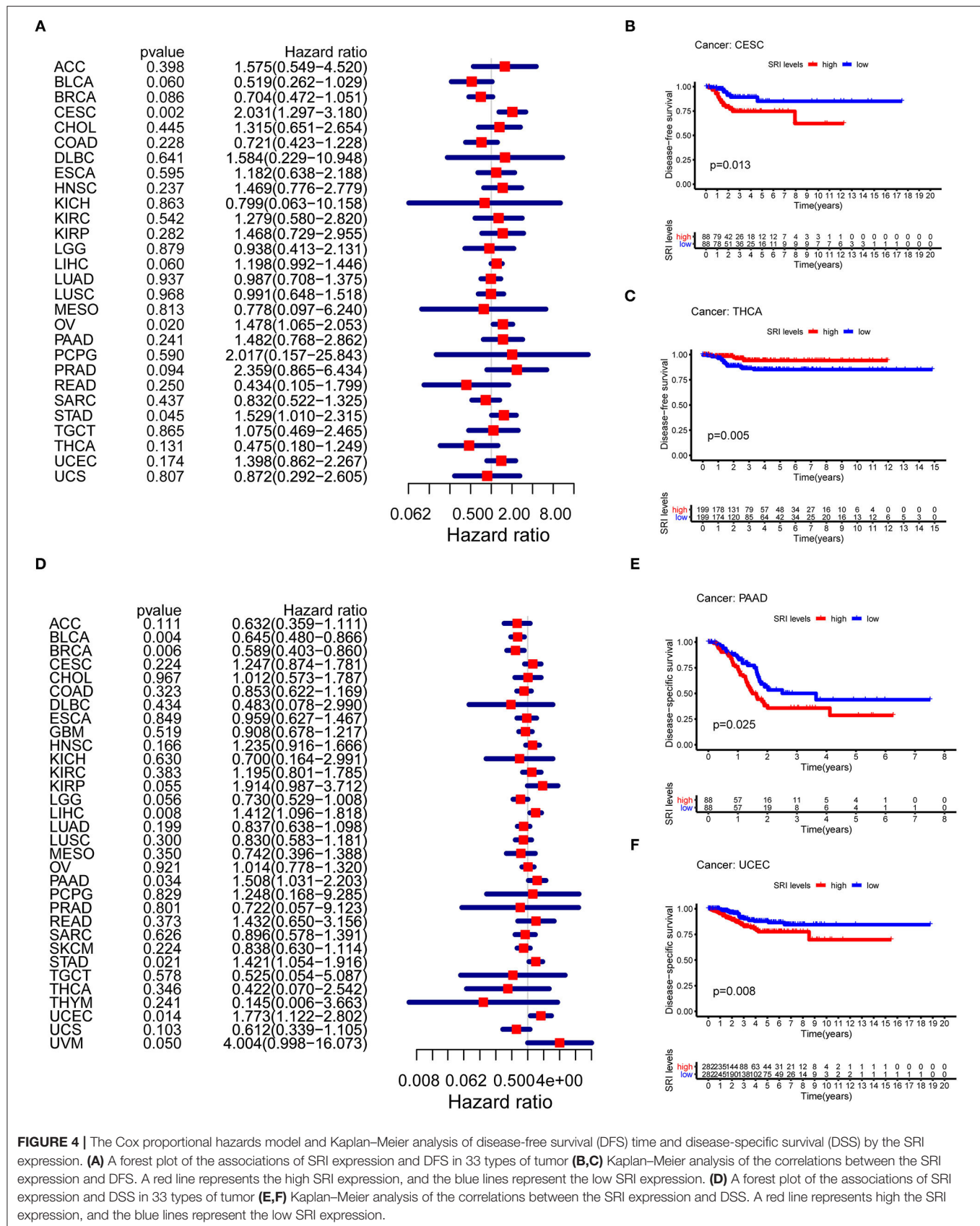
and UVM ($P = 0.008$), while the high SRI level was considered as a favorable factor in BLCA ($P = 0.002$), BRCA ($P = 0.004$), and LGG ($P = 0.017$) (Figure 3A). The Kaplan–Meier curves for PFS revealed a correlation between the high SRI

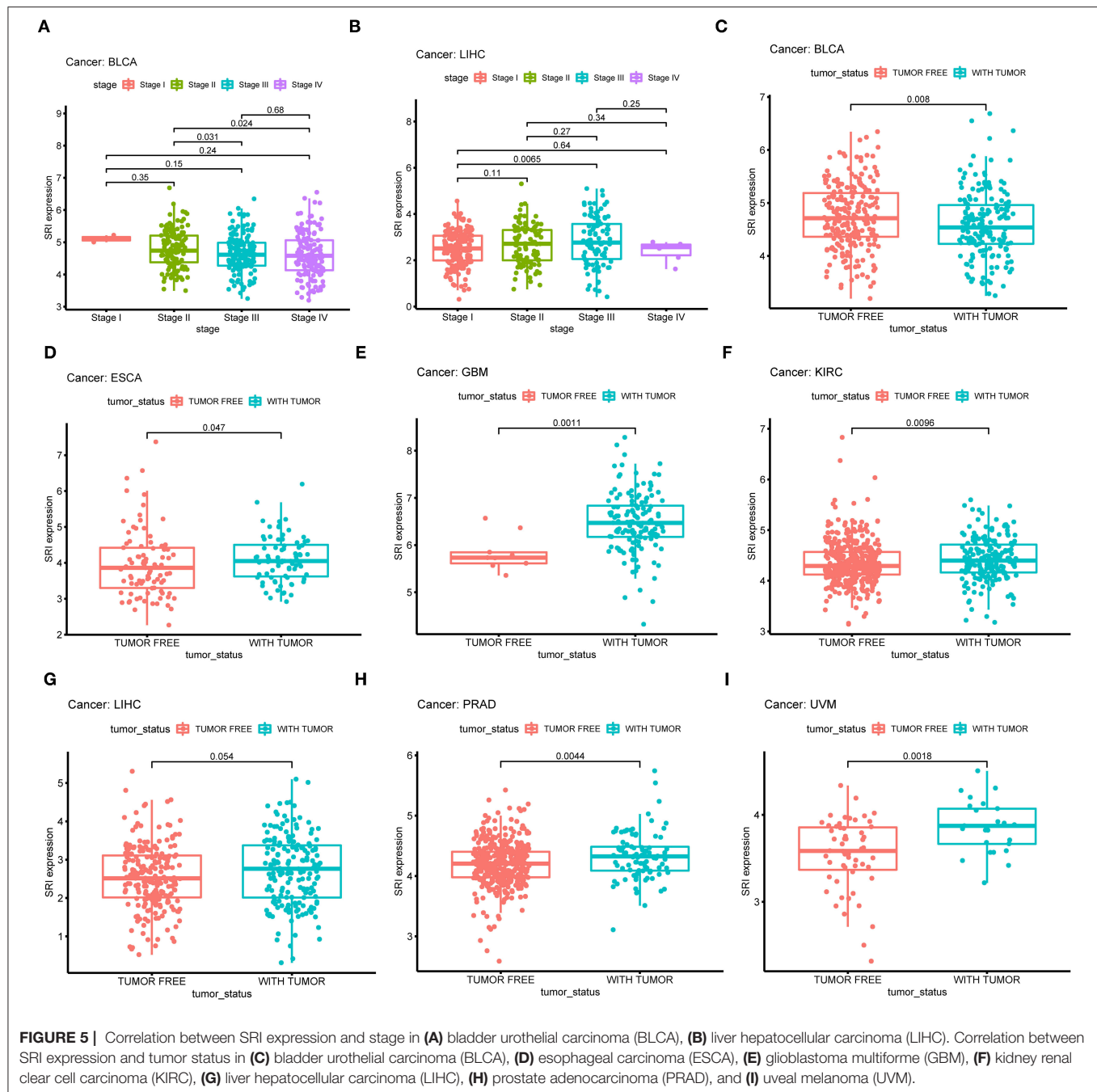
expression level and poor survival time in the patients with CESC (Figure 3C, $P = 0.038$), HNSC (Figure 3D, $P = 0.040$), KIRC (Figure 3E, $P = 0.009$), and UCEC (Figure 3G, $P = 0.016$). The patients with high SRI expression had significantly



longer PFS than the patients with low expression in BLCA (Figure 3B, $P = 0.012$) and THCA (Figure 3F, $P = 0.028$). Regarding the associations between the SRI expression and

DFS, the forest plots showed high SRI expression predicted poor DFS in CESC ($P = 0.002$), OV ($P = 0.020$), and STAD ($P = 0.040$) (Figure 4A). Significant relationships between the





SRI expression and DFS were observed in CESC (Figure 4B, $P = 0.013$) and THCA (Figure 4C, $P = 0.005$) by the Kaplan–Meier survival analysis. Furthermore, SRI exhibited a significant prognostic value in BLCA ($P = 0.004$), BRCA ($P = 0.006$), LIHC ($P = 0.008$), PAAD ($P = 0.034$), STAD ($P = 0.021$), UCEC ($P = 0.014$), and UVM ($P = 0.05$) in a Cox proportional hazards regression model for DSS (Figure 4D). The Kaplan–Meier survival analysis found that the PAAD (Figure 4E, $P = 0.025$) and UCEC (Figure 4F, $P = 0.008$) patients with high expression of SRI had shortened DSS. Additionally, the survival

analysis of SRI in some types of cancer was validated in the GEO database (Supplementary Figure 2).

Correlation Between the SRI Expression and Clinicopathological Phenotypes in Cancers

Next, the correlations between the SRI expression and the clinicopathological features of patients were investigated in pan-cancer. Patients with age ≥ 65 years had higher expression of

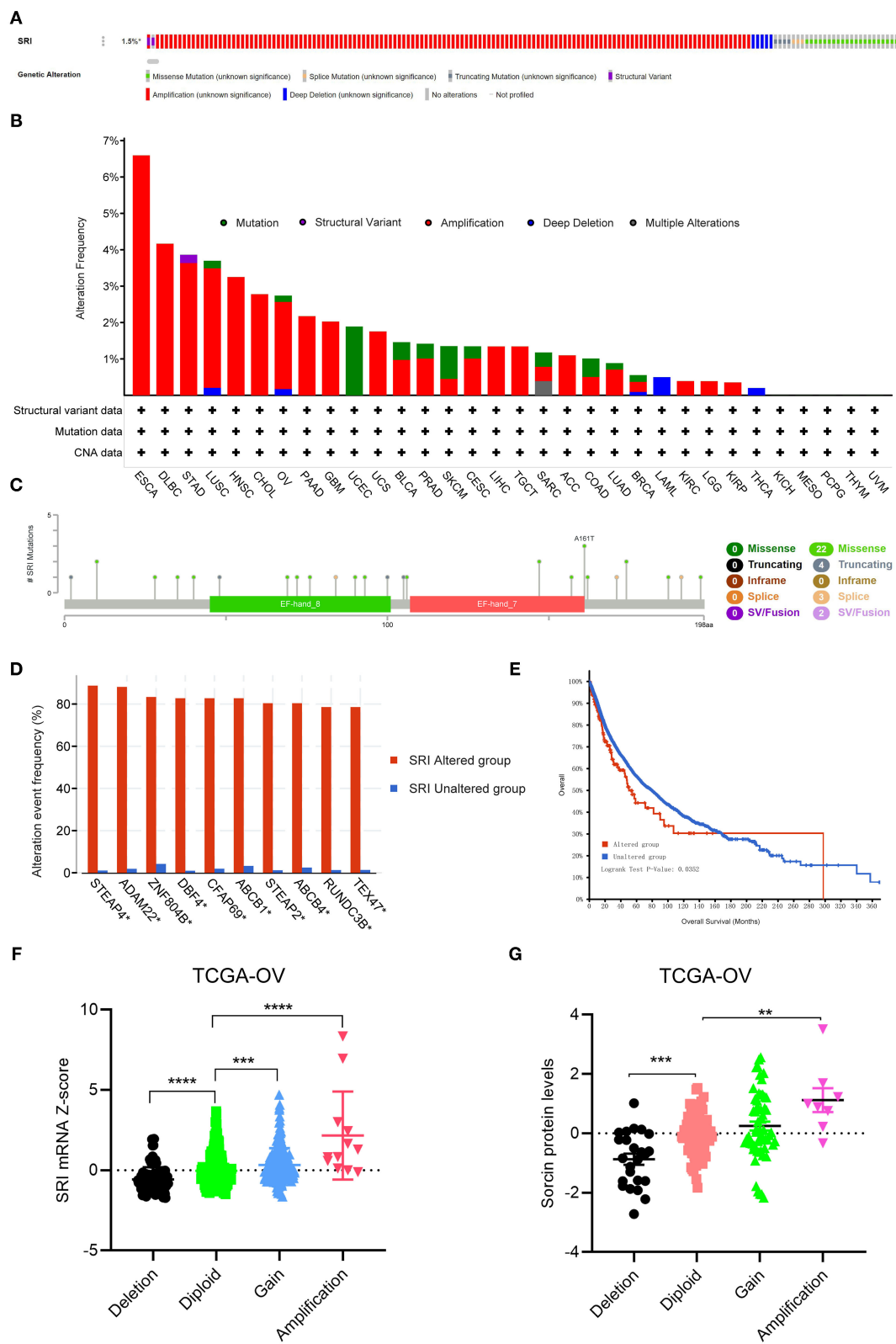
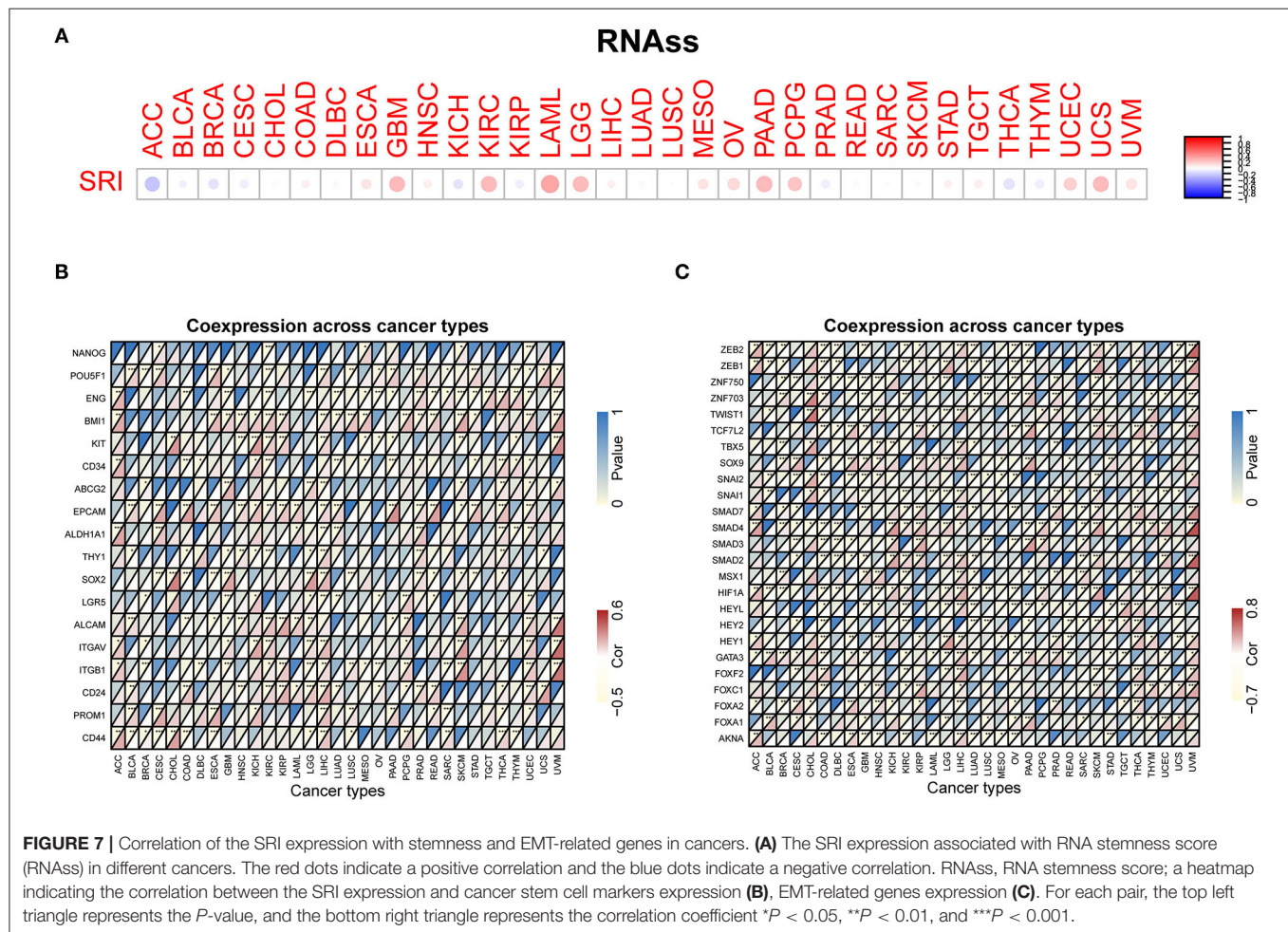


FIGURE 6 | The genetic alterations of SRI in TCGA pan-cancer. **(A)** OncoPrint summary of the alterations on SRI in TCGA pan-cancer datasets. **(B)** Summary of the alteration frequency of SRI derived from structural variant, mutations, and copy-number alterations data in TCGA pan-cancer datasets. **(C)** The mutation types, (Continued)

FIGURE 6 | number, and sites of the SRI genetic alterations. **(D)** The analysis of gene mutation co-occurrence comparing the altered group and unaltered group of SRI. **(E)** Kaplan–Meier overall survival of TCGA pan-cancer cohort with altered or unaltered SRI. **(F)** Association of the SRI copy number alterations with its mRNA expression in the TCGA ovarian cancer cohort. $***P < 0.001$, $****P < 0.0001$, and by one-way ANOVA followed by Tukey's test. **(G)** Association of SRI copy number alterations with its protein expression in the TCGA ovarian cancer cohort. $**P < 0.01$, $***P < 0.001$, and by one-way ANOVA followed by Tukey's test.



SRI in KIRC (Supplementary Figure 3A; $P = 0.011$) and PRAD (Supplementary Figure 3B; $P = 0.012$). No significant difference was observed between the SRI expression and age in other cancers. We further compared the differential mRNA level of SRI in different tumor stages. SRI expression tended to decrease from stage I to stage IV in BLCA (Figure 5A). In contrast, the high SRI expression tended to associate with the advanced tumor stages in LIHC (Figure 5B). The difference between stage I and IV tumors was not statistically significant in LIHC, one reason that might be responsible is the small patient numbers in the advanced stage. Tumor status after treatment was closely associated with disease recurrence. We found that the high level of SRI was significantly correlated with-tumor status in ESCA (Figure 5D), GBM (Figure 5E), KIRC (Figure 5F), LIHC (Figure 5G), PRAD (Figure 5H), and UVM (Figure 5I). While a high expression of SRI was significantly related to the tumor-free status in BLCA (Figure 5C).

Genetic Alteration Analysis of SRI in Pan-Cancer

Next, the cBioPortal database was applied to investigate the genetic alterations of SRI in the TCGA pan-cancer datasets. As shown in Figure 6A, the SRI gene was altered in 168 patients with cancer which accounted for only 1.5% across 10,953 samples. Regarding the SRI alterations in different cancer types (Figure 6B), the most frequent alterations of SRI gene were amplification in ESCA, DLBC, STAD, LUSC, HNSC, CHOL, OV, PAAD, GBM, UCS, BLCA, PRAD, CESC, LIHC, TGCT, ACC, LUAD, BRCA, KIRC, LGG, and KIRP. Patients with UCEC, SKCM harbored High frequency of SRI mutations was observed in UCEC, SKCM, while patients with LAML and THCA harbored high frequency of SRI deep deletion of SRI gene was the most frequent mutation type in LAML and THCA. The mutation types, number, and sites of the SRI genetic alterations are displayed in Figure 6C. The results showed that the missense

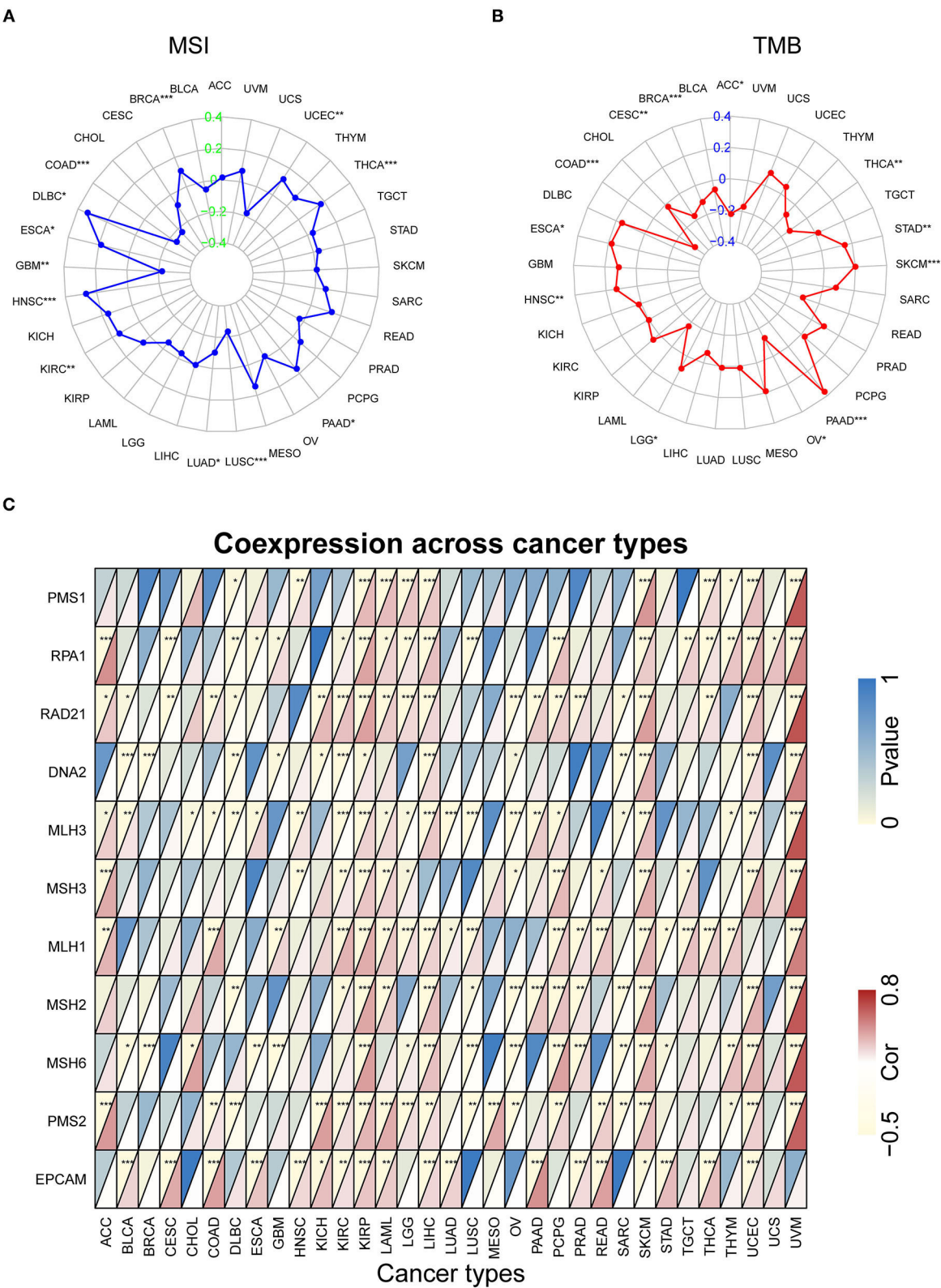


FIGURE 8 | The correlation of SRI expression with microsatellite instability (MSI), tumor mutational burden (TMB), and mismatch repair (MMR) genes. **(A)** Radar map of correlation between the SRI expression and MSI. **(B)** Radar map of correlation between the SRI expression and TMB. **(C)** A heatmap indicating the correlation between the SRI expression and MMR genes. For each pair, the top left triangle indicates the *P*-value, and the bottom right triangle indicates the correlation coefficient **P* < 0.05, ***P* < 0.01, and ****P* < 0.001.

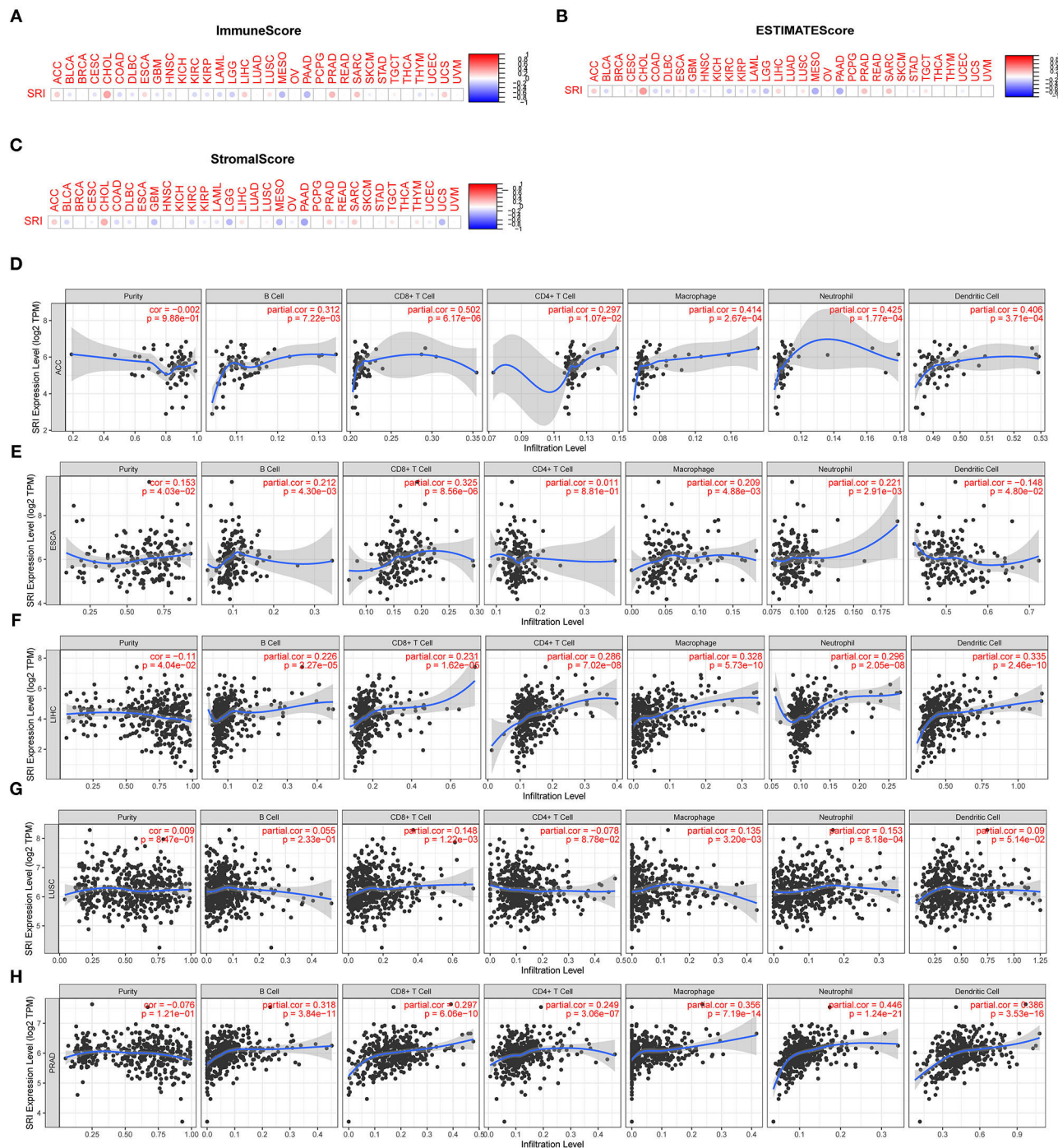
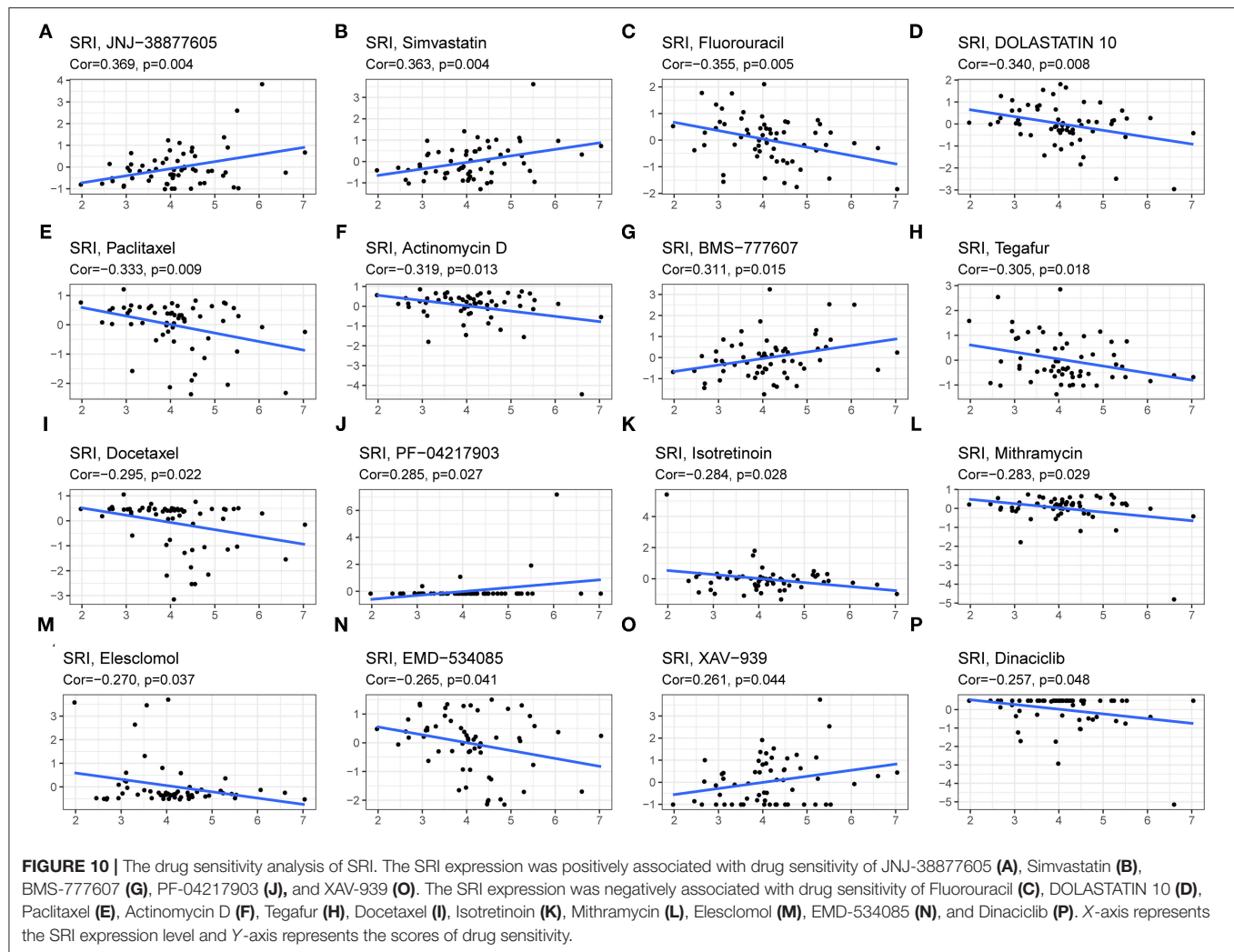


FIGURE 9 | Correlation between the SRI gene expression and tumor immune microenvironment in TCGA. **(A–C)** The SRI expression associated with ImmuneScore, ESTIMATEScore, and StromaScore in different cancers. The red dots indicate a positive correlation and the blue dots indicate a negative correlation. **(D–H)** Correlation of the SRI expression with immune infiltration level in adrenocortical carcinoma (ACC), ESCA, LIHC, lung squamous cell carcinoma (LUSC), and PRAD in TIMER database.

mutations were the major mutation type of SRI. Intriguingly, co-occurrence of STEAP4, ADAM22, ZNF804B, DBF4, CFAP69, ABCB1, STEAP2, ABCB4, RUNDC3B, and TEX47 alterations was observed with the SRI alterations (**Figure 6D**). While the

patients with SRI alterations represented only a small part, the patients in the SRI genetic altered group had poorer OS than those in the SRI unaltered group (**Figure 6E**). To explore whether the SRI amplification had an influence on its mRNA and protein



level, the copy number alterations data and expression data of SRI were acquired in TCGA ovarian cancer. The results indicated that SRI was amplified along with the significantly high mRNA and protein level in TCGA-OV cohort (Figures 6F,G). Regarding the co-expression genes of SRI, the co-expression analysis was conducted using the Linkedomics database in the TCGA ovarian cancer dataset. The results are presented in Supplementary Table 1. The top 50 positive and negative co-expression genes of SRI were visualized in the STRING database (Supplementary Figure 4).

Correlation Among the SRI Expression and Stemness Score, EMT-Related Genes, MSI, TMB, and MMR-Related Genes in Pan-Cancer

Our previous work revealed that SRI promoted the stemness and EMT process in ovarian cancer (35), we, therefore, wanted to investigate the association of SRI expression with stemness score and EMT-related genes in cancers. First, the sorcin

expression was upregulated in the OVCAR-3 and SKOV-3 spheres cells compared with the respective adherent cells (Supplementary Figure 5). Correlation analysis indicated that the SRI expression was positively associated with the RNAss in GBM, KIRC, LAML, LGG, OV, PADD, pheochromocytoma and paraganglioma (PCPG), UCEC, and UCS. While a negative relationship between the SRI level and RNAss was observed in ACC (Figure 7A). Figure 7B depicts the correlations between the SRI expression and 18 tumor stem cell markers. More than 10 cancer stem cell markers expression were significantly positively correlated with the SRI levels in GBM, KIRC, LGG, LIHC, SKCM, THCA, and UCEC. Furthermore, a heatmap showed a positive correlation between the SRI expression and 25 EMT-related genes (Figure 7C). The expression of over 15 EMT-related genes was positively associated with the SRI expression in COAD, GBM, KIRC, LIHC, and OV. The correlations among the TMB, MSI, and SRI expression were further analyzed in cancers. The results demonstrated that the SRI expression was significantly associated with increased MSI in BRCA, DLBC, ESCA, HNSC, KIRC, PAAD, THCA, and UCEC, while the SRI expression was negatively associated with MSI in COAD, GBM, LUAD, and

LUSC (**Figure 8A**). In addition, SRI was positively correlated with TMB in ESCA, HNSC, LGG, PAAD, SKCM, and STAD, while a negative correlation between the SRI expression and TMB was found in ACC, BRCA, CESC, COAD, OV, and THCA (**Figure 8B**). As shown in **Figure 8C**, the SRI expression was significantly positively correlated with the MMR-related genes level in most of the tumors, especially in KIRP, LIHC, SKCM, UCEC, and UVM.

Correlation of the SRI Expression With Tumor Immune Microenvironment

Presently, the predictive role of SRI in tumor immune microenvironment has received little attention. Here, the association of SRI expression with the tumor immune microenvironment was evaluated according to the ESTIMATE algorithm and TIMER database. Our findings showed that the SRI expression had a positive correlation with the immune scores and estimate scores in ACC, ESCA, CHOL, LIHC, PRAD, and SARC (sarcoma). A strong negative correlation between the SRI and immune scores and estimate scores was found in LGG, MESO, and PAAD (**Figures 9A,B**). In addition, the correlation analyses revealed that the stromal scores were negatively correlated with the SRI expression in GBM, LGG, MESO, PADD, and UCS (**Figure 9C**). Then, the relationship between the SRI expression and immune cells infiltration was investigated in pan-cancer. The results indicated that the SRI expression was significantly associated with tumor purity in five cancer types. Furthermore, the SRI expression was significantly correlated with the infiltration levels of B cells in 14 cancer types, CD8+T cells in 14 cancer types, CD4+T cells in 13 cancer types, the macrophages in 21 cancer types, the neutrophils in 16 cancer types, and the dendritic cells in 12 cancer types (**Figures 9D–H; Supplementary Figures 6, 7**). Combined results from the correlation analysis and TIMER database, the SRI expression was strongly correlated with the immune infiltrating level in ACC, ESCA, LIHC, LUSC, and PRAD. The correlation between the immune checkpoint genes expression and SRI expression was explored. The results demonstrated that the most immune checkpoint genes were positively correlated with the SRI expression, especially in LIHC, PRAD, UVM, and ACC (**Supplementary Figure 8**), which suggested that the high level of SRI might mediate immune escape.

Drug Sensitivity Analysis of SRI

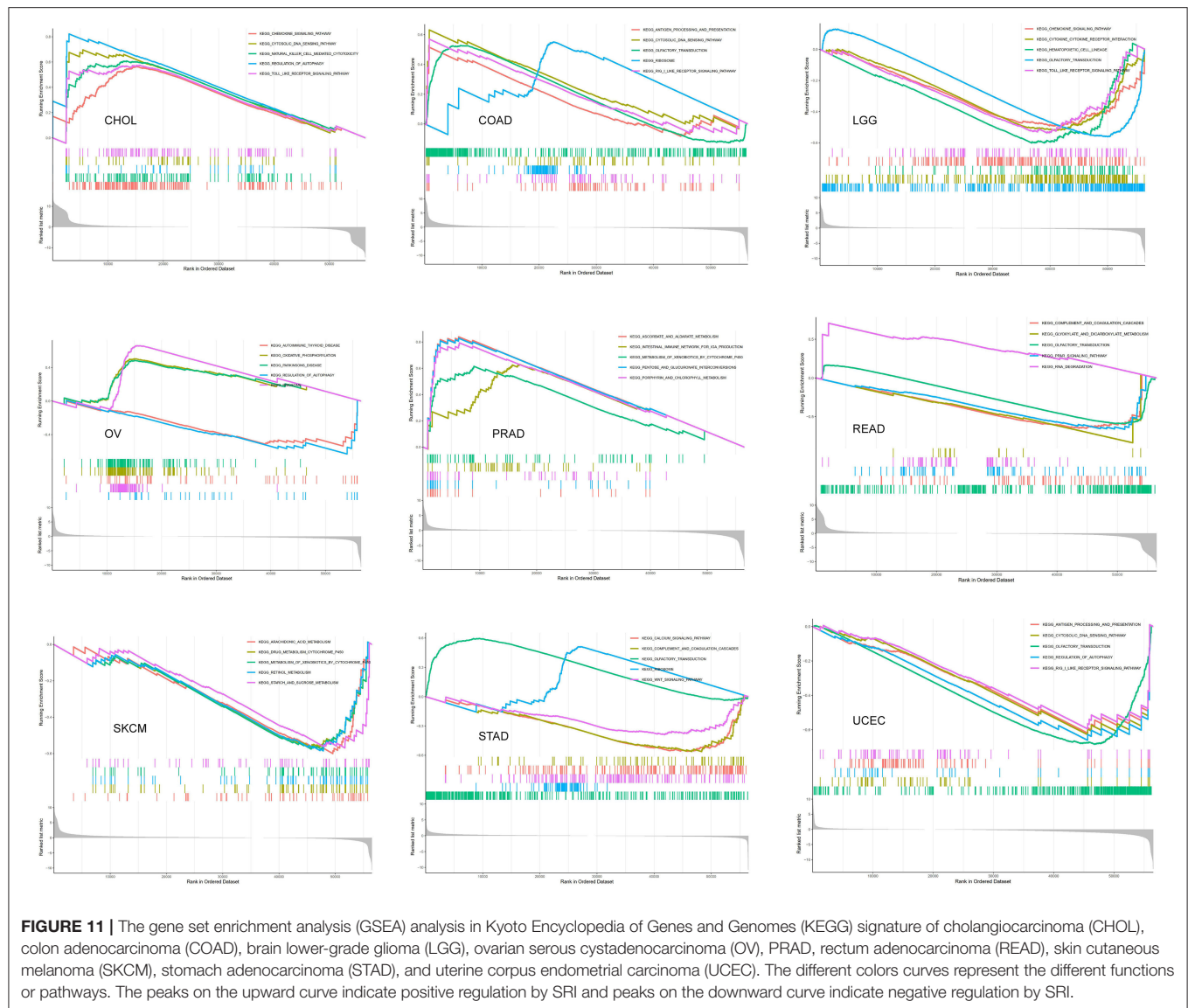
Since the drug resistance role of SRI in cancer has been gradually valued, we further investigated the potential correlation analysis between the drug sensitivity and SRI expression using the CellMiner™ database. Our results indicated that the SRI expression was positively related to JNJ-38877605, Simvastatin, BMS-777607, PF-04217903, and XAV-939 sensitivity (**Figures 10A,B,G,J,O**). Notably, the SRI expression was negatively correlated with the drug sensitivity of fluorouracil, dolastatin 10, paclitaxel, actinomycin D, tegafur, docetaxel, isotretinoin, mithramycin, elesclomol, EMD-534085, and dinaciclib (**Figures 10C–F,H,I,K–N,P**).

Biological Function of SRI in Cancer

The GSEA was then performed to explore the main biological process affected by SRI in cancer. In the KEGG pathway gene set analysis, our results suggested that SRI positively regulated the immune-related pathways in CHOL, COAD, PRAD, such as natural cell-mediated cytotoxicity, antigen processing, and presentation. SRI was positively enriched in the metabolism-related pathways of OV and PRAD, such as oxidative phosphorylation, ascorbate and aldarate metabolism, and chlorophyll metabolism. In contrast, SRI was negatively enriched in the glyoxylate and dicarboxylate metabolism and drug metabolism cytochrome P450 in READ and SKCM. In addition, SRI was identified as a negative regulator for olfactory transduction, autophagy, WNT signaling pathway in LGG, READ, OV, UCEC, and STAD (**Figure 11**). The GO results of GSEA analysis of SRI in pan-cancer are displayed in **Supplementary Figures 9, 10**. In ACC, CESC, HNSC, LIHC, LAML, and SKCM, the SRI expression showed positive enrichment in the gene regulatory mechanisms, such as gene silencing, alternative mRNA splicing, transcription activator activity, and methylation CPG binding. Several immune response pathways, such as immune response regulating cell surface receptor signaling, regulation of immune effector process, and response to interleukin 12 were positively correlated to the SRI expression in STAD, CHOL, and SARC. Furthermore, a negative enrichment among the cell cycle G1/S transition, regulation of epidermal cell growth, and SRI expression was observed in OV, LUSC, PRAD, UCS, PCPG, READ, and ESCA.

DISCUSSION

In the present study, high expression of SRI in multiple cancers was observed and its dysregulation could predict worse prognosis in the patients with cancer, which indicated that SRI could serve as a robust prognostic factor among a variety of cancers. SRI exerts its function *via* the regulation of stemness, MSI, TMB, tumor immune microenvironment, and drug resistance. Sorcin, one of the most abundant calcium-binding proteins, plays an essential role in excitable cells, such as neurons and cardiomyocytes (36). Regarding the SRI expression in normal tissues, our research revealed that the small intestine, brain tissues, and bone marrow had a high expression abundance of SRI, which was consistent with the reports of PaxDb database (<https://pax-db.org/>) (18). Cytosolic calcium (Ca^{2+}), one of the most fascinating cell signaling, organizes the diverse physiological activities from cell division, cellular motility to cell death (37). As an overexpressed Ca^{2+} binding protein, sorcin exquisitely controls the intracellular calcium content and exchange under normal physiological conditions. Multiple mechanisms are reported in the regulation of calcium balance affected by sorcin, some of which depend on the interaction with other calcium-related proteins (38). For a long time, substantial evidence suggests that Ca^{2+} signaling is implicated in the cancer cells' uncontrolled proliferation, angiogenesis, immune surveillance, and drug resistance (39). The previous studies have reported overexpression of SRI in the cell lines



of gastric cancer, colorectal cancer, and breast cancer (13–15). Meanwhile, the high expression level of SRI was observed in the resistance cells of ovarian cancer, myeloma, lung cancer, leukemia, and nasopharyngeal carcinoma (20, 40–44). In this study, the high expression of SRI was verified in 25 tumor types compared with the normal tissues. Until now, literature on the prognostic value of SRI in patients with cancer is scarce. The SRI overexpression was reported to be closely related to the poor clinical outcomes and the complete remission rate in acute myeloid leukemia (45). Our Kaplan–Meier survival and the Cox regression analyses first demonstrated that SRI had predictive value on the survival outcomes in BLCA, BRCA, CESC, CHOL, HNSC, KIRC, KIRC, LGG, LIHC, OV, PAAD, PRAD, SKCM, STAD, THYM, THCA, UCEC, and UVM. In addition, the SRI expression was negatively associated with the disease stage in BLCA but positively correlated to the tumor stage in HNSC, LIHC, and MESO. Sorcin was previously reported to be positively

related with the TNM stage in gastric cancer (46). Further, our study revealed that the high expression of SRI was significantly related to the with-tumor status in ESCA, GBM, KIRC, LIHC, PRAD, and UVM, suggesting SRI might have the potential for reflecting the tumor status.

Regarding the SRI genomic alterations in cancer, our data found that amplification was the most frequent mutation type in cancer. As reported in many studies, genomic amplification of the chromosomal region 7q21, such as ABCB1 and SRI occurred in the multidrug-resistant cancers (47–51). In treating with the chemotherapeutic drugs, amplification of the ABCB1-related amplicon region containing SRI was sufficient to drive tumor chemoresistance. Genomic amplification of SRI has long been recognized as an occasional event of such genomic co-amplification (52). However, accumulating evidence indicated that the pathways involved in the tumor malignant behaviors, such as TGF- β and JAK-STAT3 signaling were affected by the SRI

gene amplification (35, 53). In addition, we found co-occurrence of ADAM22, DBF4, ABCB1, ABCB4, and RUNDC3B alterations was observed with the SRI alterations, which was consistent with the previous reports (48). The full-length sorcin could lead to a low level of paclitaxel resistance in ovarian cancer cells (54). In this study, our analysis showed that the amplification was associated with the high mRNA and protein level of SRI in ovarian cancer, which indicated the sorcin overexpression in ovarian cancer might be due in part to the SRI genomic amplification. A survival analysis further showed that the patients with cancer with the SRI genetic alterations had poorer OS than those with SRI unaltered group. The above findings suggested that the SRI genomic alterations were considered as a risk factor for prognosis in cancer.

The cancer stem cells (CSCs), a pool of specialized cancer cells, are at the root of tumor initiation and responsible for chemoresistance of malignant tumors (55). The occurrence of EMT endows cancer cells with the mesenchymal phenotypes and stem cell-like characteristics and thus confers invasiveness and chemoresistance (56). Sorcin silencing in the breast cancer cells decreases the pool of CD44+/CD24- and ALDH1 high CSCs *in vitro* (13). Sorcin was also reported to facilitate the migration, invasion, and EMT in breast cancer, ovarian cancer, and colorectal cancer (13, 14, 35). Sorcin overexpression is tightly associated with the increased local invasion and lymph node metastasis of gastric cancer (46). Mechanistically, sorcin silencing effectively decreased the expression of matrix metalloproteinases 2 and 9 (MMP2 and MMP9), and eventually suppressed gastric cancer metastasis (53). The previous studies have demonstrated that sorcin induced EMT-related phenotype by the regulation of PI3K/Akt/mTOR pathway and vascular endothelial growth factor (VEGF) (13, 14). Our analysis of on cancer stemness and EMT-related genes further supported the oncogenic and stemness-related role of SRI in cancer. MSI and TMB have recently captured widespread attention as promising predictive biomarkers for immunotherapy efficacy, especially in colorectal cancers and lung cancer (57, 58). Our results demonstrated that the SRI expression was significantly related to MSI in 12 cancer types and TMB in 13 cancer types. These results strongly implied that the SRI expression might affect the response to immune checkpoint therapy in patients with cancer, which will shed new light on the prognosis of immunotherapy. Further analysis revealed that the SRI expression was positively correlated with the MMR-related genes expression in most of the tumors. Thus, the patients with cancer with low expression of SRI, high MSI, and TMB may benefit from immunotherapy.

At present, the role of SRI in the tumor immune microenvironment remains a research gap worth investigation in further research. According to ESTIMATE algorithm, the correlation between the SRI expression and immune cell content might depend on the tumor types. TIMER database mining further found that the SRI expression was significantly correlated with the infiltration levels of various immune cells, particularly in ACC, ESCA, LIHC, LUSC, and PRAD. In addition, the correlation analysis demonstrated that the immune checkpoint genes were positively correlated with the SRI expression

in most tumor types, suggesting SRI might be involved in immune escape. Further *in vitro* and *in vivo* studies exploring the relationship between the SRI expression and immune infiltrations are warranted. Currently, the SRI roles in multi-drug resistance have become increasingly appreciated. The previous studies reported sorcin knockdown resulted in the increased cisplatin, paclitaxel, doxorubicin, fluorouracil, and vincristine sensitivity in the cancer cells (21, 54, 59–62), which was also consistent with our drug sensitivity of SRI. Nevertheless, more experimental validation needs to be further studied to evaluate the influence of SRI on new drugs in clinical trials. Our findings may be useful in prioritizing further research on drug screening. Furthermore, our GSEA analyses suggested that SRI was closely associated with the metabolism-related pathways, transcription activator activity, immune-related pathways, and cell cycle G1/S transition. Relevant literature has reported the SRI affected glucose metabolism (63), STAT3 transcriptional activity (53), cell cycle progression in mitosis (64), and immuno-inflammatory responses (65). The results of GSEA further indicated that SRI was involved in immune-related pathways in certain types of tumor, which was consistent with our previous immune-related analysis on SRI. Our findings revealed that a signaling pathway was significantly enriched in various types of tumors. We, therefore, considered that the signaling pathways mediated by SRI were not specific for different cancers.

The pre-clinical studies have found that the natural compounds, such as dihydromyricetin and triptolide specifically reversed drug resistance through the downregulation of SRI expression *in vitro*, which indicates the clinical transfer value of SRI as a good candidate to prevent chemoresistance (22, 44). The significance of our work is that the multifaceted functions of SRI were unveiled in cancer, which not only further verified the previous findings but also advanced our understanding of the role mediated by SRI in the tumor immune microenvironment. Since our study was a comprehensive bioinformatics analysis and relied on multiple databases, several limitations are inevitable. First, the results are not experimental, and thus future experimental validations are required. Second, the majority of our analyses were focused on the mRNA expression of SRI. It is worth mentioning that the analyses based on the protein levels of SRI would make the results more convincing. Third, this work solely presented the correlation analysis, and the molecular mechanisms of SRI in tumor stemness and immune infiltration require further investigation in the future. To sum up, our pan-cancer analyses systematically probe the characteristics of SRI in multiple aspects, such as expression pattern, survival prognosis, genetic mutation, stemness, TMB, MSI, tumor immune microenvironment, and drug resistance. SRI might be a potential target for cancer therapy since it displayed abnormal high expression in multiple cancers and predicted worse prognosis in patients with cancer. Frequent amplification of SRI genomic was observed and the SRI expression correlated with amplification. Moreover, the aberrant SRI expression was related to the stemness score, EMT-related genes, MSI, TMB, and tumor immune microenvironment across various types of cancer. SRI was able to predict the sensitivity to chemotherapeutic agents as a novel resistance gene. The present study may provide insights for

illustrating the role of SRI in tumorigenesis and drug resistance. At the same time, our work also points to several directions for future prospective studies focusing on SRI in cancer.

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

JZ performed most of the analyses and wrote the manuscript. XH and JZ conceived the idea and led the project. JC, BS, LL,

JD, QS, and QZ performed some of the analyses and edited the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by grants from the Key Science and Technology Program of Anhui province, China (1401042007).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.752619/full#supplementary-material>

REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin.* (2021) 71:7–33. doi: 10.3322/caac.21654
2. Bergholz JS, Wang Q, Kabraji S, Zhao JJ. Integrating immunotherapy and targeted therapy in cancer treatment: mechanistic insights and clinical implications. *Clin Cancer Res.* (2020) 26:5557–66. doi: 10.1158/1078-0432.CCR-19-2300
3. Meyers MB, Schneider KA, Spengler BA, Chang TD, Biedler JL. Sorcin (V19), a soluble acidic calcium-binding protein overproduced in multidrug-resistant cells. Identification of the protein by anti-sorcin antibody. *Biochem Pharmacol.* (1987) 36:2373–80. doi: 10.1016/0006-2952(87)90606-X
4. Ilari A, Johnson KA, Nastopoulos V, Verzili D, Zamparelli C, Colotti G. The crystal structure of the sorcin calcium binding domain provides a model of Ca²⁺-dependent processes in the full-length protein. *J Mol Biol.* (2002) 317:447–58. doi: 10.1006/jmbi.2002.5417
5. Zheng BB, Zhang P, Jia WW, Yu LG, Guo XL. Sorcin, a potential therapeutic target for reversing multidrug resistance in cancer. *J Physiol Biochem.* (2012) 68:281–7. doi: 10.1007/s13105-011-0140-0
6. Matsumoto T, Hisamatsu Y, Ohkusa T, Inoue N, Sato T, Suzuki S. Sorcin interacts with sarcoplasmic reticulum Ca(2+)-ATPase and modulates excitation-contraction coupling in the heart. *Basic Res Cardiol.* (2005) 100:250–62. doi: 10.1007/s00395-005-0518-7
7. Fowler MR, Colotti G, Chiancone E, Smith GL, Fearon IM. Sorcin modulates cardiac L-type Ca²⁺ current by functional interaction with the α_1C subunit in rabbits. *Exp Physiol.* (2008) 93:1233–8. doi: 10.1113/expphysiol.2008.043497
8. Yabuki N, Sakata K, Yamasaki T, Terashima H, Mio T, Miyazaki Y. Gene amplification and expression in lung cancer cells with acquired paclitaxel resistance. *Cancer Genet Cytogenet.* (2006) 173:1–9. doi: 10.1016/j.cancergencyto.07.020
9. Gupta K, Sirohi VK, Kumari S, Shukla V, Manohar M, Popli P. Sorcin is involved during embryo implantation via activating VEGF/PI3K/Akt pathway in mice. *J Mol Endocrinol.* (2018) 60:119–32. doi: 10.1530/JME-17-0153
10. Genovese I, Giamogante F, Barazzuol L, Battista T, Fiorillo A, Vicario M. Sorcin is an early marker of neurodegeneration, Ca(2+) dysregulation and endoplasmic reticulum stress associated to neurodegenerative diseases. *Cell Death Dis.* (2020) 11:861. doi: 10.1038/s41419-020-03063-y
11. He Q, Zhang G, Hou D, Leng A, Xu M, Peng J. Overexpression of sorcin results in multidrug resistance in gastric cancer cells with up-regulation of P-gp. *Oncol Rep.* (2011) 25:237–43. doi: 10.3892/or_0001066
12. Gong Z, Sun P, Chu H, Zhu H, Sun D, Chen J. Overexpression of sorcin in multidrug-resistant human breast cancer. *Oncol Lett.* (2014) 8:2393–8. doi: 10.3892/ol.2014.2543
13. Hu Y, Li S, Yang M, Yan C, Fan D, Zhou Y. Sorcin silencing inhibits epithelial-to-mesenchymal transition and suppresses breast cancer metastasis in vivo. *Breast Cancer Res Treat.* (2014) 143:287–99. doi: 10.1007/s10549-013-2809-2
14. Tong W, Sun D, Wang Q, Suo J. Sorcin enhances metastasis and promotes epithelial-to-mesenchymal transition of colorectal cancer. *Cell Biochem Biophys.* (2015) 72:453–9. doi: 10.1007/s12013-014-0486-3
15. Deng LM, Tan T, Zhang TY, Xiao XF, Gu H. miR-1 reverses multidrug resistance in gastric cancer cells via downregulation of sorcin through promoting the accumulation of intracellular drugs and apoptosis of cells. *Int J Oncol.* (2019) 55:451–61. doi: 10.3892/ijo.2019.4831
16. Robey RW, Pluchino KM, Hall MD, Fojo AT, Bates SE, Gottesman MM. Revisiting the role of ABC transporters in multidrug-resistant cancer. *Nat Rev Cancer.* (2018) 18:452–64. doi: 10.1038/s41568-018-0005-8
17. Zhang H, Xu H, Ashby CR, Assaraf YG, Chen ZS, Liu HM. Chemical molecular-based approach to overcome multidrug resistance in cancer by targeting P-glycoprotein (P-gp). *Med Res Rev.* (2021) 41:525–555. doi: 10.1002/med.21739
18. Battista T, Fiorillo A, Chiarini V, Genovese I, Ilari A, Colotti G. Roles of sorcin in drug resistance in cancer: one protein, many mechanisms, for a novel potential anticancer drug target. *Cancers (Basel).* (2020) 12:887. doi: 10.3390/cancers12040887
19. Van der Bliek AM, Meyers MB, Biedler JL, Hes E, Borst P. A 22-kd protein (sorcin/V19) encoded by an amplified gene in multidrug-resistant cells, is homologous to the calcium-binding light chain of calpain. *Embo J.* (1986) 5:3201–8. doi: 10.1002/j.1460-1986.tb04630.x
20. Gao Y, Li W, Liu X, Gao F, Zhao X. Reversing effect and mechanism of soluble resistance-related calcium-binding protein on multidrug resistance in human lung cancer A549/DDP cells. *Mol Med Rep.* (2015) 11:2118–24. doi: 10.3892/mmr.2014.2936
21. Genovese I, Fiorillo A, Ilari A, Masciarelli S, Fazi F, Colotti G. Binding of doxorubicin to Sorcin impairs cell death and increases drug resistance in cancer cells. *Cell Death Dis.* (2017) 8:e2950. doi: 10.1038/cddis.2017.342
22. Sun Y, Wang C, Meng Q, Liu Z, Huo X, Sun P. Targeting P-glycoprotein and SORCIN: dihydromyricetin strengthens anti-proliferative efficiency of adriamycin via MAPK/ERK and Ca(2+)-mediated apoptosis pathways in MCF-7/ADR and K562/ADR. *J Cell Physiol.* (2018) 233:3066–79. doi: 10.1002/jcp.26087
23. Zhang J, Guan W, Xu X, Wang F, Li X, Xu G. A novel homeostatic loop of sorcin drives paclitaxel-resistance and malignant progression via Smad4/ZEB1/miR-142-5p in human ovarian cancer. *Oncogene.* (2021) 40:4906–18. doi: 10.1038/s41388-021-01891-6
24. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Varambally R, Yu J, Briggs BB. OncoPrint 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia.* (2007) 9:166–80. doi: 10.1593/neo.07112
25. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* (2013) 6:l1. doi: 10.1126/scisignal.2004088

26. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*. (2018) 173:338–54.e315. doi: 10.1016/j.cell.03.034
27. Forde PM, Chaft JE, Smith KN, Anagnostou V, Cottrell TR, Hellmann MD. Neoadjuvant PD-1 blockade in resectable lung cancer. *N Engl J Med*. (2018) 378:1976–86. doi: 10.1056/NEJMoa1716078
28. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nat Rev Cancer*. (2019) 19:133–50. doi: 10.1038/s41568-019-0116-x
29. Bonneville R, Krook MA, Kautto EA, Miya J, Wing MR, Chen HZ. Landscape of microsatellite instability across 39 cancer types. *JCO Precis Oncol*. (2017) 1:1–5. doi: 10.1200/PO.17.00073
30. Yu Y, Carey M, Pollett W, Green J, Dicks E, Parfrey P. The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Med*. (2019) 17:150. doi: 10.1186/s12916-019-1379-5
31. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-García W. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. (2013) 4:2612. doi: 10.1038/ncomms3612
32. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res*. (2017) 77:e108–10. doi: 10.1158/0008-5472.CAN-17-0307
33. Shankavaram UT, Varma S, Kane D, Sunshine M, Chary KK, Reinhold WC. CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*. (2009) 10:277. doi: 10.1186/1471-2164-10-277
34. Zhang J, Wang F, Xu X, Li X, Guan W. Overexpressed COL5A1 is correlated with tumor progression, paclitaxel resistance, and tumor-infiltrating immune cells in ovarian cancer. *J Cell Physiol*. (2021). 236:6907–19. doi: 10.1002/jcp.30350
35. Zhang J, Guan W, Xu X, Wang F, Li X, Xu G. A novel homeostatic loop of sorcin drives paclitaxel-resistance and malignant progression via Smad4/ZEB1/miR-142-5p in human ovarian cancer. *Oncogene*. (2021). 40:4906–18. doi: 10.1038/s41388-021-01891-6
36. Meyers MB, Pickel VM, Sheu SS, Sharma VK, Scotto KW, Fishman GI. Association of sorcin with the cardiac ryanodine receptor. *J Biol Chem*. (1995) 270:26411–8. doi: 10.1074/jbc.270.44.26411
37. Bruce JIE, James AD. Targeting the calcium signalling machinery in cancer. *Cancers (Basel)*. (2020) 12:2351. doi: 10.3390/cancers12092351
38. Colotti G, Poser E, Fiorillo A, Genovese I, Chiarini V, Ilari A. Sorcin, a calcium binding protein involved in the multidrug resistance mechanisms in cancer cells. *Molecules*. (2014) 19:13976–89. doi: 10.3390/molecules190913976
39. Tajada S, Villalobos C. Calcium permeable channels in cancer hallmarks. *Front Pharmacol*. (2020) 11:968. doi: 10.3389/fphar.2020.00968
40. Qi J, Liu N, Zhou Y, Tan Y, Cheng Y, Yang C. Overexpression of sorcin in multidrug resistant human leukemia cells and its role in regulating cell apoptosis. *Biochem Biophys Res Commun*. (2006) 349:303–9. doi: 10.1016/j.bbrc.08.042
41. Liu X, Chen L, Feng B, Liu G. Reversing effect of sorcin in the drug resistance of human nasopharyngeal carcinoma. *Anat Rec (Hoboken)*. (2014) 297:215–21. doi: 10.1002/ar.22832
42. Yamagishi N, Nakao R, Kondo R, Nishitsuji M, Saito Y, Kuga T. Increased expression of sorcin is associated with multidrug resistance in leukemia cells via up-regulation of MDR1 expression through cAMP response element-binding protein. *Biochem Biophys Res Commun*. (2014) 448:430–6. doi: 10.1016/j.bbrc.04.125
43. Xu P, Jiang YF, Wang JH. shRNA-mediated silencing of sorcin increases drug chemosensitivity in myeloma KM3/DDP and U266/ADM cell lines. *Int J Clin Exp Pathol*. (2015) 8:2300–10.
44. Hu H, Zhu S, Tong Y, Huang G, Tan B, Yang L. Antitumor activity of triptolide in SKOV3 cells and SKOV3/DDP in vivo and in vitro. *Anticancer Drugs*. (2020) 31:483–91. doi: 10.1097/CAD.0000000000000894
45. Tan Y, Li G, Zhao C, Wang J, Zhao H, Xue Y. Expression of sorcin predicts poor outcome in acute myeloid leukemia. *Leuk Res*. (2003) 27:125–31. doi: 10.1016/S0145-2126(02)00083-8
46. Deng L, Su T, Leng A, Zhang X, Xu M, Yan L. Upregulation of soluble resistance-related calcium-binding protein (sorcin) in gastric cancer. *Med Oncol*. (2010) 27:1102–8. doi: 10.1007/s12032-009-9342-5
47. Chao CC, Ma CM, Lin-Chao S. Co-amplification and over-expression of two mdr genes in a multidrug-resistant human colon carcinoma cell line. *FEBS Lett*. (1991) 291:214–8. doi: 10.1016/0014-5793(91)81287-I
48. Flahaut M, Mühlethaler-Mottet A, Martinet D, Fattet S, Bourlond KB, Auderset K. Molecular cytogenetic characterization of doxorubicin-resistant neuroblastoma cell lines: evidence that acquired multidrug resistance results from a unique large amplification of the 7q21 region. *Genes Chromosomes Cancer*. (2006) 45:495–508. doi: 10.1002/gcc.20312
49. Kitada K, Yamasaki T. The MDR1/ABCB1 regional amplification in large inverted repeats with asymmetric sequences and microhomologies at the junction sites. *Cancer Genet Cytogenet*. (2007) 178:120–7. doi: 10.1016/j.cancergencyto.06.014
50. Patch AM, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S. Whole-genome characterization of chemoresistant ovarian cancer. *Nature*. (2015) 521:489–94.
51. Steuer CE, Ramalingam SS. Tumor mutation burden: leading immunotherapy to the era of precision medicine? *J Clin Oncol*. (2017) 36:631–2. doi: 10.1200/JCO.76.8770
52. Van der Bliek AM, Baas F, Van der Velde-Koerts T, Biedler JL, Meyers MB, Ozols RF. Genes amplified and overexpressed in human multidrug-resistant cell lines. *Cancer Res*. (1988) 48:5927–32.
53. Tuo H, Shu F, She S, Yang M, Zou XQ, Huang J. Sorcin induces gastric cancer cell migration and invasion contributing to STAT3 activation. *Oncotarget*. (2017) 8:104258–71. doi: 10.18632/oncotarget.22208
54. Parekh HK, Deng HB, Choudhary K, Houser SR, Simpkins H. Overexpression of sorcin, a calcium-binding protein, induces a low level of paclitaxel resistance in human ovarian and breast cancer cells. *Biochem Pharmacol*. (2002) 63:1149–58. doi: 10.1016/S0006-2952(02)00850-X
55. Jones CL, Inguva A, Jordan CT. Targeting energy metabolism in cancer stem cells: progress and challenges in leukemia and solid tumors. *Cell Stem Cell*. (2021) 28:378–93. doi: 10.1016/j.stem.02.013
56. Rodriguez-Aznar E, Wiesmüller L, Sainz B, Hermann PC. EMT and stemness-key players in pancreatic cancer stem cells. *Cancers (Basel)*. (2019) 11:1136. doi: 10.3390/cancers11081136
57. Margetis P, Antonelou M, Karababa F, Loutradi A, Margaritis L, Papassideri I. Physiologically important secondary modifications of red cell membrane in hereditary spherocytosis-evidence for *in vivo* oxidation and lipid rafts protein variations. *Blood Cells Mol Dis*. (2006) 38:210–20. doi: 10.1016/j.bcmd.10.163
58. André T, Shiu KK, Kim TW, Jensen BV, Jensen LH, Punt C. Pembrolizumab in microsatellite-instability-high advanced colorectal cancer. *N Engl J Med*. (2020) 383:2207–18. doi: 10.1056/NEJMoa2017699
59. Padar S, van Breemen C, Thomas DW, Uchizono JA, Livesey JC, Rahimian R. Differential regulation of calcium homeostasis in adenocarcinoma cell line A549 and its Taxol-resistant subclone. *Br J Pharmacol*. (2004) 142:305–16. doi: 10.1038/sj.bjp.0705755
60. Maddalena F, Laudiero G, Piscazzi A, Secondo A, Scorziello A, Lombardi V. Sorcin induces a drug-resistant phenotype in human colorectal cancer by modulating Ca(2+) homeostasis. *Cancer Res*. (2011) 71:7659–69. doi: 10.1158/0008-5472.CAN-11-2172
61. Qinghong S, Shen G, Lina S, Yueming Z, Xiaou L, Jianlin W. Comparative proteomics analysis of differential proteins in respond to doxorubicin resistance in myelogenous leukemia cell lines. *Proteome Sci*. (2015) 13:1. doi: 10.1186/s12953-014-0057-y
62. Wang D, Shi S, Hsieh YL, Wang J, Wang H, Wang W. Knockdown of sorcin increases HEI-OC1 cell damage induced by cisplatin *in vitro*. *Arch Biochem Biophys*. (2021) 701:108752. doi: 10.1016/j.abb.2021.108752
63. Rutti S, Arous C, Schwartz D, Timper K, Sanchez JC, Dermitzakis E. Fractalkine (CX3CL1), a new factor protecting β -cells against TNF α . *Mol Metab*. (2014) 3:731–41. doi: 10.1016/j.molmet.07.007
64. Lalioti VS, Ilari A, O'Connell DJ, Poser E, Sandoval IV, Colotti G. Sorcin links calcium signaling to vesicle trafficking, regulates Polo-like kinase 1 and is necessary for mitosis. *PLoS ONE*. (2014) 9:e85438. doi: 10.1371/journal.pone.0085438

65. Li X, Liu Y, Wang Y, Liu J, Cao H. Negative regulation of hepatic inflammation by the soluble resistance-related calcium-binding protein *via* signal transducer and activator of transcription 3. *Front Immunol.* (2017) 8:709. doi: 10.3389/fimmu.2017.00709

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

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of

the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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



Identification and Structure Prediction of Human Septin-4 as a Biomarker for Diagnosis of Asthenozoospermic Infertile Patients—Critical Finding Toward Personalized Medicine

A. S. Vickram¹, K. Anbarasu², Palanivelu Jeyanthi³, G. Gulothungan⁴, R. Nanmaran⁴, S. Thanigaivel¹, T. B. Sridharan⁵ and Karunakaran Rohini^{6*}

OPEN ACCESS

Edited by:

Balu Kamaraj,
Imam Abdulrahman Bin Faisal
University, Saudi Arabia

Reviewed by:

Manigandan Venkatesan,
The University of Texas Health Science
Center at San Antonio, United States
Abdullahi Hassan Ndanusa,
Skyline University Nigeria
(SUN), Nigeria
Saravanan Parameswaran,
Gwangju Institute of Science and
Technology, South Korea
Subbaiya Ramasamy,
Copperbelt University, Zambia

*Correspondence:

Karunakaran Rohini
rohini@aimst.edu.my

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 09 June 2021

Accepted: 25 October 2021

Published: 03 December 2021

Citation:

Vickram AS, Anbarasu K, Jeyanthi P,
Gulothungan G, Nanmaran R,
Thanigaivel S, Sridharan TB and
Rohini K (2021) Identification and
Structure Prediction of Human
Septin-4 as a Biomarker for Diagnosis
of Asthenozoospermic Infertile
Patients—Critical Finding Toward
Personalized Medicine.
Front. Med. 8:723019.
doi: 10.3389/fmed.2021.723019

¹ Department of Biotechnology, Saveetha School of Engineering (SSE), SIMATS, Chennai, India, ² Department of Bioinformatics, Saveetha School of Engineering (SSE), SIMATS, Chennai, India, ³ Department of Biotechnology, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India, ⁴ Department of Biomedical Engineering, Saveetha School of Engineering (SSE), SIMATS, Chennai, India, ⁵ Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology (VIT), Vellore, India, ⁶ Unit of Biochemistry, Faculty of Medicine, AIMST University, Bedong, Malaysia

Semen parameters are been found as a key factor to evaluate the count and morphology in the given semen sample. The deep knowledge of male infertility will unravel with semen parameters correlated with molecular and biochemical parameters. The current research study is to identify the motility associated protein and its structure through the *in-silico* approach. Semen samples were collected and initial analysis including semen parameters was analyzed by using the World Health Organization protocol. Semen biochemical parameters, namely, seminal plasma protein concentration, fructose content, and glucosidase content were calculated and evaluated for correlation. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) were carried out for identification of Septin-4 presence in the semen sample. Mascot search was done for protein conformation and *in-silico* characterization of Septin-4 by structural modeling in Iterative Threading Assembly Refinement (I-TASSER). Twenty-five nanoseconds molecular dynamics (MD) simulations results showed the stable nature of Septin-4 in the dynamic system. Overall, our results showed the presence of motility-associated protein in normospermia and control samples and not in the case of asthenospermia and oligoasthenospermia. Molecular techniques characterized the presence of Septin-4 and as a novel biomarker for infertility diagnosis.

Keywords: human semen, seminal plasma, motility associated protein, septin-4, *in-silico* characterization, molecular dynamics simulations

INTRODUCTION

Human infertility affects <15% of all couples, <6% of Indian couples. Among these, male partner contributes 40–50% of total infertility (1–3). This gave a clear picture of the contribution of males toward human infertility. Semen parameters, namely, spermatozoa concentration, sperm motility, morphology, etc. plays a major role and act as a deciding factor for fertility rate.

So, the andrologists majorly focus on these issues primarily toward the diagnosis of male infertility (4, 5). For analyzing these factors, a multiple-omics approach is in need to diagnosis male infertility by having a strong focus on parameter analysis. Semen parameters were found to be only the primary trump card, with these, we can just tell about the count and morphology, wherein the deepest knowledge of male infertility will come only when semen parameters were correlated with many molecular and biochemical parameters (6). One such approach is proteomics of semen, correlating with motility-associated proteins. Motility is the major parameter analyzed during semen analysis, the cluster of proteins involved in giving mobility to the sperm cells when entered into the female reproductive system (7–11). Many potential biomarkers could be elucidated here (proteomic approach) which strengthens the diagnosing part (12). A biomarker is a marker derived from any biological substances which could be used to study, analyze, and compare various conditions and strategies. Biomarkers were non-invasive, with minimal side effects, and could be used for various diagnostics and therapeutics values (13–17). Currently, the basic andrology laboratory, various semen analysis parameters, endocrine research, and antisperm antibodies where assisting clinicians for diagnosis (18–20).

In addition, the proteomic approach will strengthen the patient-specific diagnosis and prognosis. Already we studied the role and influence of many proteins like Semenogelin II, prostasomes proteins, and epididymal proteins as markers for various diagnostic approaches. Septin is one of the flagellar proteins that produce the energy in the annular region and helps the sperm to move forward in the female reproductive tract (21–23). Septins are the major cytoskeletal protein with major and unique filament-forming capabilities (24, 25). Many mice model studies proved that the downregulated or missing septin family protein in ejaculated semen will fall under sick without mobility and thus producing the immotile sperms will not help further for natural conception (26). So far 14 different septin genes were identified since the first was 35 years back. Disruption of septin and its functions shows many abnormalities to humankind, namely, neoplasia, breast cancer, Parkinson's disease, neurodegenerative disease, and human male infertility (27, 28). Each septin presence is important for other septins to do their functions properly. These septins will bind together to produce a higher order structure, to form a filament, membranes, or ring-like structure (29). The septin-rich part of sperm is the annulus, it is a submembranous ring that separates the middle and peripheral regions of the sperm flagella. The role of septin is still in debate whether it is an active GTPases or just as a guanosine triphosphate binding protein (30, 31). Septin gives much more energy and the ring structure gives the circulatory force that drives the sperm to move forward and not immotile in the female reproductive tract (32, 33).

The functions of septin start at spermatogenesis itself, during this time it helps in establishing the mitochondrial architecture and cytoskeleton to the annulus. The absence of Septin-4 and–12 in the sperm cell, lacking with the functions of mobility, midpiece damage, rounded sperm head, acrosomal defects, etc. (34, 35). Many studies revealed the insufficient energy

for a sperm cell to move forward in the absence of septin proven by *in vitro* and *in vivo* mice models. The absence of septin in sperm cells is shown with lots of annuli and the connection between midpiece and head, this will malfunction the sperm and not able to fuse the ovum as it fails the forward motility (36, 37). The functions of septin in male fertility were more, but still, the mechanism of understating these family proteins was very tough, and correlating with male infertility diagnosis could be elucidated further (38, 39). Due to the lack of experimental structure of human Septin-4, the structural prediction methods using *in-silico* characterization will help in elucidating the structure–function relationship at the molecular level.

MATERIALS AND METHODS

Semen Sample Collection

Semen samples were collected from the patients who visited Bangalore Assisted Conception Center, Bangalore, Karnataka at the Andrology lab. The samples were collected from them in a wide mounted, sterile, non-toxic plastic ware, they have been provided with a neat room to collect the samples. The method followed was 7 days abstinence time and masturbation technique. Strictly the abstinence time was asked with them as it influences the results in a great manner. The patients were provided with all necessary infrastructures for collection as this also influence the results. Once, the collection was over, the patient details, namely, the name, hospital number, andrology number, abstinence time, method of collection, smoking habits, alcohol habits, last visit date, last collection date, age, and region were asked for and observed. The sample container was marked with a patient number, hospital number for further processing (40).

Ethical Consent

Ethical clearance was done for this work to carry on human semen samples. Informed consent was also obtained from the patients in their own language. The patients were explained with the motive of this work and only after semen analysis report were ready, and then the remaining samples were utilized for this work.

Semen Analysis Report Preparation

Soon after the arrival of samples from the patients to andrologists, a semen analysis report was prepared. World Health Organization (41) procedure was strictly followed to prepare the report. Computer-assisted semen analysis, Germany made, was used to compute the number of spermatozoa, motility, morphology, etc. (42).

Categorization of Semen Samples

Semen samples were segregated into groups by prepared semen analysis report; the categories were asthenozoospermia, oligozoospermia, normozoospermia, and healthy volunteers or controls. The segregation was done purely by using semen parameter values and semen analysis reports (42).

Statistical Analysis

We have used Graphpad prism (GraphPad Software, USA), version 5.1 for this research statistical data. Values were mentioned with mean \pm standard error of the mean for experiments repeated (43).

Separation of Spermatozoa and Seminal Plasma for Biochemical and Molecular Analysis

For this research, after semen sample analysis, samples were collected according to the standard protocol followed by WHO and as per Rao et al. (44).

Spermatozoa Disruption for Obtaining the Intracellular Protein Content

Spermatozoa separated from seminal plasma; sperm pellets were suspended which was supplied with buffers with various detergents. The standard protocol is followed for spermatozoa disruption (42).

Protein Estimation

Protein estimation was done on each fraction of seminal plasma and spermatozoa with the standard protocol followed by standard protocol (45).

Fructose Content Estimation

Fructose content in each sample was evaluated with the standard protocol given by WHO (41), with some modifications done (46).

Enzyme α -Glucosidase Estimation

α -Glucosidase estimation in each sample was evaluated with the standard protocol given by WHO (41), with some modifications done (47).

Estimation of Trace Element Zn

Zinc (Zn) plays a major role in human male fertility. Estimation of Zn was done with standard protocol by using atomic absorption spectroscopy and followed standard protocol (48). Trace element concentrations were estimated using the standard curve.

Identification of Septin-4 Protein in Spermatozoa

The centrifuged and ultrasonicated samples were used to identify the fertility-associated protein in spermatozoa (Septin-4); intracellular proteins isolated from different semen samples' categories (asthenozoospermia, oligospermia, normospermia, and control) were subjected to SDS-PAGE analysis. The silver staining protocol was used to stain the gel. To the extend, the protein band which was differentially expressed (downregulated) in the asthenozoospermia category was subjected to matrix-assisted laser desorption/ionization- time of flight- mass spectrometry (MALDI-TOF-MS) analysis and then Mascot search for identification of the protein.

The differentially expressed band from the gel was excised and dehydrated with a minimum of 50% 50 mM ammonium bicarbonate and 50% acetonitrile. Then follows the standard

protocol overnight. Voyager-DE STR instrument (PerSeptive Biosystems, Inc., USA) in linear mode was used to acquire MALDI-TOF-MS spectra. Positive ions accelerated to 20 V were calculated. Both matrix and sample were dissolved in milliQ water and equal ratios of matrix and sample were mixed and spotted onto MALDI plate for analysis.

In-silico Characterization

In addition to the wet-lab experiments, the *in-silico* structural analysis was evaluated for human Septin-4. The primary analysis based on the Swissprot database screen proved Septin-4 consists of 478 amino acids (Uniprot/Swissprot id: O43236). Septin-4 consists of eight isoforms and isoform 1 (identifier: O43236-1) was selected for the analysis consisting of molecular weight 55,098 Daltons (55 KDa). From the structural database screening, the absence of an experimental 3D structure of Septin-4 was identified. The *in-silico* structural modeling of Septin-4 was performed using the Iterative Threading Assembly Refinement (I-Tasser) server (49). Iterative Threading Assembly Refinement is a fully automated 3D structural prediction of protein server based on the threading/fold recognition methodology. It ranked no. 1 among the structural prediction server evaluated by a critical assessment of structure prediction (CASP14 experiment in 2020) and also ranked top for the function prediction (CASP9). The server chooses the suitable structural templates from database protein data bank (PDB) by a multiple-threading approach called local meta-threading server (LOMETS) and protein models constructed by iterative template-based fragment assembly simulations. The prediction is mainly based on critical parameters like C-score, TM score, and root mean square deviation (RMSD). C-score, a scoring function mainly based on the theoretical concepts were also done. C-score with a range of $[-5, 2]$ signifies the higher value confirmed the protein model with the confidence level. The output showed the five best protein models based on optimal C-score, TM-score, RMSD, and SD.

Molecular Dynamics Simulation

Molecular dynamics (MD) simulations study on human Septin-4 was carried out using GROMACS 5.0 package (David van der Spoel, Sweden) (50). Simple point charge (SPC21) water molecules of 0.9 nm were used for the solvation of protein models in the simulation box. The neutralization of the system was obtained by adding six sodium ions to replace the initial SPC water molecule in all directions. Energy minimization of all systems was carried out by steepest descent energy minimization with tolerance limit 100 kJ/mol and GROMOS96 43a1 force field was used for the simulations of protein (51). A cutoff of 14 Å for van der Waals interactions and 12 Å for electrostatic interactions was used for the process. Electrostatic interactions were computed using the particle mesh Ewald method. The LINCS algorithm was used to constrain all bond lengths and the SETTLE algorithm was applied to constrain the geometry of water molecules in the system. The energy minimization was done in two equilibration phases, number of particles, volume, and temperature (NVT) ensemble with a constant temperature of 300 K and with a coupling constant of 0.1 ps for duration 100 ps, and number of particles, pressure, and temperature

(NPT) ensemble with a constant pressure of 1 bar was employed with a coupling constant of 5 ps for duration 100 ps. For both ensembles of equilibration, the coupling scheme of Berendsen was employed. Finally, the systems were subjected to production MD simulation for 25 ns run. MD trajectories of human Septin-4 were analyzed by GROMACS utilities. The analysis included RMSD, solvent accessible surface, the radius of gyration (Rg), and principal component analysis (PCA). The stability analysis was performed by using utilities like *g_rms*, *g_sas*, *g_gyrate*, *g_covar*, and *g_anaeig*, respectively. Principal component analysis describes a correlated motion of the protein obtained from the mass-weighted $C\alpha$ -covariance matrix. The functionally relevant motion of the protein can be computed by the collective displacement of domains called essential dynamics. To detect the collective motion mutant trajectories were subjected to PCA. The resulting covariance matrix describes the concerted coordinate motions

In this study, the first and second Cartesian principal components are considered reaction coordinates derived from PCA.

RESULTS

The first step was to categorize the semen samples based on the World health organization values, this was done by using several semen samples, and each value and its error mean was the final mark. Based on the semen analysis report oligospermia ($N = 18$) meant for less count than normal, asthenospermia (less motility $N = 24$) than normal, normospermia (normal as per WHO $N = 15$), oligoasthenospermia (both less count and motility $N = 12$), and healthy volunteer (control $N = 8$). The semen parameter values were tabulated in **Supplementary Table 1**. The results suggested that there exists a potential statistical difference exist between oligospermia and asthenospermia in the case of motility parameter. As this work will further correlate only the motility issues, the results we majorly focused on only motility issues.

Once the semen analysis report and categorization of samples were done, immediately the samples were kept in liquid nitrogen preservation. Once the need, the samples were centrifuged for separation of seminal plasma and spermatozoa. Important biochemical parameters were analyzed. The total protein content was done for both seminal plasma and spermatozoa, fructose content was estimated in seminal plasma, α -glucosidase estimation was also done in seminal plasma for all samples in all categories, and Zn content was evaluated in the same way. All these are very essential biochemical parameters that need to be evaluated for proper correlation with molecular markers during diagnosis. All these biochemical values for different categories of semen samples were tabulated in **Supplementary Table 2**.

Protein content was already evaluated through Lowry's method. After centrifugation, sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) was done for different infertile categories as mentioned earlier in the methodology section. The developed silver-stained protein SDS-PAGE was depicted in **Figure 1**. Almost eight bands were

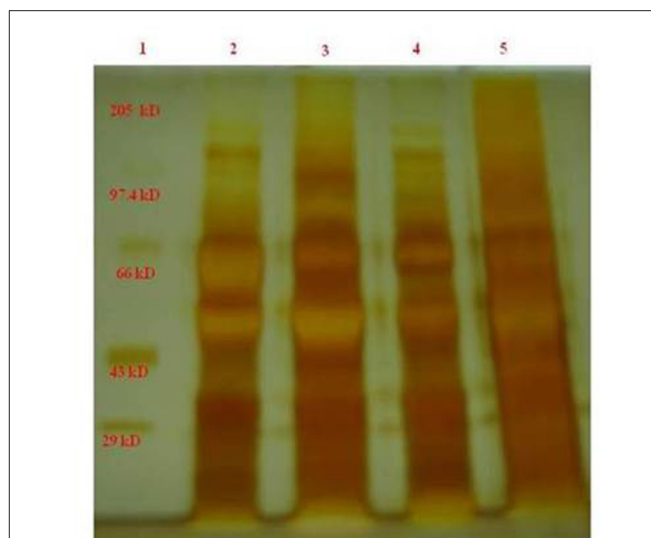


FIGURE 1 | The band around 55 kDa was less expressed in the case of asthenospermia, but present in the case of oligospermia, normospermia, and healthy volunteers. We guessed the importance of missed 55-kDa protein and further we want to investigate on this protein. GelAnalyzer was used to analyze this 1D SDS PAGE bands and all the interpretation has been done by the standard protocol. 1, marker standard; 2, Normospermia; 3, healthy volunteer; 4, Asthenospermia; 5, Oligoasthenospermia.

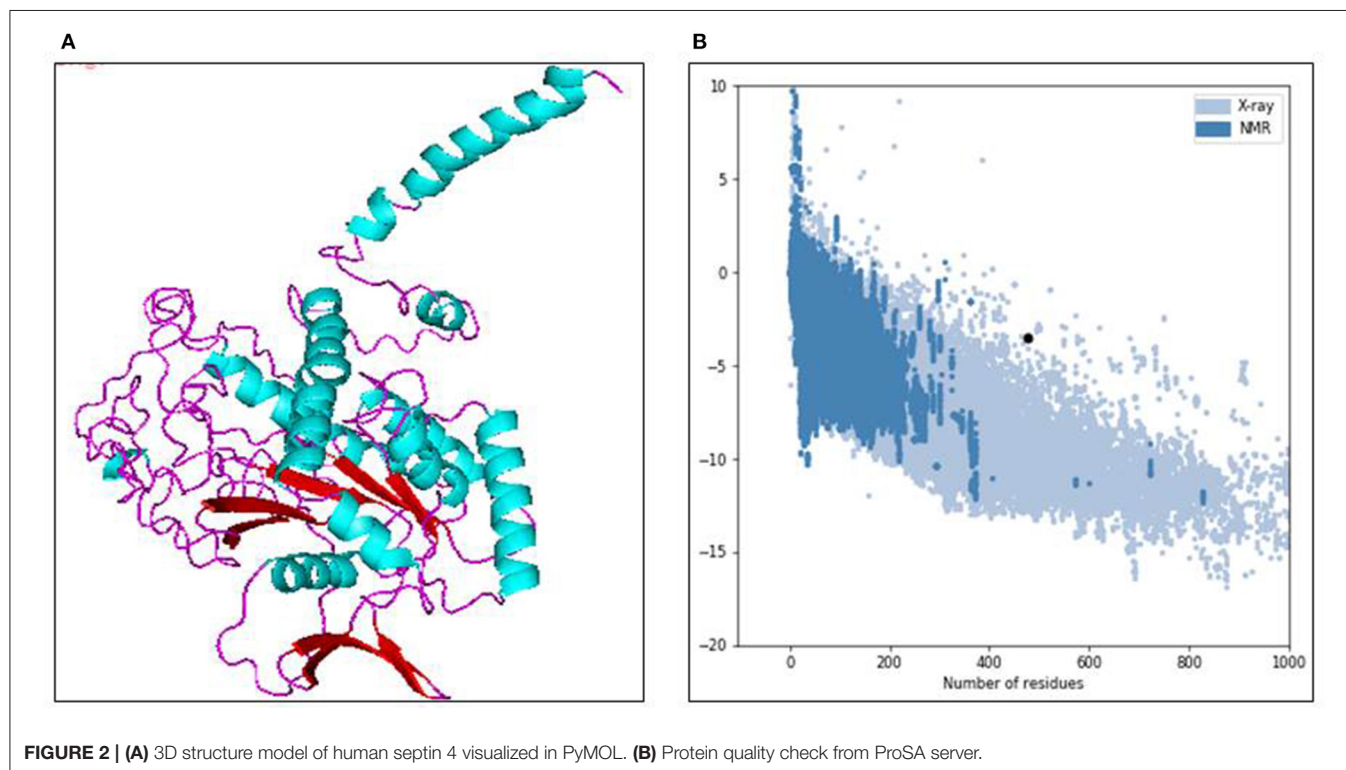
found to be visible in the SDS-PAGE, with a maximum of bands existing in the case of 50 and 110 kDa proteins. The band around 55 kDa was missing in the case of asthenospermia, but present in the case of oligospermia, normospermia, and healthy volunteers. We guessed the importance of missed 55-kDa protein and further, we want to investigate this protein. The missed protein was isolated from normospermia and healthy volunteers and then MALDI-TOF analysis was done for eight samples to access the similarity in the results. Also, a Mascot search was done by using the MALDI-TOF results. The missing protein in asthenospermia was identified as Septin-4. It has already been evidenced that this protein had played a major role in Alzheimer's disease, male infertility, and Down syndrome. The role of Septin-4 in male infertility is enormous and more molecular work is in need for the prediction of the pathway mechanism behind male infertility. The correlation of motility and its implications with male infertility diagnosis is the key to success.

Mascot Search and Its Implications

The data got through m/z values were analyzed for each sample was searched in mascot MALDI-TOF-MS ions search. The database used as SwissProt, humans as chosen for taxonomy and enzyme as trypsin in the search tool. The parameters used for searching the protein of interest through mascot search were tabulated in **Table 1**. We looked for a maximum of hits and were obtained against the Septin-4 protein. The functions of the query protein were reviewed in Swissprot and involves in male infertility if downregulated in certain patients. Database

TABLE 1 | Mascot Search for the identified protein by MALDI TOF and its parameter search.

Variable modification	Protein	Fixed modification	Sequence coverage (%)	Significance score
Carbamidomethyl (C)	Septin 4 (<i>Homo sapiens</i>)	Carbamidomethyl (N-term)	91	85
Carbamidomethyl (C)	Septin family (Mice)	Carbamidomethyl (N-term)	85	57
Carbamyl (K)		Carbamyl (N-Term)		
Carbamidomethyl (C)	Semenogelin II	Carbamidomethyl (N-term)	80	52
Carbamyl (K)	(<i>Homo sapiens</i>)	Carbamyl (N-Term)		



search was performed in PDB and observed that the absence of experimental structure of human Septin-4. The *in-silico* approach has been used to predict the structure of protein for further research studies.

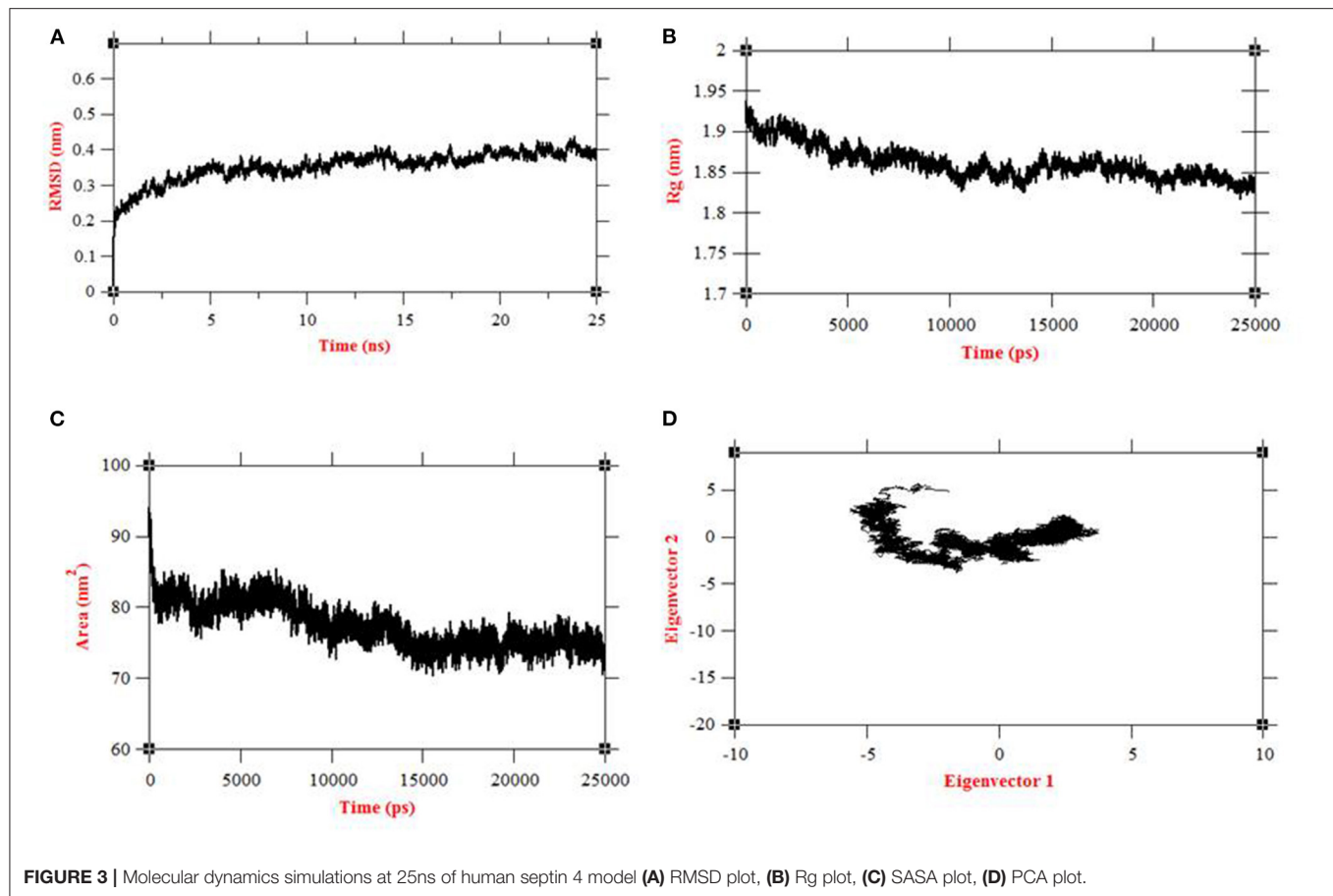
Structure Prediction

Using the *in-silico* structural study on human Septin-4, the 3D structural model was predicted from the I-Tasser server. Out of five models, the model with the least C-score -3.19 was selected as the best structure of Septin-4. The other parameters also supported the model with an estimated TM score of 0.36 ± 0.12 and an estimated RMSD of $15.2 \pm 3.5 \text{ \AA}$. The threading/fold recognition method screened the structure of the GTPase domain of human Septin-12 (PDB code: 6MQ9) as the template for the Septin-4 modeling. The Septin-4 model falls under the structural classification of alpha + beta, the architecture of the three-layer ($\alpha\beta\alpha$) sandwich, and the topology of the Rossmann fold, and is visualized in PyMol in **Figure 2A**. The major molecular function of the septin family was catalytic activity, GTPase activity, hydrolase activity, protein binding, lipid binding, and

protein dimerization activity. The quality of the model was deciphered by the ProSA server and results showed the Z-score of -3.5 that related to experimental structures in **Figure 2B**. The above predicted human Septin-4 structural model can be used for further annotation studies related to male infertility mechanisms.

Molecular Dynamics Simulations

The convergence of the protein system during simulations was measured by RMSD of all C α atoms from the initial structure. The initial equilibration of the native structure of human Septin-4 was done in 5 ns. After the equilibration phase, the structure of Septin-4 showed an RMSD range in 0.3–0.4 nm during 25 ns simulations (**Figure 3A**). The structure was well-converged and confirmed the protein stability of Septin-4 at end of simulations and structure with a stable trajectory in the dynamic system. Radius of gyration was the property of the overall dimension of protein during simulations. The Rg is termed as a measure of mass-weighted root mean square distance of all atoms from the center of mass. Radius of gyration of Septin-4 native structure started with 1.92 nm but gradually decrease to equilibrate



with 1.85 nm (Figure 3B). Thus, the overall protein folding pattern of human Septin-4 protein was observed. A solvent-accessible surface (SASA) plot was constructed and results showed the accessibility area around 75–80 nm² confirmed the behavior of the hydrophilic and hydrophobic residues in Septin-4 (Figure 3C). Principal component analysis was performed based on two steps. In the first step, the covariance matrix was constructed and diagonalized based on C α atoms using g_covar and trace value of 5.52816 nm². The eigenvectors and corresponding eigenvalues were evaluated from the covariance matrix using the motion of protein at the atom level. Then PCA was done using g_anaeig with the projection of the first two eigenvectors (eigenvector 1 vs. eigenvector 2) and the maximum motion extracted from the production run of 20 ns. The local motion of the PCA plot showed the overall motion of human Septin-4 in the dynamic system related to eigenvector 1 vs. eigenvector 2. The cluster was more compact and deciphered the motion of protein with covariance matrix (Figure 3D).

DISCUSSION

Homozygous Septin-4 (Human semen Septin-4) deletion or its downexpression was shown to have a complete or partial defect in the structure of the sperm flagellum; this means it helps a lot

for the forward motility (52, 53). In our results also, the Septin-4 absent or less expressed yield with less motility and especially with forward type. The defect in the flagella or neck region always yields these types of results (54, 55). Other researchers worked with Septin-4-null sperm or flagella modified with no annulus, this structure has been replaced by thin segment missing cortical material, acts like an abnormal-flagella conferring a hairpin-like structure (56–58). Two major hypothetical utilities have long been ascribed to the annulus of the spermatozoa: one is a diffusion barrier function; it is a very essential function for the fertilization, detaining proteins to various compartments of the sperm tail to the neck (59, 60). The second one is might be on morphological planner function given guidance to the growth of the flagellum and the association of the mitochondria along the axoneme. Both of these mechanisms were found to be failed in the case of Septin-4 null sperm. Morphology of human sperm annulus/flagellum has been known for a long time, but the mechanism by which it is correlating is poorly studied (56, 61). Sperm flagella biogenesis, the biochemical composition of the sperm tail to neck, and its functions remained as same in the case of rigorous research. For the last decade, septins have appeared and been explored as constitutive components of the annulus/flagella of spermatozoa and persuasive evidence has been evidenced by many researchers and suggest that a very stable septin complex/Septin-4 is the prerequisite for morphological

differentiation of the sperm tail, neck and with an important mechanism of diffusion barrier function (56). Although current evidence suggests that septins bind to the plasma membrane via interaction with phosphoinositides, our previous research with prostasomes suggest that the Zn present on prostasomes may transfer the essentials of needed motility factors and phospholipids for proper movement (62), this achieved through the fusion process of prostasomes and spermatozoa by means of protein dependent or pH dependent (63, 64). This finding suggests that binding to integral membrane proteins could also be involved. Moreover, the advance of *in-silico* studies deciphered the structural annotation of human Septin-4 that can be used to understand the role of septin in male infertility. Molecular modeling is the current best method used in the 3D structure prediction of key protein/enzymes/drug targets in proteomics. From the model structure, the major mechanism of Septin-4 has been studied using the structural arrangements of helix and sheets. The structure–function relationship is highly critical in the research area of male infertility, as very few 3D experimental structures are available. Also, advancements in MDs simulation deciphered the behavior of novel biomarker protein Septin-4 in the all-atom dynamics. *In-silico* finding acts as a critical point that can initiate various structure-function studies on human Septin-4 toward male infertility mechanism and pharmacology aspects.

CONCLUSION

Septins are the most important constituents of the annulus in spermatozoa, a submembranous ring that disconnects the middle and primary pieces of spermatozoa. This is believed to be an important protein Septin-4 that plays a major role in motility and its absence may be associated with asthenospermia. Many researchers previously reported its essential role in spermatogenesis and reproduction in animal models. Till now many researchers worked with labeling techniques and identified the importance of Septin-4 in the case of male infertility. In this current research work, we elucidated and identified the presence of Septin-4 in normal healthy sperm samples and its absence or less expression in the case of other infertile groups especially in the case of motility-related issues. The importance of Septin-4 in male fertility was proved with 3D structural modeling from *in-silico* characterization and MDs simulation confirmed the role of stable Septin-4 in the dynamic system. Less expression was found exclusively in infertile patients when compared to fertile

patients. Further research on Septin-4 with structural studies may be used to explore more on the mechanism and its role in spermatogenesis and human infertility. Hence, our findings concluded that Septin-4 was a novel biomarker for male infertility and can be used for diagnosis and pharmacology purposes.

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 Dr. Radha Saraswathy, Member Secretary-UHEC, Senior Professor, SBST, VIT, Vellore, Tamilnadu, India. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ASV involved in conceptualization, design protocol, performed experimental analysis, and wrote the manuscript on *in-vitro* studies section. KA involved in conceptualization, protocol design, experimental analysis, and wrote the manuscript on *in-silico* studies section. PJ, GG, RN, ST, and TS were involved in data validation, critical assessment, and edited the manuscript. KR involved in supervision of the research work and edited the final manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors would like to express their gratitude toward Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences for providing the necessary infrastructure and facilities to carry out this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.723019/full#supplementary-material>

REFERENCES

1. Bracke A, Peeters K, Punjabi U, Hoogewijs D, Dewilde S. A search for molecular mechanisms underlying male idiopathic infertility. *Reprod Biomed Online*. (2018) 36:327–39. doi: 10.1016/j.rbmo.2017.12.005
2. Kumar N, Singh AK, Choudhari AR. Impact of age on semen parameters in male partners of infertile couples in a rural tertiary care center of central India: a cross-sectional study. *Int J Reprod Biomed*. (2017) 15:497. doi: 10.29252/ijrm.15.8.497
3. Durairajanayagam D. Lifestyle causes of male infertility. *Arab J Urol*. (2018) 16:10–20. doi: 10.1016/j.aju.2017.12.004
4. Patel A, Sharma PS, Kumar P. Role of mental health practitioner in infertility clinics: a review on past, present and future directions. *J Hum Reprod Sci*. (2018) 11:219. doi: 10.4103/jhrs.JHRS_41_18
5. Nicola Z, Federica B, Simone P, Elettra V, Saverio CF. Infertility worldwide: the lack of global pediatric andrologists and prevention. In: Wu W, Ziglioli F, Maestroni U, editors. *Male Reproductive Health*. London: IntechOpen (2019). doi: 10.5772/intechopen.88459
6. Zedan H, Ismail S, Gomaa A, Saleh R, Henkel R, Agarwal A. Evaluation of reference values of standard semen parameters in fertile Egyptian men. *Andrologia*. (2018) 50:e12942. doi: 10.1111/and.12942

7. Suarez SS. Mammalian sperm interactions with the female reproductive tract. *Cell Tissue Res.* (2016) 363:185–94. doi: 10.1007/s00441-015-2244-2
8. Castillo J, Jodar M, Oliva R. The contribution of human sperm proteins to the development and epigenome of the preimplantation embryo. *Hum Reprod Update.* (2018) 24:535–55. doi: 10.1093/humupd/dmy017
9. Borziak K, Álvarez-Fernández A, Karr TL, Pizzari T, Dorus S. The Seminal fluid proteome of the polyandrous Red junglefowl offers insights into the molecular basis of fertility, reproductive ageing and domestication. *Sci Rep.* (2016) 6:35864. doi: 10.1038/srep35864
10. Lee RK, Tseng HC, Hwu YM, Fan CC, Lin MH, Yu JJ, et al. Expression of cystatin C in the female reproductive tract and its effect on human sperm capacitation. *Reprod Biol Endocrinol.* (2018) 16:1–0. doi: 10.1186/s12958-018-0327-0
11. Lan R, Xin M, Hao Z, You S, Xu Y, Wu J, et al. Biological functions and large-scale profiling of protein glycosylation in human semen. *J Proteome Res.* (2020) 19:3877–89. doi: 10.1021/acs.jproteome.9b00795
12. Xu Y, Lu H, Wang Y, Zhang Z, Wu Q. Comprehensive metabolic profiles of seminal plasma with different forms of male infertility and their correlation with sperm parameters. *J Pharm Biomed Anal.* (2020) 177:112888. doi: 10.1016/j.jpba.2019.112888
13. Vashisht A, Gahlay GK. Using miRNAs as diagnostic biomarkers for male infertility: opportunities and challenges. *Mol Hum Reprod.* (2020) 26:199–214. doi: 10.1093/molehr/gaaa016
14. Korbakis D, Schiza C, Brinc D, Soosaipillai A, Karakosta TD, Légaré C, et al. Preclinical evaluation of a TEX101 protein ELISA test for the differential diagnosis of male infertility. *BMC Med.* (2017) 15:60. doi: 10.1186/s12916-017-0817-5
15. Bieniek JM, Drabovich AP, Lo KC. Seminal biomarkers for the evaluation of male infertility. *Asian J Androl.* (2016) 18:426. doi: 10.4103/1008-682X.175781
16. Vickram AS, Samad HA, Latheef SK, Chakraborty S, Dhama K, Sridharan TB, et al. Human prostasomes an extracellular vesicle–Biomarkers for male infertility and prostate cancer: the journey from identification to current knowledge. *Int J Biol Macromol.* (2020) 146:946–58. doi: 10.1016/j.ijbiomac.2019.09.218
17. Baskaran S, Finelli R, Agarwal A, Henkel R. Diagnostic value of routine semen analysis in clinical andrology. *Andrologia.* (2021) 53:e13614. doi: 10.1111/and.13614
18. Mike-Antoine MA, Sihaliyolo J, Agasa SB, O'yandjo AM, Mbutu B, Bosunga GK, et al. Anti-Sperm antibodies: risk factors of positive serology among infertile men patients in Kisangani-Democratic Republic of Congo. *Open J Obstetr Gynecol.* (2019) 9:1130. doi: 10.4236/ojog.2019.98109
19. Morin SJ, Scott RT. Knowledge gaps in male infertility: a reproductive endocrinology and infertility perspective. *Transl Androl Urol.* (2018) 7:S283. doi: 10.21037/tau.2018.05.02
20. Baskaran S, Selvam MK, Agarwal A. Exosomes of male reproduction. *Adv Clin Chem.* (2020) 95:149–63. doi: 10.1016/bs.acc.2019.08.004
21. Gervasi MG, Xu X, Carbajal-Gonzalez B, Buffone MG, Visconti PE, Krapf D. The actin cytoskeleton of the mouse sperm flagellum is organized in a helical structure. *J Cell Sci.* (2018) 131:jcs215897. doi: 10.1242/jcs.215897
22. Netherton JK, Hetherington L, Ogle RA, Velkov T, Baker MA. Proteomic analysis of good- and poor-quality human sperm demonstrates that several proteins are routinely aberrantly regulated. *Biol Reprod.* (2018) 99:395–408. doi: 10.1093/biolre/iox166
23. Toure A, Martinez G, Kherraf ZE, Cazin C, Beurois J, Arnoult C, et al. The genetic architecture of morphological abnormalities of the sperm tail. *Hum Genet.* (2020) 140:1–22. doi: 10.1007/s00439-020-02113-x
24. Abbey M, Gaestel M, Menon MB. Septins: active GTPases or just GTP-binding proteins? *Cytoskeleton (Hoboken).* (2019) 76:55–62. doi: 10.1002/cm.21451
25. Phatarpekar PV, Overlee BL, Leehan A, Wilton KM, Ham H, Billadeau DD. The septin cytoskeleton regulates natural killer cell lytic granule release. *J Cell Biol.* (2020) 219:e202002145. doi: 10.1083/jcb.202002145
26. Wang WL, Tu CF, Tan YQ. Insight on multiple morphological abnormalities of sperm flagella in male infertility: what is new? *Asian J Androl.* (2020) 22:236–45. doi: 10.4103/aja.aja_53_19
27. Lin CH, Shen YR, Wang HY, Chiang CW, Wang CY, Kuo PL. Regulation of septin phosphorylation: SEPT12 phosphorylation in sperm septin assembly. *Cytoskeleton.* (2019) 76:137–42. doi: 10.1002/cm.21491
28. Condrat CE, Thompson DC, Barbu MG, Bugnar OL, Boboc A, Cretoiu D, et al. miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis. *Cells.* (2020) 9:276. doi: 10.3390/cells9020276
29. Beber A, Tavenau C, Nania M, Tsai FC, Di Cicco A, Bassereau P, et al. Membrane reshaping by micrometric curvature sensitive septin filaments. *Nat Commun.* (2019) 10:1–2. doi: 10.1038/s41467-019-08344-5
30. Spiliotis ET. Spatial effects – site-specific regulation of actin and microtubule organization by septin GTPases. *J Cell Sci.* (2018) 131:jcs207555. doi: 10.1242/jcs.207555
31. Neubauer, K., and Zieger, B. (2017), The mammalian septin interactome. *Front Cell Dev Biol.* 5:3. doi: 10.3389/fcell.2017.00003
32. Lepanto MS, Rosa L, Paesano R, Valenti P, Cutone A. Lactoferrin in aseptic and septic inflammation. *Molecules.* (2019) 24:1323. doi: 10.3390/molecules24071323
33. Valadares NE, Pereira HD, Araujo AP, Garratt RC. Septin structure and filament assembly. *Biophys Rev.* (2017) 9:481–500. doi: 10.1007/s12551-017-0320-4
34. Rafae A, Mohseni Meybodi A, Yaghmaei P, Hosseini SH, Sabbaghian M. Single-nucleotide polymorphism c. 474G> A in the SEPT12 gene is a predisposing factor in male infertility. *Mol Reprod Dev.* (2020) 87:251–9. doi: 10.1002/mrd.23310
35. Akhmetova KA, Chesnokov IN, Fedorova SA. Functional characterization of septin complexes. *Mol Biol.* (2018) 52:137–50. doi: 10.1134/S0026893317050028
36. Shen YR, Wang HY, Kuo YC, Shih SC, Hsu CH, Chen YR, et al. SEPT12 phosphorylation results in loss of the septin ring/sperm annulus, defective sperm motility and poor male fertility. *PLoS Genet.* (2017) 13:e1006631. doi: 10.1371/journal.pgen.1006631
37. Barzi NV, Kakavand K, Sodeifi N, Ghezelayagh Z, Sabbaghian M. Expression and localization of septin 14 gene and protein in infertile men testis. *Reprod Biol.* (2020). 20:164–8. doi: 10.1016/j.repbio.2020.03.007
38. Marquardt J, Chen X, Bi E. Architecture, remodeling, and functions of the septin cytoskeleton. *Cytoskeleton.* (2019) 76:7–14. doi: 10.1002/cm.21475
39. Dietrich MA, Nynca J, Ciereszko A. Proteomic and metabolomic insights into the functions of the male reproductive system in fishes. *Theriogenology.* (2019) 132:182–200. doi: 10.1016/j.theriogenology.2019.04.018
40. Cooper TM, Noonan E, Von Eckardstein S, Auger J, Baker HW, Behre HG, et al. World Health Organization reference values for human semen characteristics. *Hum Reprod Update.* (2010) 16:231–45. doi: 10.1093/humupd/dmp048
41. World Health Organization (WHO) Report. (2010).
42. Alshahrani S, Aldossari K, Al-Zahrani J, Gabr AH, Henkel R, Ahmad G. Interpretation of semen analysis using WHO 1999 and WHO 2010 reference values: abnormal becoming normal. *Andrologia.* (2018) 50:e12838. doi: 10.1111/and.12838
43. Berkman SJ, Roscoe EM, Bourret JC. Comparing self-directed methods for training staff to create graphs using Graphpad Prism. *J Appl Behav Anal.* (2019) 52:188–204. doi: 10.1002/jaba.522
44. Rao M, Wu Z, Wen Y, Wang R, Zhao S, Tang L. Humanin levels in human seminal plasma and spermatozoa are related to sperm quality. *Andrology.* (2019) 7:859–66. doi: 10.1111/andr.12614
45. Lowry OI, Rosebrough NJ, Farr A, Randall RJ. Protein determination by a modified Folin phenol method. *J Biol Chem.* (1951) 193:265–75. doi: 10.1016/S0021-9258(19)52451-6
46. Mann T. Fructose content and fructolysis in semen. Practical application in the evaluation of semen quality. *J Agric Sci.* (1948) 38:323–31. doi: 10.1017/S0021859600006109
47. Yassa DA, Idriss WK, Atassi ME, Rao SK. The diagnostic value of seminal α -glucosidase enzyme index for sperm motility and fertilizing capacity. *Saudi Med J.* (2001) 22:987–91.
48. Huang YL, Tseng WC, Cheng SY, Lin TH. Trace elements and lipid peroxidation in human seminal plasma. *Biol Trace Elem Res.* (2000) 76:207–15. doi: 10.1385/BTER:76:3:207
49. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* (2015) 43:W174–81. doi: 10.1093/nar/gkv342
50. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: high performance molecular simulations through multi-level

- parallelism from laptops to supercomputers. *SoftwareX*. (2015) 1–2:19–25. doi: 10.1016/j.softx.2015.06.001
51. Anbarasu K, Jayanthi S. Structural modeling and molecular dynamics studies on the human LMTK3 domain and the mechanism of ATP binding. *Mol Biosyst*. (2014) 10:1139–45. doi: 10.1039/c4mb00063c
 52. Kissel H, Georgescu MM, Larisch S, Manova K, Hunnicutt GR, Steller H. The Sept4 septin locus is required for sperm terminal differentiation in mice. *Dev Cell*. (2005) 8:353–64. doi: 10.1016/j.devcel.2005.01.021
 53. Lindemann CB, Lesich KA. Functional anatomy of the mammalian sperm flagellum. *Cytoskeleton*. (2016) 73:652–69. doi: 10.1002/cm.21338
 54. Moretti E, Geminiani M, Terzuoli G, Renieri T, Pascarelli N, Collodel G. Two cases of sperm immotility: a mosaic of flagellar alterations related to dysplasia of the fibrous sheath and abnormalities of head-neck attachment. *Fertil Steril*. (2011) 95:1787–e19. doi: 10.1016/j.fertnstert.2010.11.027
 55. Gilpin W, Bull MS, Prakash M. The multiscale physics of cilia and flagella. *Nat Rev Phys*. (2020) 2:74–88. doi: 10.1038/s42254-019-0129-0
 56. Sugino Y, Ichioka K, Soda T, Ihara M, Kinoshita M, Ogawa O, et al. Septins as diagnostic markers for a subset of human asthenozoospermia. *J Urol*. (2008) 180:2706–9. doi: 10.1016/j.juro.2008.08.005
 57. Lehti MS, Sironen A. Formation and function of sperm tail structures in association with sperm motility defects. *Biol Reprod*. (2017) 97:522–36. doi: 10.1093/biolre/iox096
 58. Wang YY, Lai TH, Chen MF, Lee HL, Kuo PL, Lin YH. SEPT14 Mutations and teratozoospermia: genetic effects on sperm head morphology and DNA integrity. *J Clin Med*. (2019) 8:1297. doi: 10.3390/jcm8091297
 59. De Amicis F, Perrotta I, Santoro M, Guido C, Morelli C, Cesario MG, et al. Human sperm anatomy: different expression and localization of phosphatidylinositol 3-kinase in normal and varicocele human spermatozoa. *Ultrastruct Pathol*. (2013) 37:176–82. doi: 10.3109/01913123.2013.763881
 60. La Spina FA, Stival C, Krapf D, Buffone MG. Molecular and cellular aspects of mammalian sperm acrosomal exocytosis. In: Constantinescu G, Schatten H, editors. *Animal Models and Human Reproduction*. Hoboken, NJ: John Wiley & Sons, Inc. (2017) p. 409–26. doi: 10.1002/9781118881286.ch15
 61. Devlin DJ, Agrawal Zaneveld S, Nozawa K, Han X, Moye AR, Liang Q, et al. Knockout of mouse receptor accessory protein 6 leads to sperm function and morphology defects. *Biol Reprod*. (2020) 102:1234–47. doi: 10.1093/biolre/iaaa024
 62. Bertin A, McMurray MA, Thai L, Garcia III G, Votin V, Grob P, et al. Phosphatidylinositol-4, 5-bisphosphate promotes budding yeast septin filament assembly and organization. *J Mol Biol*. (2010) 404:711–31. doi: 10.1016/j.jmb.2010.10.002
 63. Gönczi M, Dienes B, Dobrosi N, Fodor J, Balogh N, Oláh T, et al. Septins, a cytoskeletal protein family, with emerging role in striated muscle. *J Muscle Res Cell Motil*. (2020) 1–5. doi: 10.1007/s10974-020-09573-8
 64. Pereira R, Sá R, Barros A, Sousa M. Major regulatory mechanisms involved in sperm motility. *Asian J Androl*. (2017) 19:5.

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

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

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



Development and Validation of a Novel Prognosis Prediction Model for Patients With Stomach Adenocarcinoma

Tong Wang^{1†}, Weiwei Wen^{2†}, Hongfei Liu³, Jun Zhang¹, Xiaofeng Zhang^{4,5,6*} and Yu Wang^{4,5,6*}

¹ School of Life Sciences, Zhejiang Chinese Medical University, Hangzhou, China, ² Department of Dermatology, Third People's Hospital of Hangzhou, Hangzhou, China, ³ Department of Zhiweibing, Ningbo Municipal Hospital of Traditional Chinese Medicine, Ningbo, China, ⁴ Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine, Hangzhou, China, ⁵ Hangzhou Institute of Digestive Diseases, Hangzhou, China, ⁶ Key Laboratory of Integrated Traditional Chinese and Western Medicine for Biliary and Pancreatic Diseases of Zhejiang Province, Hangzhou, China

OPEN ACCESS

Edited by:

Thirumal Kumar D.,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

E. Srinivasan,
Saveetha University, India
Ibrahim Jelaidan,
King Abdulaziz Medical City,
Saudi Arabia

*Correspondence:

Yu Wang
wangyu@zcmu.edu.cn
Xiaofeng Zhang
zxf837@tom.com

[†]These authors share first authorship

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 12 October 2021

Accepted: 22 November 2021

Published: 22 December 2021

Citation:

Wang T, Wen W, Liu H, Zhang J,
Zhang X and Wang Y (2021)
Development and Validation of a Novel
Prognosis Prediction Model for
Patients With Stomach
Adenocarcinoma.
Front. Med. 8:793401.
doi: 10.3389/fmed.2021.793401

Background: Stomach adenocarcinoma (STAD) is a significant global health problem. It is urgent to identify reliable predictors and establish a potential prognostic model.

Methods: RNA-sequencing expression data of patients with STAD were downloaded from the Gene Expression Omnibus (GEO) and the Cancer Genome Atlas (TCGA) database. Gene expression profiling and survival analysis were performed to investigate differentially expressed genes (DEGs) with significant clinical prognosis value. Overall survival (OS) analysis and univariable and multivariable Cox regression analyses were performed to establish the prognostic model. Protein-protein interaction (PPI) network, functional enrichment analysis, and differential expression investigation were also performed to further explore the potential mechanism of the prognostic genes in STAD. Finally, nomogram establishment was undertaken by performing multivariate Cox regression analysis, and calibration plots were generated to validate the nomogram.

Results: A total of 229 overlapping DEGs were identified. Following Kaplan-Meier survival analysis and univariate and multivariate Cox regression analysis, 11 genes significantly associated with prognosis were screened and five of these genes, including COL10A1, MFAP2, CTHRC1, P4HA3, and FAP, were used to establish the risk model. The results showed that patients with high-risk scores have a poor prognosis, compared with those with low-risk scores ($p = 0.0025$ for the training dataset and $p = 0.045$ for the validation dataset). Subsequently, a nomogram (including TNM stage, age, gender, histologic grade, and risk score) was created. In addition, differential expression and immunohistochemistry stain of the five core genes in STAD and normal tissues were verified.

Conclusion: We develop a prognostic-related model based on five core genes, which may serve as an independent risk factor for survival prediction in patients with STAD.

Keywords: stomach adenocarcinoma, GEO, TCGA, differentially expressed genes, prognostic model

BACKGROUND

Approximately 1.4 million people die each year worldwide from adenocarcinomas of the esophagus, stomach, colon, or rectum (1), of which stomach adenocarcinoma (STAD) has the third highest incidence and second highest for cancer-related mortality, and it remains a significant global health problem (2). In 2018, STAD was estimated to cause one million new cases and 781,000 deaths worldwide (3). Since the non-specific symptoms in early stages of the disease, STAD is typically not diagnosed until the disease has progressed to a more severe state, resulting in poor prognosis due to metastasis, intratumoral heterogeneity, chemotherapy resistance, etc. (4). This raises an urgent need for the development of reliable diagnostic, prognostic, and therapeutic molecular biomarkers of STAD.

Integrative bioinformatics analysis is one of the frontiers of biological research today and can be used to identify differential genes, screen prognostic biomarkers, and select appropriate treatment approach (5). Research on single-gene prediction is very concentrated, but it is not yet effective in prognosis. Polygenic combination has been reported to possess better predictive ability for cancer prognosis than single genes (6). For instance, Lu et al. (7) revealed that the dysregulated expression of the THBS family was closely related to STAD prognosis and tumor immunity. Additionally, Liu et al. (8) demonstrated that the SFRP family was potential targets for precision therapy and prognostic biomarkers for survival of patients with STAD. Although there are some polygene bioinformatics analysis studies, most of them focus on predicting of signatures, and there is still a lack of research on polygenic risk estimation model and predict prognosis of STAD.

In this study, we aimed to develop a prognostic model for the predict prognosis for patients with STAD. A large number of mRNA expression profiles of patients with STAD were downloaded from Gene Expression Omnibus (GEO) and the Cancer Genome Atlas-Stomach Adenocarcinoma (TCGA-STAD) database. Differential expression analysis was used to identify differentially expressed genes (DEGs) between STAD-related tissue and normal tissue. Then, survival analysis and univariable Cox regression analysis were performed to screen prognostic genes, and multivariable Cox regression analysis was used to establish a prognostic risk model. Further, protein-protein interaction (PPI) network, functional enrichment analysis, differential expression, and structure investigation of the core genes were performed. Finally, a nomogram that includes age, gender, tumor TNM stage, histologic grade, and 5-gene risk prediction model as an independent clinical factor was used to

predict the 1-, 3-, and 5-year survival rate of patients with STAD. The detailed flowchart of this work is provided in **Figure 1**.

METHODS

Data Source

The GEO database (<http://www.ncbi.nlm.nih.gov/geo>) was used to retrieve data with “stomach adenocarcinoma” as the keywords and human as the species. Datasets that covered cancer tissue and normal adjacent tissue, came from the same platform, and contained at least 20 samples were selected, and then, the gene expression profiles and their clinical data were downloaded. From TCGA portal (<https://tcga-data.nci.nih.gov/tcga/>), we collected the STAD RNA-seq data and related clinical parameters.

Differential Expression Analysis

Gene level expression data were normalized and then log2 transformation is provided by the limma package of R software (version 3.6.3). For GEO datasets, data were analyzed using the GEO2R analysis tool, and the DEGs were identified at adjusted $p < 0.05$ and $|\text{Log}_2\text{FC}| > 1$. For TCGA-STAD cohort, DEGs were identified with false discovery rate (FDR) < 0.05 and $|\text{Log}_2\text{FC}| > 1$ via the edge R package (9).

Prognostic Genes' Identification

Overlapping DEGs were screened based on the p -value and fold change (FC)/log(FC), and the top 50 genes were selected for Kaplan–Meier survival analysis. Log-rank p -values for Kaplan–Meier plots were calculated using an R package for survival analysis. Then, we screened genes with log-rank $p < 0.05$ as prognostic-related genes for subsequent analysis.

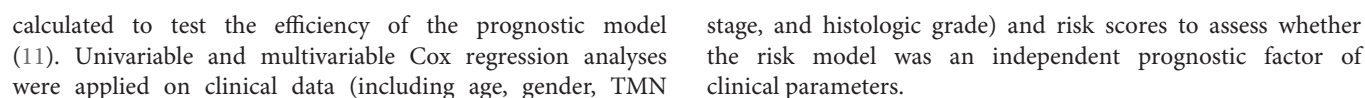
Construction and Validation Prognostic Model of STAD

To establish the prognostic model of STAD, univariable and multivariable Cox regression analyses were performed on the prognostic-related genes by the survival R package (10). Owing to the lack of survival information on GEO, we randomly divided the patients with complete survival information in TCGA-STAD dataset into training set and validation set, fit the model in the training set, and assessed its performance in the validation set. Then, the prognostic model was established based on corresponding coefficients of the prognostic genes of STAD.

$$\text{Risk score} = \sum_{i=1}^n \beta_i \times \text{Exp}_i$$

Further, the training set was divided into high- and low-risk groups according to the median value of risk score. Kaplan–Meier survival analysis was performed to estimate overall survival (OS) between the two groups by the survival R package. Time-dependent receiver operating characteristic (ROC) curves were plotted by time ROC R package, and the areas under ROC curves (AUCs) were

Abbreviations: STAD, stomach adenocarcinoma; GEO, Gene Expression Omnibus; TCGA, the Cancer Genome Atlas; DEGs, differentially expressed genes; TCGA-STAD, The Cancer Genome Atlas- Stomach Adenocarcinoma; PPI, protein–protein interaction; OS, overall survival; ROC, receiver operating characteristic; AUCs, areas under ROC curves; STRING, Search Tool for the Retrieval of Interacting Genes database; GO, gene ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; BP, biological process; MF, molecular function; CC, cell components; GSEA, gene set enrichment analysis; CRC, colorectal cancer; FDR, false discovery rate.



PPI Network Construction and Functional Enrichment Analysis

Interaction network analysis was obtained by employing STRING v11.5 database (<http://string-db.org/>), keeping default parameters. The topological properties of the PPI network included average shortest path length, betweenness centrality, closeness centrality, degree, eccentricity, neighborhood connectivity, radiality, stress, and topological coefficient. Molecular complex detection (MCODE) analysis was applied to the prognostic-related gene network to identify densely connected subnetwork modules. Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis were performed to identify significant pathways *via* the “cluster Profiler” package in R (12). The items of biological processes were further analyzed by GO classifications. The adjusted $p < 0.05$ was considered to indicate a statistically significant difference. In addition, gene set enrichment analysis (GSEA) was utilized to determine the core gene-related signaling pathways by the “cluster Profiler” package in R. Results with absolute value of normalized enrichment score > 1 , FDR < 0.25 , and adjust $p < 0.05$ were considered statistically significant. 1D linear domain structures and 3D structures of proteins were visualized using cBioPortal (<http://www.cbioportal.org/>).

Prognostic Gene Expression Investigation in STAD and Nomogram Construction

Differential expression of the prognostic genes between normal and STAD-related tissues was verified. Additionally, immunohistochemistry staining of the prognostic genes in STAD and normal tissues was acquired from the Human Protein Atlas database (<https://www.proteinatlas.org/>). According to the results of univariate and multivariate Cox regression analyses, a nomogram was created using the rms and survival package of R (13). Additionally, a calibration plots were generated to validate the nomogram.

RESULTS

Differential Expression Analysis

The clinical data of GSE27342, GSE63089, and TCGA-STAD were shown in **Table 1**. We found 474 DEGs in GSE27342 profile (287 upregulated and 187 downregulated, **Figure 2A**), 732 DEGs in GSE63089 profile (622 upregulated and 110 downregulated, **Figure 2B**), and 5,494 DEGs in TCGA-STAD cohort (2,659 upregulated and 2,835 downregulated, **Figure 2C**). Subsequently, a total of 229 overlapping DEGs (159 upregulated and 70 downregulated) were screened among the three datasets (**Figure 2D** and **Additional File 1**).

Prognostic-Related Genes' Identification

We selected the top 50 overlapping DEGs (cutoff: $p < 0.05$ and $|\log_2FC| > 1.75$) as candidate genes. According to log-rank $p < 0.05$ by Kaplan–Meier survival analysis, 11 genes (ADAM2, BGN, COL10A1, MMP1, MMP7, MFAP2, CTHRC1, P4HA3, SFRP4, TNFRSF11B, and FAP) were screened as prognostic-related genes for following research, and the Kaplan–Meier plots were shown in **Figure 3**.

TABLE 1 | Clinical or characteristics of patients with STAD in different datasets.

Characteristic	TCGA data (n, %)	GSE27342 (n, %)	GSE63089 (n, %)
Platform	Illumina HiSeq2000 RNA sequencing platform	Affymetrix Human Exon 1.0 ST Array	Affymetrix Human Exon 1.0 ST Array
Samples	407 (100.0%)	160 (100.0%)	90 (100.0%)
Normal	32 (7.9%)	80 (50.0%)	45 (50.0%)
Tumor	375 (92.1%)	80 (50.0%)	45 (50.0%)
Survival status	377 (92.6%)	NA	NA
Death	145 (35.6%)	NA	NA
Survival	232 (57.0%)	NA	NA
Age	366 (89.9%)	77 (48.1%)	70 (77.8%)
≤65	155 (38.1%)	59 (36.9%)	51 (56.7%)
>65	211 (51.8%)	18 (11.3%)	19 (21.1%)
Gender	380 (93.3%)	80 (50.0%)	70 (77.8%)
Female	137 (33.7%)	27 (16.9%)	25 (27.8%)
Male	243 (59.7%)	53 (33.1%)	45 (50.0%)
Stage	356 (87.5%)	80 (50.0%)	NA
I	55 (13.5%)	4 (2.5%)	NA
II	112 (27.5%)	7 (4.4%)	NA
III	150 (36.9%)	54 (33.8%)	NA
IV	39 (9.6%)	15 (9.4%)	NA
T classification	372 (91.4%)	NA	70 (77.8%)
T1	20 (4.9%)	NA	12 (13.3%)
T2	84 (20.6%)	NA	16 (17.8%)
T3	168 (41.3%)	NA	25 (27.8%)
T4	100 (24.6%)	NA	17 (18.9%)
N classification	362 (88.9%)	NA	NA
N0	113 (27.8%)	NA	NA
N1	99 (24.3%)	NA	NA
N2	76 (18.7%)	NA	NA
N3	74 (18.2%)	NA	NA
M classification	358 (88.0%)	NA	NA
M0	332 (81.6%)	NA	NA
M1	26 (6.4%)	NA	NA

TCGA, The Cancer Genome Atlas; NA, not available; T, primary tumor; N, regional lymph nodes; M, distant metastasis.

Construction and Validation Prognostic Model of STAD

Since the expression value of ADAM2 was zero in half of the samples, it was impossible to group by the median. The results of the univariate and multivariate proportional hazards regression analyses of the 10 prognostic-related genes associated with clinical outcomes are shown in **Table 2**. Multivariate regression analysis revealed COL10A1, MFAP2, CTHRC1, P4HA3, and FAP as the risk factor, and the risk score formula for OS was as follows:

$$\begin{aligned} \text{Riskscore} = & (-0.0013 \times \text{COL10A1Exp}) + (0.2709 \\ & \times \text{MFAP2Exp}) + (0.1869 \times \text{CTHRC1Exp}) + (0.1649 \\ & \times \text{P4HA3Exp}) + (9e - 04 \times \text{FAPExp}) \end{aligned}$$

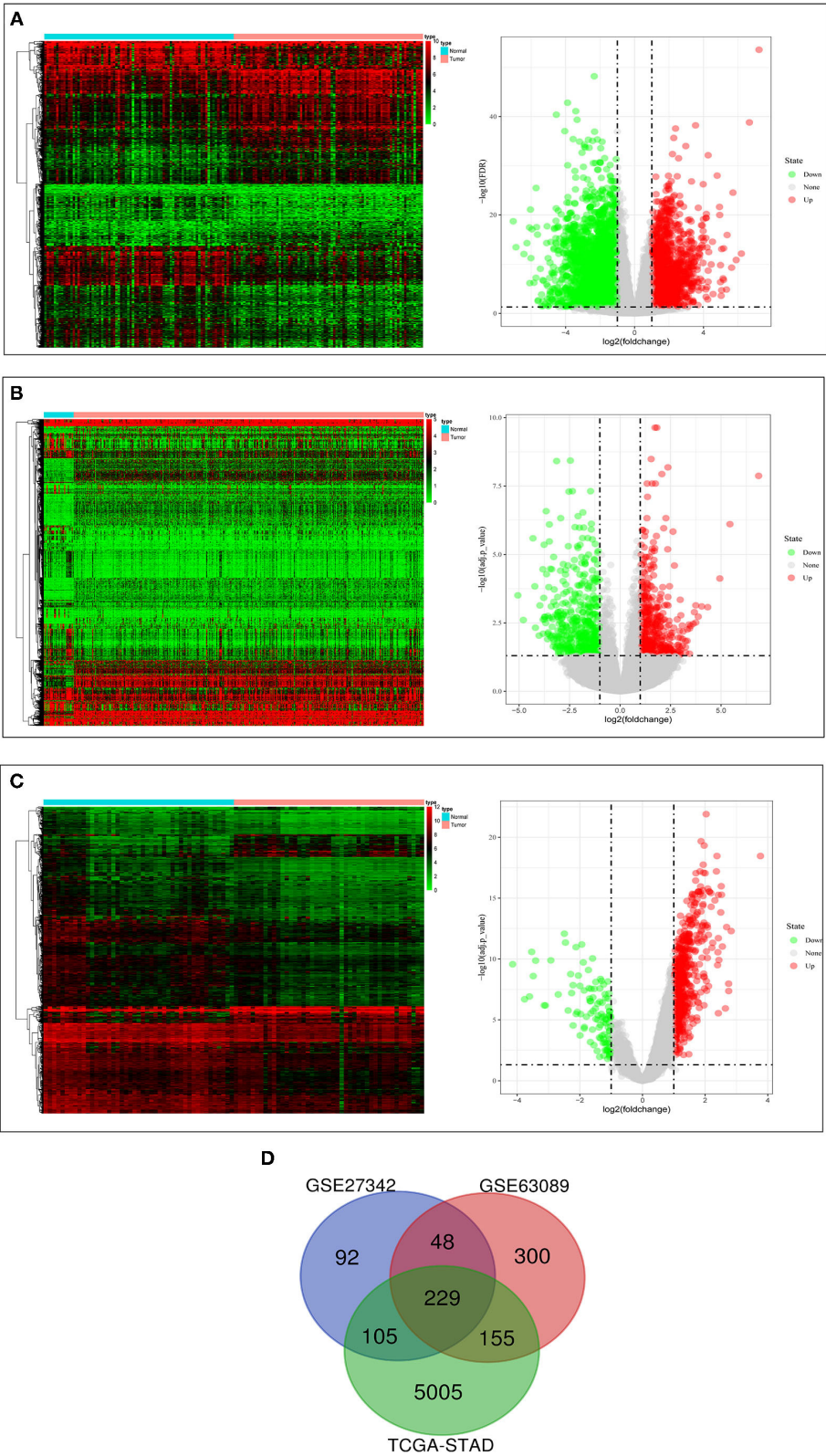


FIGURE 2 | The results of differential expression analysis. **(A)** The heatmap and volcano plots visualizing the DEGs in TCGA-STAD. **(B)** The heatmap and volcano plots visualizing the DEGs in GSE27342. **(C)** The heatmap and volcano plots visualizing the DEGs in GSE63089. **(D)** Venn diagram showing the overlapping DEGs in the three datasets.

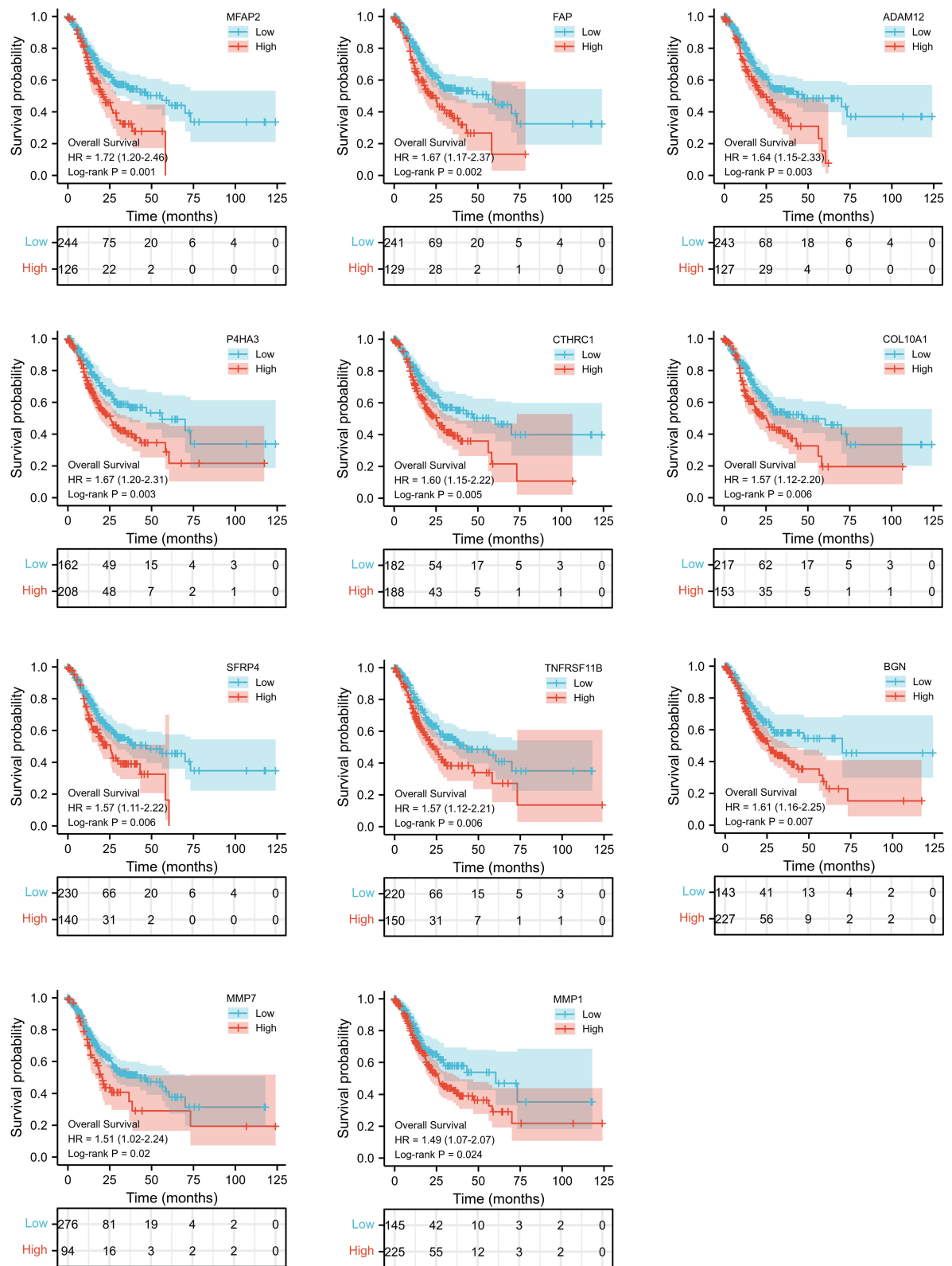
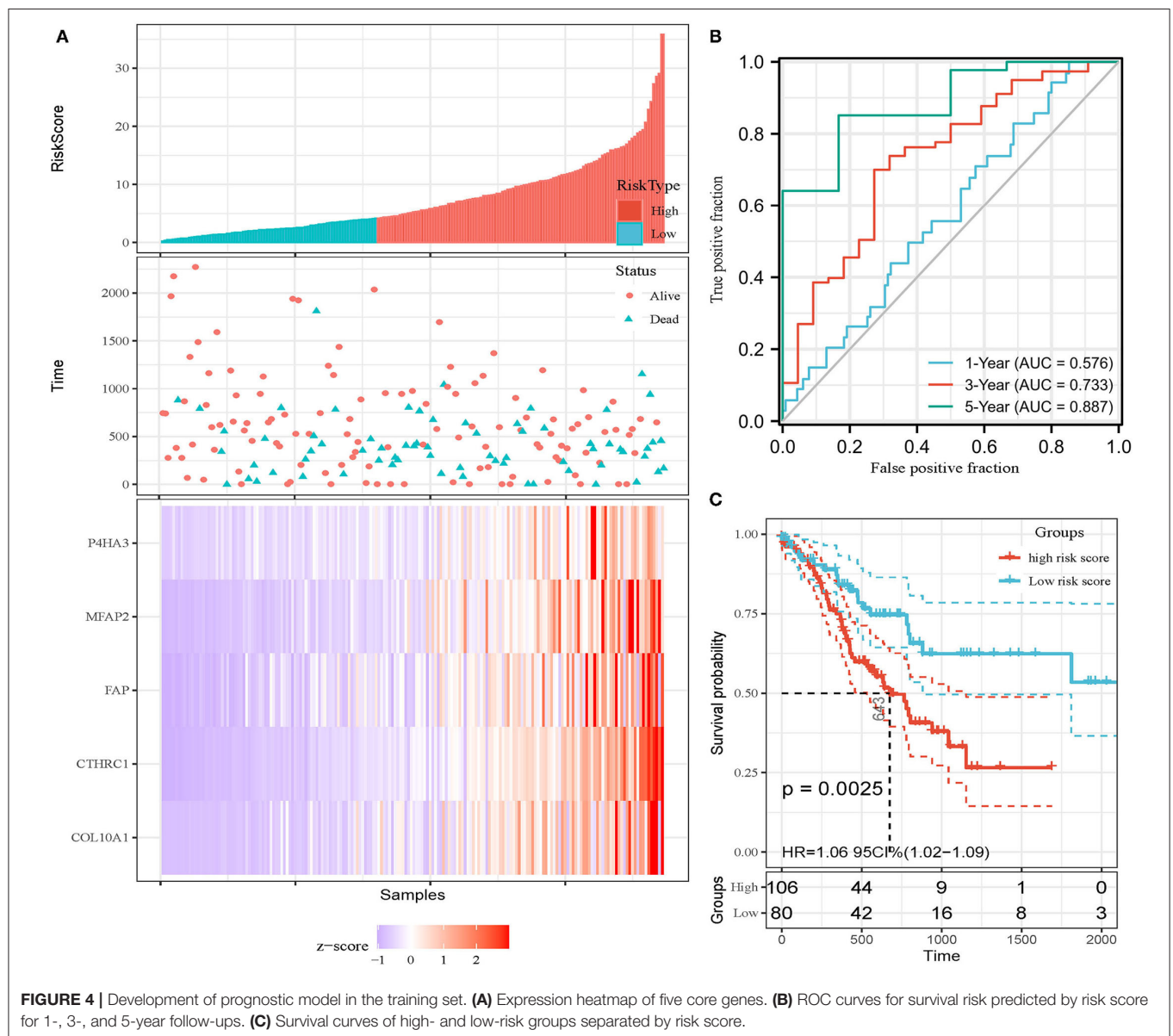
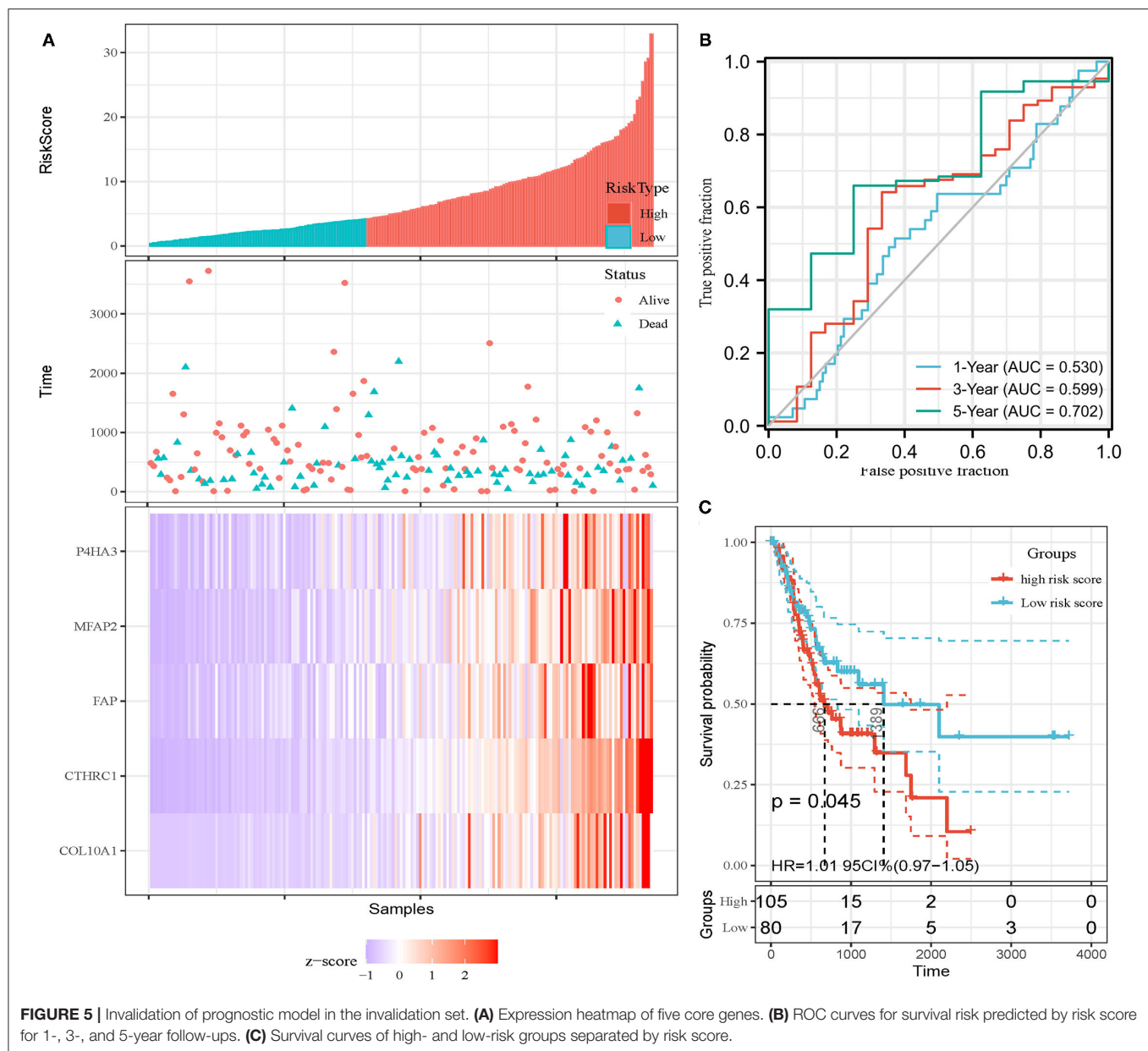


FIGURE 3 | Kaplan–Meier curves of 11 gene with prognostic value.

TABLE 2 | Univariate and multivariate Cox regression analyses.

Characteristics	Total (N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	p-value	Hazard ratio (95% CI)	p-value
COL10A1 (high vs. low)	370	1.434 (1.030–1.996)	0.033	0.999 (0.595–1.676)	0.996
MFAP2 (high vs. low)	370	1.593 (1.140–2.226)	0.006	1.311 (0.851–2.021)	0.220
CTHRC1 (high vs. low)	370	1.559 (1.118–2.174)	0.009	1.206 (0.708–2.052)	0.491
P4HA3 (high vs. low)	370	1.511 (1.084–2.107)	0.015	1.179 (0.740–1.878)	0.488
FAP (high vs. low)	370	1.415 (1.017–1.970)	0.039	1.001 (0.598–1.676)	0.997
BGN (high vs. low)	370	1.326 (0.954–1.844)	0.094		
MMP1 (high vs. low)	370	1.187 (0.855–1.648)	0.305		
MMP7 (high vs. low)	370	1.084 (0.781–1.504)	0.631		
SFRP4 (high vs. low)	370	1.267 (0.913–1.758)	0.157		
TNFRSF11B (high vs. low)	370	1.341 (0.966–1.862)	0.080		





In addition, the patients of TCGA-STAD dataset were divided into training set and validation set. Training set consisted of 186 STAD cases whereas validation set consisted of 185 STAD cases. Patients with STAD were divided into high- and low-risk subgroups according to the median value of risk score (cutoff = 14.9). In the training set, the survival analysis showed that the OS rates in the high-risk group were significantly lower than those in the low-risk group ($p = 0.0025$, **Figure 4C**). The time-dependent ROC curves offered a survival prediction that the AUCs were 0.576 (1-year OS), 0.733 (3-year OS), and 0.887 (5-year OS). Result showed that the risk model had a good ability to predict long-term prognosis of STAD (**Figure 4B**). The heatmap showed that the expression levels of five core genes were higher in patients with STAD with high-risk scores than

those with low-risk scores (**Figures 4A, 5A**). Meanwhile, data in the validation set showed the similar results: OS rates in the high-risk group were significantly lower than those in the low-risk group ($p = 0.045$, **Figure 5C**); the time-dependent ROC curves (**Figure 5B**) predicted that the AUCs were 0.530 (1-year OS), 0.599 (3-year OS), and 0.702 (5-year OS). Moreover, using multivariate Cox regression analysis, the prognostic model was identified as an independent predictor for patients with STAD ($p = 0.008$, **Table 3**).

PPI Network Construction and Functional Enrichment Analysis

The PPI network of the five core genes was shown in **Figure 6A**. The topological properties of the PPI network for each gene were

TABLE 3 | Univariate and multivariate Cox regression analyses for risk score of patients with STAD.

Characteristics	Total (N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Age	370				
≥65	213	Reference			
≤65	157	0.607 (0.430–0.856)	0.004	0.538 (0.369–0.785)	0.001
Gender	370				
Male	237	Reference			
Female	133	0.789 (0.554–1.123)	0.188		
T stage	362				
T1&T2	96	Reference			
T3&T4	266	1.719 (1.131–2.612)	0.011	1.410 (0.894–2.226)	0.140
M stage	352				
M0	327	Reference			
M1	25	2.254 (1.295–3.924)	0.004	2.489 (1.337–4.634)	0.004
N stage	352				
N0&N1	204	Reference			
N2&N3	148	1.650 (1.182–2.302)	0.003	1.517 (1.060–2.172)	0.023
Histologic grade	361				
G1&G2	144	Reference			
G3	217	1.353 (0.957–1.914)	0.087	1.399 (0.956–2.048)	0.084
Risk score	370	1.036 (1.012–1.062)	0.004	1.037 (1.010–1.065)	0.008

shown in **Additional File 6**. Highly interconnected subcluster of the five core genes was shown in **Figure 6B**, and the subcluster consisted of 53 nodes and 305 edges (score = 11.731), which represented relatively stable of protein in the network.

The results of GO enrichment analysis (**Figure 7A**) showed that the five core genes significantly focused on extracellular matrix organization, extracellular structure organization (biological process, BP); extracellular matrix structural constituent, dipeptidyl-peptidase activity (molecular function, MF); and collagen-containing extracellular matrix, collagen trimer (cell components, CC). Meanwhile, we found that in terms of biological processes, the genes were mainly focused on extracellular matrix, cell cycle, and Wnt signaling pathways. According to the *p*-value, the top five items from the three categories were selected to plot a histogram (**Figure 7B**). KEGG enrichment analysis (**Figure 7A**) indicated that prognostic genes were significantly enriched with arginine and proline metabolism and protein digestion and absorption, etc.

Gene set enrichment analysis was applied to determine their related signaling pathways (**Figures 7C–G**). COL10A1 was significantly enriched with olfactory transduction and nitrogen metabolism pathways, etc. CTHRC1 was significantly enriched in olfactory transduction and metabolism of xenobiotics by cytochrome p450 pathways, etc. MFAP2 was significantly enriched with olfactory transduction and fatty acid metabolism pathways, etc. P4HA3 was significantly enriched with ribosome and nitrogen metabolism pathways, etc. FAP was significantly enriched in nitrogen metabolism and metabolism of xenobiotics by cytochrome p450 pathways, etc.

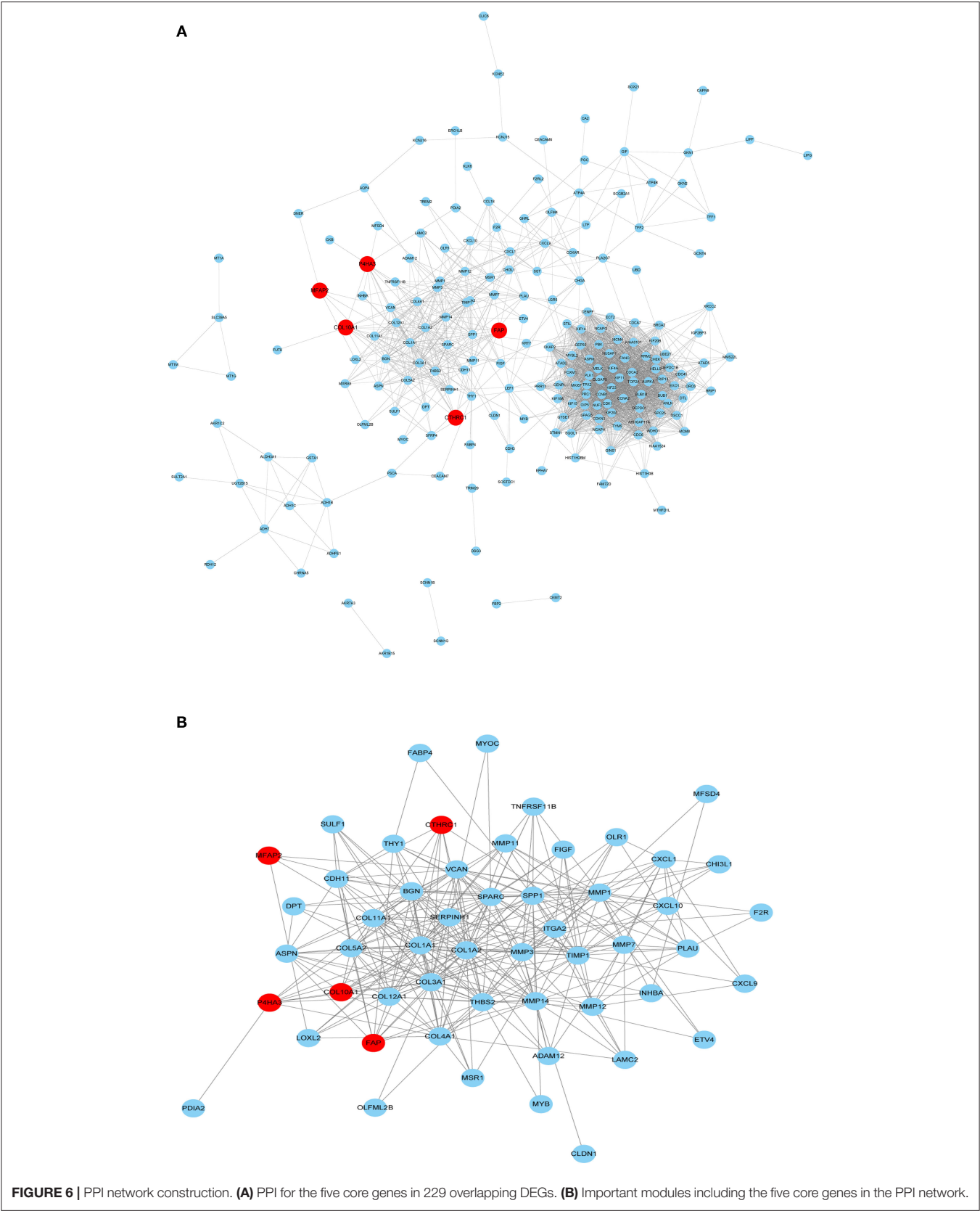
The mutation site and structure of the five core genes were shown in **Figure 8**.

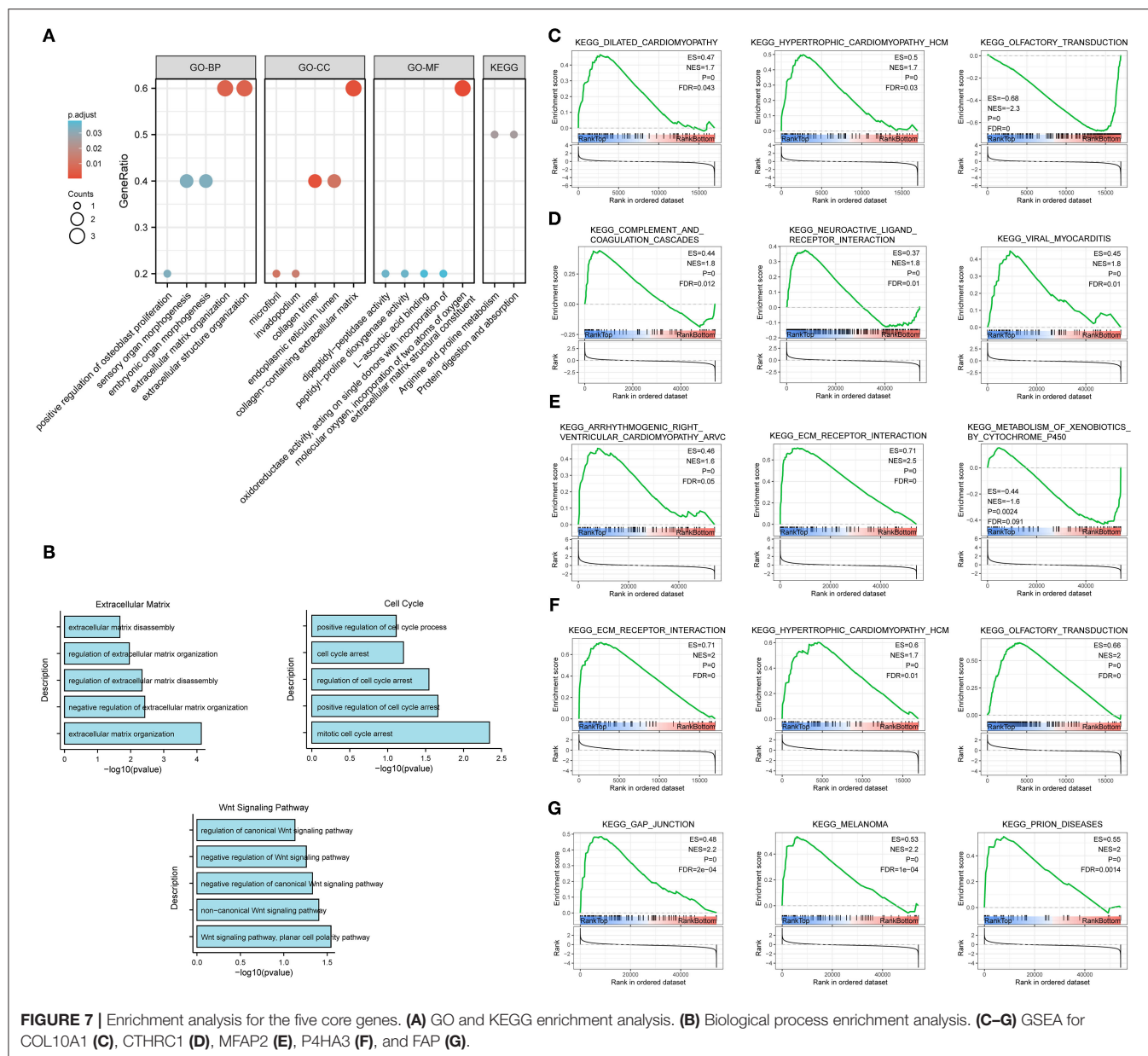
Prognostic Gene Expression Investigation in STAD and Nomogram Construction

Differential expression of the prognostic genes between normal and STAD-related tissues was verified. Results demonstrated that COL10A1, MFAP2, CTHRC1, P4HA3, and FAP were significantly upregulated in STAD-related tissues compared with normal tissues (**Figure 9A**). Additionally, immunohistochemistry staining of five core genes in STAD and normal tissues was acquired from the Human Protein Atlas database, which showed that differential expression of protein was consistent with gene expression (**Figure 9B**). However, the immunohistochemical images of COL10A1 were not found. Then, a nomogram (**Figure 10A**, including TNM stage, age, gender, histologic grade, and risk score) was created to predict the survival rate of patients with STAD at 1, 3, and 5 years. It was found that high total points predicted low 1-, 3-, and 5-year survival rates; however, a low total points did the opposite. The nomogram calibration plots (**Figure 10B**) indicate that the nomogram was well-calibrated, with mean predicted probabilities for 1- and 3-year OS close to observed probabilities.

DISCUSSION

The genetic background of STAD is complicated. Mining genes related to the prognosis of STAD from the genetic and molecular

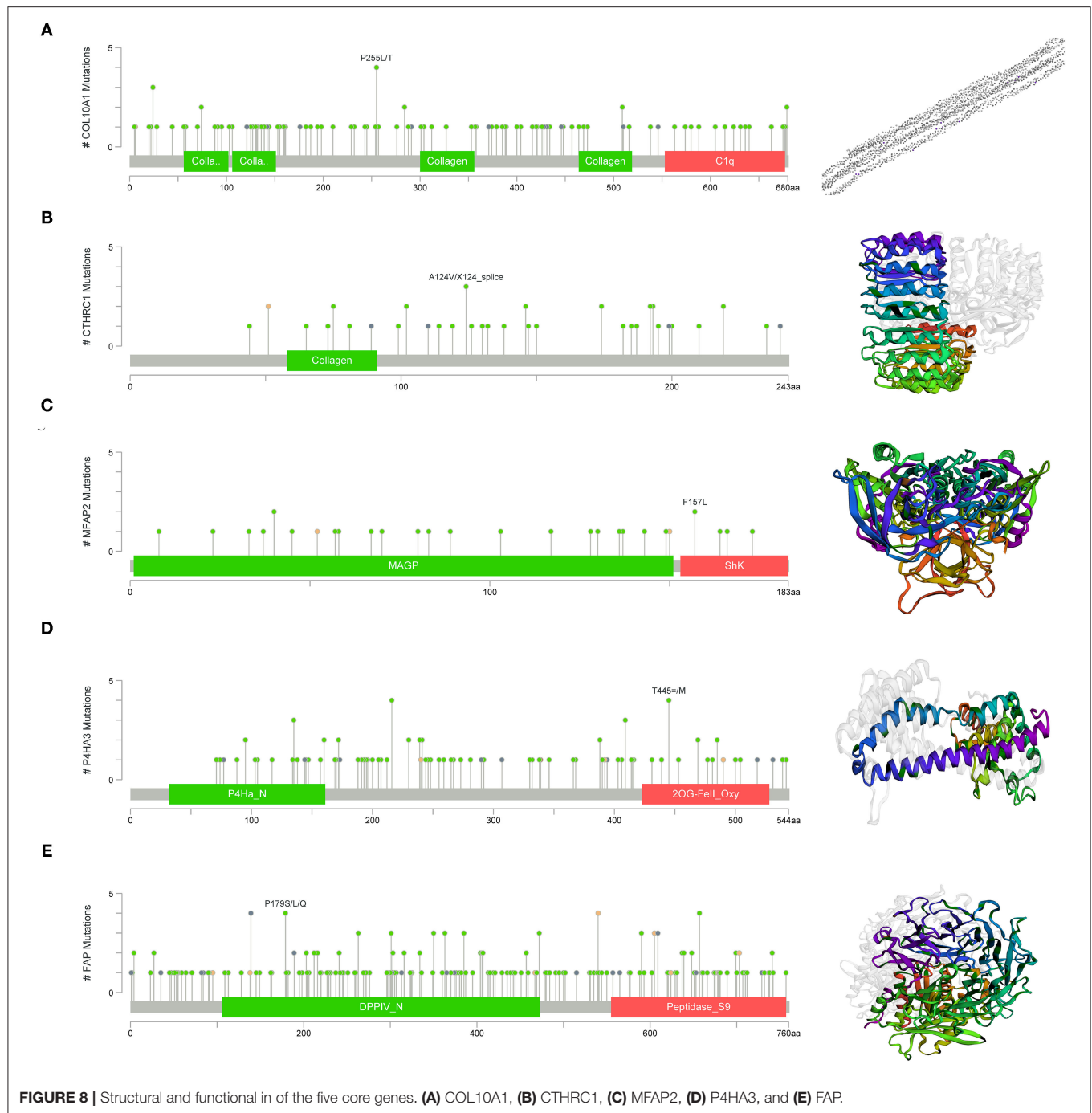




level is of great significance for the treatment and prognosis prediction of STAD. Bioinformatics analysis based on large databases has pointed out the direction for tumor research. In this study, we downloaded gene expression profiling and clinical data from the TCGA and GEO databases, identified DEGs, screened the prognostic-related genes, and then constructed a prognostic model based on five core genes (COL10A1, MFAP2, CTHRC1, P4HA3, and FAP).

COL10A1 is a member of the collagen family involved in tissue architecture and acts as a barrier to the migration of epithelial cells under normal conditions (14). Necula et al. (14) identified a significant increase in COL10A1 plasma level in patients with STAD and concluded that COL10A1 shows an elevated expression from the beginning of carcinogenesis, in the early

stages, and its increased level remains elevated during gastric cancer progression. Aktas et al. also found that COL10A1 is abnormally upregulated in gastric cancer and its high expression can be used as a diagnostic and/or prognostic biomarker (15). It has been reported that MFAP2 is upregulated in STAD, negatively correlated with OS, and can be used as a prognostic biomarker of STAD (16, 17), which is consistent with our results. Further, Yao et al. revealed that MFAP2 is overexpressed in gastric cancer and promotes motility *via* the MFAP2/integrin $\alpha 5 \beta 1$ /FAK/ERK pathway (18). CTHRC1 is a cancer-related gene that can promote cancer recurrence or metastasis *via* diverse signaling pathways, including TGF- β , MEK/ERK, and PKC- δ /ERK (19). Ding et al. (20) found that CTHRC1 promoted STAD metastasis through HIF-1 α /CXCR4 signaling pathway, which can be used



as a biomarker for STAD, and is consistent with our results. Moreover, CTHRC1 was demonstrated that overexpressed in hepatocellular carcinoma tissues significantly correlating with poor survival rate, which can be used as a prognostic marker for liver cancer (21). Consistent with the results of this study, P4HA3 has been repeatedly reported to be overexpressed in STAD and is related to the poor prognosis of STAD (22). Song et al. found that P4HA3 can be apparently activated by Slug in STAD tissues, of which imbalance and metastasis were related to poor survival

rates (23). FAP is a fibroblast activating protein, which has found to be involved in the growth and formation of a variety of cancers. Research revealed that FAP promoted the growth of intrahepatic cholangiocarcinoma through the recruitment of myeloid derived suppression cells (24). Additionally, the high expression of FAP in colorectal cancer is related to angiogenesis and immune regulation (25).

In this study, we established a polygene risk factors model for predicting prognostic of STAD, which is more rational than

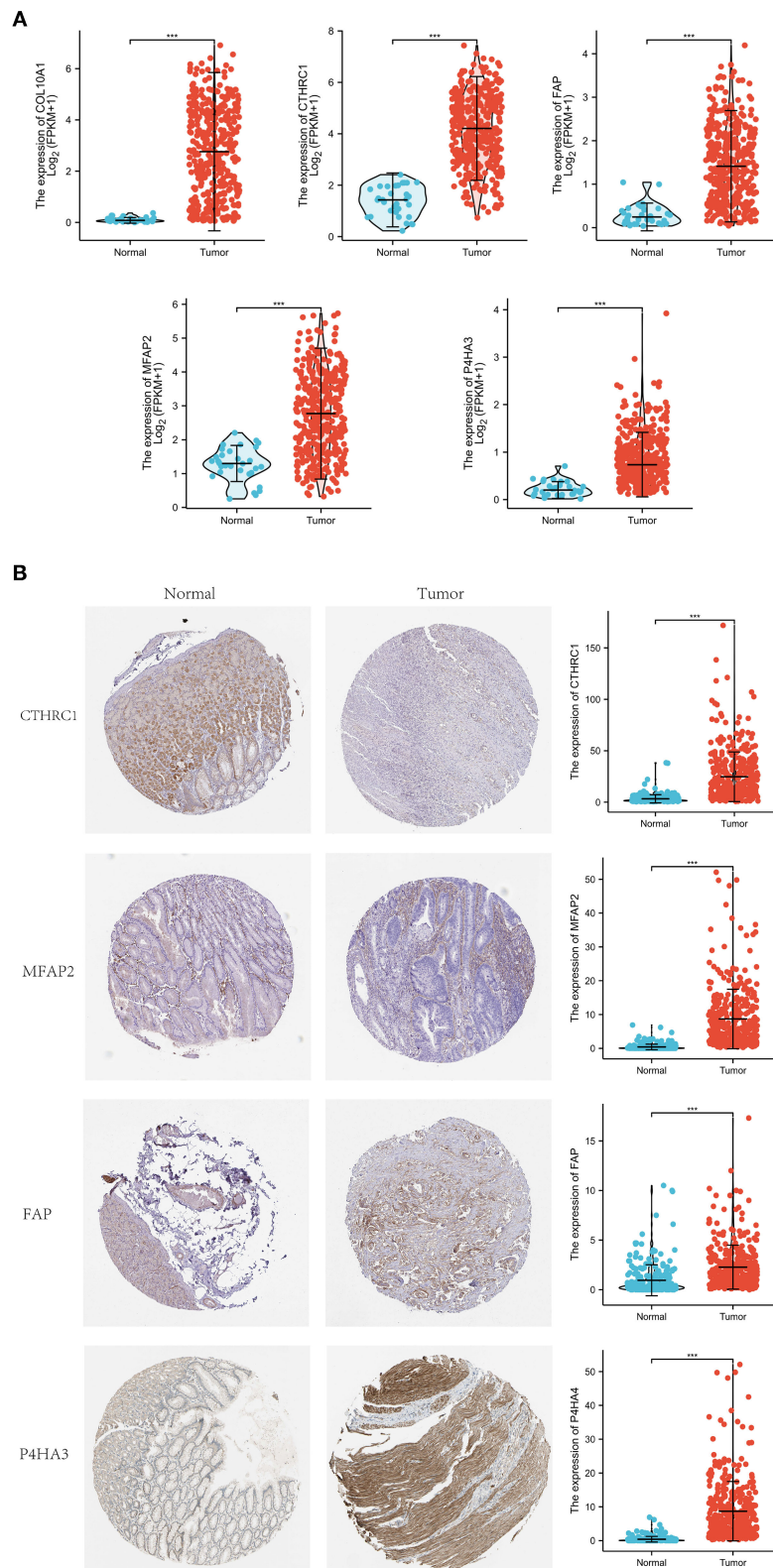


FIGURE 9 | Expression investigation of five core genes. **(A)** Differential expression of the five core genes between normal and STAD-related tissues. **(B)** Immunohistochemistry staining and their mRNA expression in normal and STAD-related tissues based on The Human Protein Atlas. ***p < 0.001.

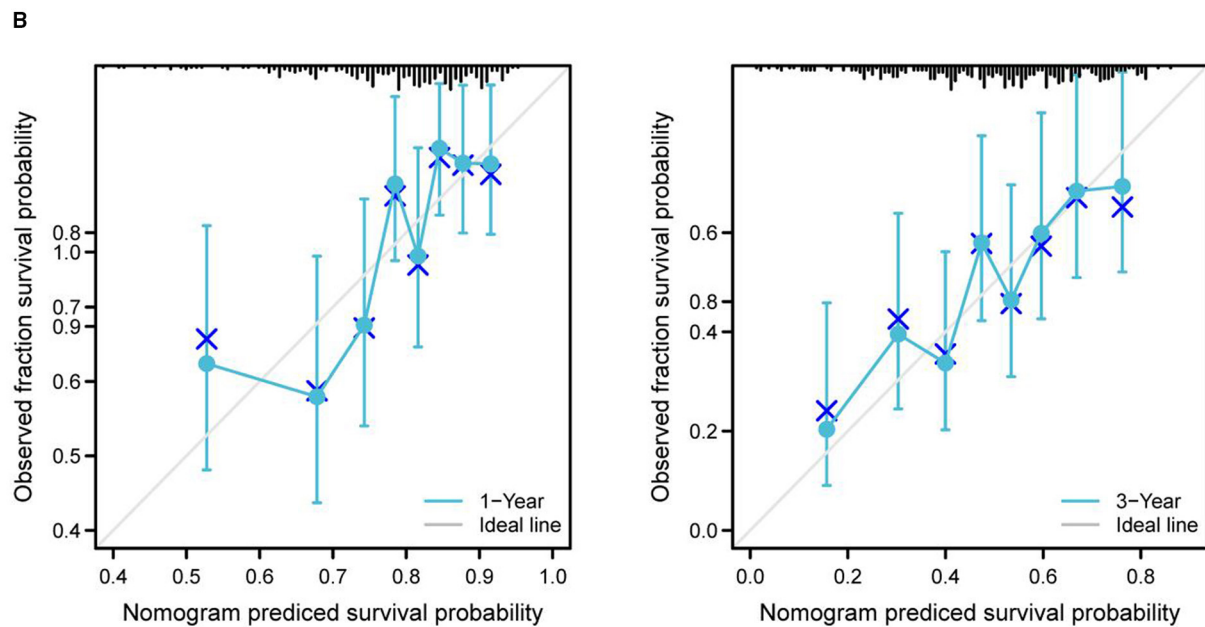
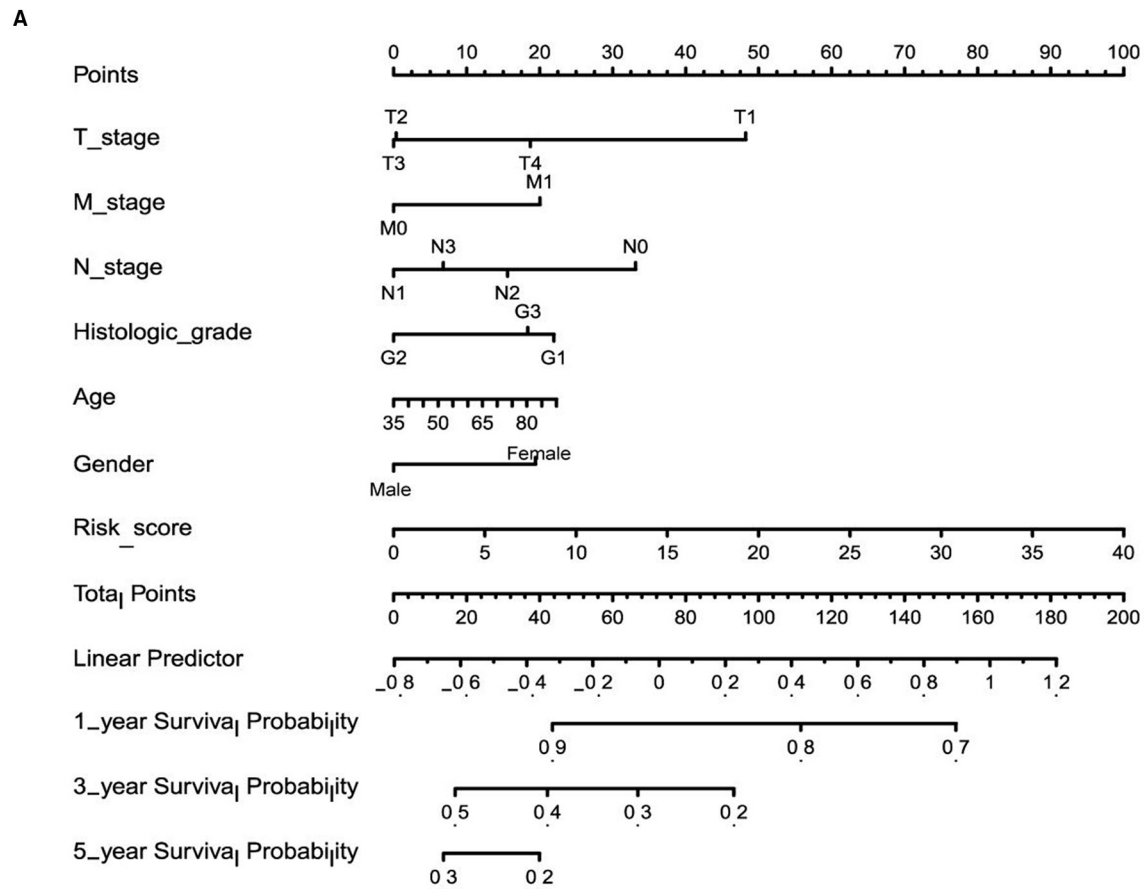


FIGURE 10 | Nomogram predicted the 1-, 3-, and 5-year survival rates of patients with STAD. **(A)** Nomogram predicting the 1-, 3-, and 5-year OS rates of patients with STAD. **(B)** Calibration plots for the 1- and 3-year OS nomogram.

single-risk factor. However, there are potential limitations to our analysis. First, this study has limitations inherent to a bioinformatics analysis. The construction of prognostic model is based on the TCGA and GEO database analysis and lacks clinical or cellular or animal functional experimental verification. Second, due to some patients with incomplete details are excluded, there may be selection bias in this study.

CONCLUSION

We developed a prognostic model for patients with STAD based on COL10A1, MFAP2, CTHRC1, P4HA3, and FAP, and a nomogram to predict the survival rate of patients with STAD at 1, 3, and 5 years. The evidence from this study comes from bioinformatics, as with other studies of a similar nature. It is still necessary to conduct further experiments to verify these findings.

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**.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

YW designed the study. TW and WW analyzed the data and wrote the original manuscript. HL and JZ

collected the data. XZ contributed to the manuscript revision. All authors have read and approved the final manuscript.

FUNDING

This study was supported by grants from the Zhejiang Provincial National Science Foundation (LY21H270013 and LGF21H310004), China Postdoctoral Science Foundation (2020M671818), Zhejiang Medical and Health Science and Technology Plan (WKJ-ZJ-2136 and 2019RC068), and Hangzhou Medical and Health Science and Technology Plan (2016ZD01, OO20190610, and A20200174).

ACKNOWLEDGMENTS

We are extremely grateful for reviewers' comments in helping this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.793401/full#supplementary-material>

Additional File 1 | Supplementary Table 1.

Additional File 2 | Supplementary Table 2.

Additional File 3 | Supplementary Table 3.

Additional File 4 | Supplementary Table 4.

Additional File 5 | Supplementary Table 5.

Additional File 6 | Supplementary Table 6.

Additional File 7 | Supplementary Table 7.

Additional File 8 | Supplementary Table 8.

REFERENCES

- Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell*. (2018) 33:721–35.e8. doi: 10.1016/j.ccell.2018.03.010
- Machlowska J, Baj J, Sitarz M, Maciejewski R, Sitarz R. Gastric cancer: epidemiology, risk factors, classification, genomic characteristics and treatment strategies. *Int J Mol Sci*. (2020) 21:114012. doi: 10.3390/ijms21114012
- Ho SWT, Tan P. Dissection of gastric cancer heterogeneity for precision oncology. *Cancer Sci*. (2019) 110:3405–14. doi: 10.1111/cas.14191
- Petryszyn P, Chapelle N, Matysiak-Budnik T. Gastric cancer: where are we heading? *Dig Dis*. (2020) 38:280–5. doi: 10.1159/000506509
- Matsuoka T, Yashiro M. Biomarkers of gastric cancer: current topics and future perspective. *World J Gastroenterol*. (2018) 24:2818–32. doi: 10.3748/wjg.v24.i26.2818
- Zhao H, Zhang S, Shao S, Fang H. Identification of a prognostic 3-gene risk prediction model for thyroid cancer. *Front Endocrinol*. (2020) 11:510. doi: 10.3389/fendo.2020.00510
- Lu Y, Kong X, Zhong W, Hu M, Li C. Diagnostic, therapeutic, and prognostic value of the thrombospondin family in gastric cancer. *Front Mol Biosci*. (2021) 8:647095. doi: 10.3389/fmolb.2021.647095
- Liu D, Sun C, Kim N, Bhan C, Tuason JPW, Chen Y, et al. Comprehensive analysis of SFRP family members prognostic value and immune infiltration in gastric cancer. *Life*. (2021) 11:522. doi: 10.3390/life11060522
- Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. (2010) 26:139–40. doi: 10.1093/bioinformatics/btp616
- Terry M, Therneau, Patricia M. *Grambsch Modeling Survival Data: Extending the Cox Model*. New York, NY: Springer (2000). doi: 10.1007/978-1-4757-3294-8
- Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*. (2013) 32:5381–97. doi: 10.1002/sim.5958
- Yu G, Wang LG, Han Y, He QY. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. (2012) 16:284–7. doi: 10.1089/omi.2011.0118
- Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. Cancer genome atlas research network. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*. (2018) 173:400–16.e11. doi: 10.1158/1538-7445.AM2018-3287
- Necula L, Matei L, Dragu D, Pitica I, Neagu AI, Bleotu C, et al. High plasma levels of COL10A1 are associated with advanced tumor

- stage in gastric cancer patients. *World J Gastroenterol.* (2020) 26:3024–33. doi: 10.3748/wjg.v26.i22.3024
15. Aktas SH, Taskin-Tok T, Al-Khafaji K, Akin-Bali DF. A detailed understanding of the COL10A1 and SOX9 genes interaction based on potentially damaging mutations in gastric cancer using computational techniques. *J Biomol Struct Dyn.* (2021) 2021:1–12. doi: 10.1080/07391102.2021.1960194
 16. Sun T, Wang D, Ping Y, Sang Y, Dai Y, Wang Y, et al. Integrated profiling identifies SLC5A6 and MFAP2 as novel diagnostic and prognostic biomarkers in gastric cancer patients. *Int J Oncol.* (2020) 56:460–9. doi: 10.3892/ijo.2019.4944
 17. Dong SY, Chen H, Lin LZ, Jin L, Chen DX, Wang OC, et al. MFAP2 is a potential diagnostic and prognostic biomarker that correlates with the progression of papillary thyroid cancer. *Cancer Manag Res.* (2020) 12:12557–67. doi: 10.2147/CMAR.S274986
 18. Yao LW, Wu LL, Zhang LH, Zhou W, Wu L, He K, et al. MFAP2 is overexpressed in gastric cancer and promotes motility via the MFAP2/integrin $\alpha 5 \beta 1$ /FAK/ERK pathway. *Oncogenesis.* (2020) 9:17. doi: 10.1038/s41389-020-0198-z
 19. Mei D, Zhu Y, Zhang L, Wei W. The role of CTHRC1 in regulation of multiple signaling and tumor progression and metastasis. *Mediators Inflamm.* (2020) 2020:9578701. doi: 10.1155/2020/9578701
 20. Ding X, Huang R, Zhong Y, Cui N, Wang Y, Weng J, et al. CTHRC1 promotes gastric cancer metastasis via HIF-1 α /CXCR4 signaling pathway. *Biomed Pharmacother.* (2020) 123:109742. doi: 10.1016/j.biopha.2019.109742
 21. Zhou H, Su L, Liu C, Li B, Li H, Xie Y, et al. CTHRC1 may serve as a prognostic biomarker for hepatocellular carcinoma. *Onco Targets Ther.* (2019) 12:7823–31. doi: 10.2147/OTT.S219429
 22. Zhang X, Zheng P, Li Z, Gao S, Liu G. The somatic mutation landscape and RNA prognostic markers in stomach adenocarcinoma. *Onco Targets Ther.* (2020) 13:7735–46. doi: 10.2147/OTT.S263733
 23. Song H, Liu L, Song Z, Ren Y, Li C, Huo J. P4HA3 is epigenetically activated by slug in gastric cancer and its deregulation is associated with enhanced metastasis and poor survival. *Technol Cancer Res Treat.* (2018) 17:1533033818796485. doi: 10.1177/1533033818796485
 24. Lin Y, Li B, Yang X, Cai Q, Liu W, Tian M, et al. Fibroblastic FAP promotes intrahepatic cholangiocarcinoma growth via MDSCs recruitment. *Neoplasia.* (2019) 21:1133–42. doi: 10.1016/j.neo.2019.10.005
 25. Coto-Llerena M, Ercan C, Kancherla V, Taha-Mehlitz S, Eppenberger-Castori S, Soysal SD, et al. High expression of FAP in colorectal cancer is associated with angiogenesis and immunoregulation processes. *Front Oncol.* (2020) 10:979. doi: 10.3389/fonc.2020.00979

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

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

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



Immune Checkpoint Gene Expression Profiling Identifies Programmed Cell Death Ligand-1 Centered Immunologic Subtypes of Oral and Squamous Cell Carcinoma With Favorable Survival

OPEN ACCESS

Edited by:

Thirumal Kumar D,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

E Srinivasan,
Saveetha University, India
Udhaya Kumar S,
Vellore Institute of Technology, India
Rituraj Purohit,
Institute of Himalayan Bioresource
Technology (CSIR), India

*Correspondence:

Yang Yu
dr_yangyu@163.com

[†]These authors have contributed
equally to this work and share first
authorship

[‡]These authors have contributed
equally to this work and share senior
authorship

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 16 August 2021

Accepted: 20 December 2021

Published: 20 January 2022

Citation:

Yu Y, Tang H, Franceschi D,
Mujagond P, Acharya A, Deng Y,
Lethaus B, Savkovic V, Zimmerer R,
Ziebolz D, Li S and Schmalz G (2022)
Immune Checkpoint Gene Expression
Profiling Identifies Programmed Cell
Death Ligand-1 Centered
Immunologic Subtypes of Oral and
Squamous Cell Carcinoma With
Favorable Survival.
Front. Med. 8:759605.
doi: 10.3389/fmed.2021.759605

Yang Yu^{1†}, Huiwen Tang^{1†}, Debora Franceschi², Prabhakar Mujagond³,
Aneesha Acharya⁴, Yupei Deng⁵, Bernd Lethaus⁶, Vuk Savkovic⁶, Rüdiger Zimmerer⁶,
Dirk Ziebolz⁷, Simin Li^{8*} and Gerhard Schmalz^{7*}

¹ Department of Stomatology, Qunli Branch, The First Affiliated Hospital of Harbin Medical University, Harbin, China,

² Department of Experimental and Clinical Medicine, University of Florence, Florence, Italy, ³ Regional Centre for
Biotechnology, 3rd Milestone Gurgaon-Faridabad Expressway, Faridabad, India, ⁴ Dr. D. Y. Patil Dental College and Hospital,

Dr. D. Y. Patil Vidyapeeth, Pune, India, ⁵ Laboratory of Molecular Cell Biology, China Tibetology Research Center, Beijing
Tibetan Hospital, Beijing, China, ⁶ Department of Cranio Maxillofacial Surgery, University Clinic Leipzig, Leipzig, Germany,

⁷ Department of Cariology, Endodontology and Periodontology, University of Leipzig, Leipzig, Germany, ⁸ Stomatological
Hospital, Southern Medical University, Guangzhou, China

Objective: This study aimed to identify the programmed death ligand-1 (PDL1, also termed as CD274) and its positively correlated immune checkpoint genes (ICGs) and to determine the immune subtypes of CD274-centered ICG combinations in oral and squamous cell carcinoma (OSCC).

Materials and Methods: Firstly, the 95 ICGs obtained *via* literature reviews were identified in the Cancer Genome Atlas (TCGA) database in relation to OSCC, and such 88 ICG expression profiles were extracted. ICGs positively correlated with CD274 were utilized for subsequent analysis. The relationship between ICGs positively correlated with CD274 and immunotherapy biomarkers (tumor mutation burden (TMB), and adaptive immune resistance pathway genes) was investigated, and the relationships of these genes with OSCC clinical features were explored. The prognostic values of CD274 and its positively correlated ICGs and also their associated gene pairs were revealed using the survival analysis.

Results: Eight ICGs, including CTLA4, ICOS, TNFRSF4, CD27, B- and T-lymphocyte attenuator (BTLA), ADORA2A, CD40LG, and CD28, were found to be positively correlated with CD274. Among the eight ICGs, seven ICGs (CTLA4, ICOS, TNFRSF4, CD27, BTLA, CD40LG, and CD28) were significantly negatively correlated with TMB. The majority of the adaptive immune resistance pathway genes were positively correlated with ICGs positively correlated with CD274. The survival analysis utilizing the TCGA-OSCC data showed that, although CD274 was not significantly associated with overall survival (OS), the majority of ICGs positively correlated with CD274

(BTLA, CD27, CTLA4, CD40LG, CD28, ICOS, and TNFRSF4) were significantly correlated with OS, whereby their low-expression predicted a favorable prognosis. The survival analysis based on the gene pair subtypes showed that the combination subtypes of CD274_low/BTLA_low, CD274_low/CD27_low, CD274_low/CTLA4_low, CD8A_high/BTLA_low, CD8A_high/CD27_low, and CD8A_high/CTLA4_low predicted favorable OS.

Conclusion: The results in this study provide a theoretical basis for prognostic immune subtyping of OSCC and highlight the importance of developing future immunotherapeutic strategies for treating oral cancer.

Keywords: PDL1, immune checkpoint genes, oral and squamous cell carcinoma, immune subtypes, prognosis

INTRODUCTION

Oral and squamous cell carcinoma (OSCC) is one of the most common oral cancers and is characterized by high morbidity and mortality (1). Currently applied treatment approaches include tumor resection, radiotherapy, and adjuvant chemotherapy but patients with OSCC continue to display an unsatisfactory prognosis after such routine therapy (2). The median survival period of patients with OSCC is 515 days, which is <1.5 years (3). This fact underscores a urgent need for innovative and effective therapeutic strategies for treating OSCC. Immunotherapy based on the drugs targeting immune checkpoint genes (ICGs) has gained considerable attention in recent years, with a surge in the development of novel immune strategies to treat and improve the survival of patients with cancer (4). Two main approaches have been proposed to enhance the antitumor immunity. The most well-investigated approach is the blockade of coinhibitory molecules with monoclonal Abs directed to T-cell surface ICG biomarkers, namely programmed cell death protein-1 (PD1)/programmed death ligand-1 (PDL1 also named as CD274), CTLA4, LAG3, TIM3, and B- and T-lymphocyte attenuator (BTLA) (5). The second approach is based on activating costimulatory molecules with agnostic Abs directed to T-cell surface ICG biomarkers, namely CD27, CD40, OX40, and CD137 (6). Most current research is focused on coinhibitory ICGs due to the fundamental safety challenges accompanying the triggering costimulatory immune pathways, as well as the dose-limiting toxicities of mAbs agents to costimulatory ICGs (7). Therefore, the primary focus of researchers has been directed at investigating coinhibitory ICGs, particularly PD1 and its ligand-PDL1 (8).

It is well-established that PDL1 expressed on the surface of cancer cells can bind with PD1 on the surface of T-cells, and the interaction between PDL1 and PD1 can inhibit the function of T-cells by inhibiting proliferation, promoting apoptosis, and inhibiting the cytokine secretion (9). The inhibitor drugs targeting PDL1, pembrolizumab and nivolumab, have been approved for use in many solid cancer treatments, including the first-line treatment for patients with recurrent or metastatic head and neck squamous cell carcinomas (HNSCC) (10). However, currently, there are no clinical trials of PDL1 inhibitors for treating OSCC. Prior clinical trials using a PDL1

inhibitor in HNSCC showed unsatisfactory response rates in an unselected population of patients with cancer, suggesting that not all patients respond well to the PDL1 inhibitor-based immunotherapy and a subset of patients are resistant to such immunotherapy (11). The purported cause underlying this phenomenon is to vary the expression levels of PDL1 among patients; therefore, patients with a high expression of PDL1 might achieve clinical responses, while those with a low expression are likely to be resistant. Another important influencing factor might be the density, composition, and activation state of the CD8+ effector T-cells, which play a central role in antitumor immunity. Patients with low infiltration of CD8+ T-cells might not respond to such immunotherapy (12). Based on these factors, there is an urgent need for the differentiation of patients with cancer into immune subtypes according to the expression level of PDL1 and the infiltration of CD8+ T-cells. The subtyping of an immune microenvironment in OSCC can be beneficial for identifying patients who may benefit from ICG-targeting therapies, and aid in optimizing the agent design for future clinical trials.

Furthermore, a few studies have shown that coinhibitory ICGs' inhibitor drugs, when administered as synergistic combination strategies, induced a dramatic increase in durable response rates as compared to monotherapy (13). The combination of PDL1 and its synergistic ICG blockers was shown to significantly increase the response rates and enhance the treatment efficacy in patients with cancer. For example, the combination of ipilimumab (anti-CTLA-4) plus nivolumab (anti-PD-1) was shown to significantly enhance efficacy in patients with metastatic melanoma; therefore, ipilimumab plus nivolumab was approved for the treatment of metastatic melanoma, advanced renal cell carcinoma, and metastatic colorectal cancer (14). The treatment efficacy of PDL1-centered ICG combination inhibitors has been reported in many research studies on melanoma, breast cancer, and lung cancer (15, 16). The success of this combination in other cancer types encourages its investigation in OSCC. While previous studies have addressed the synergistic combination of coinhibitory ICG inhibitors in cancer treatment, it has not yet been investigated in the context of OSCC.

To address this research gap, the present study utilized the Cancer Genome Atlas (TCGA) (17) and the Gene Expression

Omnibus (GEO) database (18). The TCGA is well-established as the most comprehensive cancer genomics program, which has produced, evaluated, and made public data related to the genomic sequencing, expression, methylation, and copy number variation of over 11,000 patients with cancer who have been diagnosed with more than 30 distinct forms of cancer (17). Many previous bioinformatics studies have followed the traditional study design of analyzing the TCGA data and followed by the verification of the computationally predicted results using GEO data sets (19–24). The GEO database is a freely accessible resource for the functional genomics data that contain original data sets from tens of thousands of published microarray or sequencing experiments (25). Because GEO data sets vary in the aspects of experimental design, country, race, laboratories, experimental platform, sample size, and disease severity, these multiple factors allow the GEO data sets to be heterogeneous. If the results determined based on the TCGA data can be validated using the heterogeneous GEO data sets, the predicted results can be considered reliable. Therefore, in the present investigation, four oral cancer GEO data sets were used as independent cohorts to validate the prognostic immune subtypes results obtained using the TCGA data.

Based on the clinical survival data and RNA expression data from the TCGA and GEO database, the present research aimed to identify the ICGs that work synergistically with PDL1 in OSCC, as well as the PDL1-centered ICGs combinations that have prognostic values for OSCC. The ICG inhibitor combinations and prognostic immune subtypes identified in this research could provide novel strategies for the immunotherapy of OSCC, and bear the potential for clinical translation.

MATERIALS AND METHODS

Study Design

The work flowchart is shown in **Figure 1**. In brief, 95 ICGs obtained from the literature review were mapped into the OSCC-TCGA data set, and 88 ICGs were found in the data set. The tumor-infiltrating immune cell (TIIC) analysis was performed based on the expression profiles of 88 ICGs in OSCC samples, and it included distribution proportion, a heat map analysis, and a correlation analysis. The correlation among the 88 ICGs in OSCC samples was investigated, particularly focusing on the ICGs positively correlated with CD274. Afterward, the relationship between ICGs positively correlated with CD274 and several immunotherapy-related aspects was investigated. Moreover, the OSCC sample subtypes with prognostic values were identified by investigating the prognostic values of the sole genes (CD274 and its positively correlated ICGs); the prognostic values of the combination subtypes defined by CD274 and its positively correlated ICGs; and the prognostic values of the combination subtypes defined by ICGs positively correlated with CD8A and CD274. To validate the prediction accuracy of the prognostic immune subtypes identified by the TCGA data analysis, four independent cohort data sets (i.e., GSE41613, GSE42743, GSE75538, and GSE85446) were used for verification.

Procurement of 95 ICGs by Literature Reviews

The ICG list, including 95 ICGs, was obtained by collecting the union of ICG list in several literatures investigating the involvement of ICGs in cancers (26–29). ICGs are displayed in **Supplementary Table S1**.

OSCC Data Downloading and Preprocessing

The HNSCC data from the TCGA database were downloaded from the University of California Santa Cruz's official webpage (<https://xenabrowser.net/datapages/>). The downloaded data regarding HNSCC consisted of gene expression RNAseq data, survival data, somatic mutation (SNPs and small INDELs) data, curated clinical data, and phenotype data. The OSCC data were collected by selecting the OSCC-related anatomic sites in the column of "site_of_resection_or_biopsy.diagnoses" of the excel file regarding the clinical data of HNSCC. The OSCC-related anatomic sites include buccal mucosa, retromolar trigone, alveolar ridge, the floor of the mouth, hard palate, oral cavity, gingiva (upper and lower), and oral tongue (anterior 2/3) (30). Afterward, the OSCC data were pre-processed by removing the samples without clinical information, particularly survival information, and the samples with the overall survival (OS) of <1 month, and the adjacent healthy control samples, as well as the samples in which genes' Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value was 0. After performing such preprocessing, the 326 OSCC samples were obtained. **Supplementary Table S2** presents the sample size belonging to the different anatomic sites of the 326 OSCC samples. As seen from **Supplementary Table S2**, the most frequent site is the tongue (42.33%); the overlapping lesion of the lip as the second frequent site (25.15%); and the floor of the mouth as the third frequent site (16.87%).

Procurement of the Expression Profiles of 88 ICGs in OSCC Samples

The ensemble IDs of genes in the expression profile of OSCC were converted to the gene SYMBOL by using the Bioconductor package (31). Regarding the genes that repeatedly appear after conversion, the average expression values of these genes were taken to ensure that the genes were unique in the expression profile. Afterward, 95 ICGs were mapped into the expression profile of OSCC, and 88 ICGs showed an expression in the TCGA-OSCC data set, and thus the expression profile of these 88 ICGs was obtained.

Analysis of TIICs in OSCC Samples

Twenty-two TIICs were obtained based on the CIBERSORT webtool (<https://cibersort.stanford.edu/>) (32). Firstly, the expression profiles of 88 ICGs in TCGA-OSCC samples were normalized, and the proportion of TIICs in OSCC samples was predicted using the CIBERSORT webtool (32). A total of 115 OSCC samples were obtained from 326 TCGA-OSCC samples by selecting the value of $p < 0.05$. Secondly, the expression levels of varying ICGs in each type of cell were obtained. The average

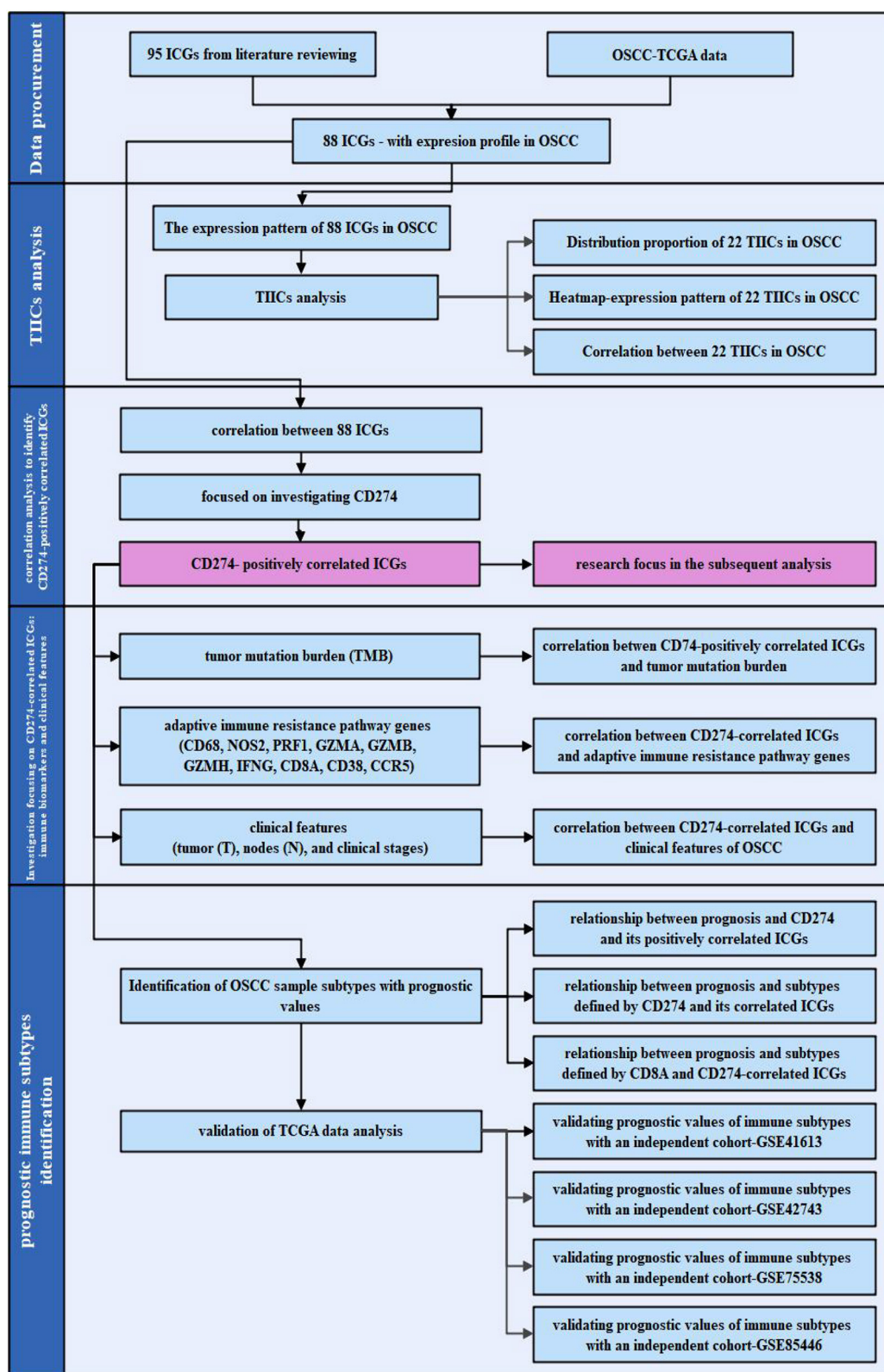


FIGURE 1 | The work flowchart of the present study. The flowchart is divided into four steps: data procurement for obtaining the 88 immune checkpoint genes (ICGs) with the expression profile in the Cancer Genome Atlas (TCGA) oral and squamous cell carcinoma (OSCC) data set; tumor-infiltrating immune cells (TIICs) analysis; a correlation analysis for identifying ICGs positively correlated with CD274; and subsequent analysis focusing on the ICGs positively correlated with CD274 particularly (Continued)

FIGURE 1 | from the aspect of the tumor mutation burden (TMB), adaptive immune resistance pathway, and clinical features; as well as the identification of the prognostic immune subtypes.

value of all ICGs in a certain type of cell was regarded as the expression level of this type of cell in samples. The heat map was plotted to show the expression levels of 22 TIICs in 115 OSCC samples.

Thirdly, a correlation plot was drawn based on the expression levels of TIICs in 115 samples to analyze the correlation between TIICs in the pathogenesis of OSCC. The Pearson correlation coefficient was used for calculating the correlation between any two types of TIICs. The correlation relationship between the two ICGs is represented by the letter r and quantified with a number, which varies between -1 and $+1$. Zero means that there is no correlation, whereas 1 means a complete or perfect correlation. The sign of r shows the direction of a correlation: a positive r means that the certain two ICGs were positively correlated and can play a synergistic role and vice versa. The interpretation of the Pearson's correlation coefficients value should be referred to the literature by the users' guide provided by Haldun Akoglu in 2018 (33). The $|r|$ value ≥ 0.8 indicates a very strong correlation; $0.5 \leq |r| < 0.8$ indicates a moderate correlation; $0.3 \leq |r| < 0.5$ indicates a fair correlation; and $|r| < 0.3$ indicates a poor correlation.

Heat Map Shows the Prognostic Values of 88 ICGs in OSCC

Eighty-eight ICGs with expression values in the TCGA-OSCC data set were obtained. Differentially expressed ICGs in the OSCC-TCGA data were identified by performing a differential expression analysis and using the edgeR package (version 3.14) (34) in the R software (version 3.6.3). Genes with $\log FC > 0$ and the value of $p < 0.05$ were regarded as upregulated differentially expressed genes (DEGs); while genes with $\log FC < 0$ and the value of $p < 0.05$ were regarded as downregulated DEGs (35). The relationship between these 88 ICGs and the OS of patients with OSCC was analyzed using a univariate Cox regression analysis ($\log\text{-rank } p < 0.05$). Heat maps were plotted using the pheatmap package (version 1.0.12) in R (36).

The Correlation Relationship Among 88 ICGs

To evaluate the correlation relationship among 88 ICGs, the Spearman algorithm was used to calculate the correlation of the expression value of 88 ICGs. The correlation relationship among 88 ICGs was displayed by using the corrplot package (version 0.92) in R (37). The correlation between PDL1 (CD274) and other ICGs was particularly marked as the research focus in subsequent analysis aimed to identify the ICGs, which were either positively (synergistic) or negatively (antagonistic) correlated with CD274.

Identification of Prognostic ICGs Positively Correlated With CD274

Among the 88 ICGs, ICGs, which were positively correlated to CD274 and had the value of $p < 0.05$, were selected. Based

on the expression values of these selected genes, the univariate Cox regression analysis was performed by using the survival package (version 3.2-13) in R (38). The relationship between the expression values of selected ICGs and prognosis was shown by the forest plot, which was plotted by using the forestplot package (version 2.0.1) in R (39). ICGs, which were not only positively correlated with CD274 but also had significant prognostic values, were identified and used as the investigation focus in subsequent analysis. Furthermore, the functional enrichment analysis was performed to determine the significant biological processes (BPs), and signaling pathways enriched by the ICGs were identified in the last step. The gene names of these ICGs were converted to the Entrez ID using the org.Hs.eg.db package (version 3.14) in R (40). The functional enrichment analysis was performed using the clusterProfiler package (version 3.14) in R (41). The species for the analysis was selected to be *Homo sapiens*. The GO terms, particularly BP and KEGG pathways that were significantly enriched by the correlated genes, were identified by setting the threshold value of $p < 0.05$. If there are more than 20 terms that were significantly enriched by this threshold setting, then only the top 20 terms ranked by the ascending order of the value of p were obtained to plot the bubble chart; otherwise, if there are < 20 terms that were significantly enriched by this threshold setting, then all of the terms were used for plotting the bubble chart. The bubble charts were plotted to visualize the enrichment results using the ggplot2 package (version 3.3.5) in R (42).

Relationship Between TMB and CD274-Related ICG Group

Among the 326 OSCC samples obtained by data preprocessing, 320 samples appeared to be with somatic mutation. The tumor mutation burden (TMB) values of these 320 samples were first calculated. The expression profile of CD274 and its positively correlated ICGs in these 320 samples were obtained. Based on the TMB values of 320 samples and the expression profile of the CD274-related ICGs group, the correlation analysis was performed to show the relationship between TMB and CD274-related ICGs group using the Spearman correlation method (43). The value of p showing a correlation was displayed with the radar chart using the fmsb package (version 0.7.2) in R (44). In addition, Spearman's RHO correlation values were calculated, and scatterplots were created to assess the correlation between CD274-related ICGs group and TMB in 320 samples. The scatter plots were drawn by using the ggplot2 package (version 3.3.5) in R (42). The edge density maps, which were located on both sides of the scatter plot, were drawn by using the ggMarginal function of the ggExtra package (version 0.9) in R (45). The correlation coefficient and the values of p were added to the scatter plots by using the stat_cor function of the ggpubr package (version 0.4.0) in R (46).

The Correlation Relationship Between Genes Involved in Adaptive Immune Resistance Pathway and Genes in the CD274-Related ICG Group

CD8⁺ T-cells can produce IFN γ and thus activate the immune pathways, leading to the upregulation of genes involved in an adaptive immune resistance pathway (e.g., CD68, NOS2, PRF1, GZMA, GZMB, GZMH, IFNG, CD8A, CD38, and CCR5). The correlation relationship between the genes involved in an adaptive immune resistance pathway and the genes in the CD274-related ICGs group was calculated by performing the Spearman correlation analysis based on the *cor.test* function in R. The Spearman correlation values and the values of *p* were obtained, based on which a heat map was plotted.

Relationship Between ICGs and Their Clinical Features

Based on the clinical information obtained from the TCGA database, the relationship between the CD274-correlated ICGs and their clinical features was analyzed, particularly focusing on the tumor (T), node (N), and clinical stages. Box plots were drawn by using the *ggboxplot* function in the *ggpubr* package of R (46). The significance of test was performed by using the *stat_compare_means* function in the *ggpubr* package of R (46). The Kruskal–Wallis algorithm was used to examine the values of *p*, showing the relationship between the gene expression level and its clinical features.

Relationship Between Prognosis and Subtypes Defined by CD274-Related ICG Group

The survival analysis was performed in three steps: (i) Firstly, to identify the relationship between the sole gene (CD274 and its positively correlated ICGs) and 5-year OS rate; (ii) Secondly, to identify the prognostic values of the combination of the genes consisted of CD274 and its positively correlated ICGs; (iii) Thirdly, to identify the prognostic values of the combination of the genes consisted of ICGs positively correlated with CD8A and CD274.

The survival analysis was performed by using the survival package in R (38). Kaplan–Meier (KM) plots were generated using the *survfit* function (version 3.2-13) in R (47); meanwhile, the values of *p* from log-rank tests were calculated. The clinical OSCC samples with the survival time of <5 years were obtained. By using the *coxph* function of the survival package in R, a Cox risk model was constructed for each gene (48). After then, the *predict* function was used for predicting the risk score of each model, and further the risk score of each sample was obtained. By taking the median value of the risk scores, the samples were divided into a high-expression group (H) and a low-expression group (L).

Regarding the second step of the survival analysis, the high-/low-expression groups of CD274 and its positively correlated ICGs were integrated. OSCC samples were divided into four categories to analyze the survival of gene classification samples.

Regarding the third step of the survival analysis, CD8A was selected among the adaptive immune resistance pathway genes, and it was focused in this section. The subtypes constituted by CD8A and each CD274-related ICG group gene were analyzed. Likewise, OSCC samples were divided into four categories to analyze the survival of gene classification samples.

Validating the Prognostic Values of the Immune Subtypes Identified by TCGA Data

To validate the prognostic values of the immune subtypes identified by the TCGA data, four OSCC-related GEO data sets [i.e., GSE41613 (49), GSE42743 (49), GSE75538 (50), and GSE85446] with the OS information were obtained, based on which the survival analysis was performed. The survival data of patients with OSCC within 5 years were obtained for an analysis. Firstly, the relationship between ICGs positively correlated with CD274 and OS within 5 years was validated. Secondly, the prognostic values of gene pair subtypes consisting of ICGs positively correlated with CD8A and CD274 were validated.

RESULTS

TIICs in OSCC Samples

A total of 115 OSCC samples were obtained by selecting samples with the value of *p* < 0.05 from the CIBERSORT webtool. The proportion of 22 TIICs in each sample is shown in **Figure 2A**. A heat map shows the expression levels of 22 TIICs in 115 OSCC samples (**Figure 2B**). As observed from **Figure 2B**, dendritic cells who were at rest were highly expressed in OSCC samples, and the other types of cells were downregulated or nearly nonexpressed in OSCC samples.

Figure 2C shows the correlation among TIICs in the pathogenesis of OSCC. Because ICGs act as an interaction between tumor cells and T-cells, the interpretation of the results of this correlation analysis should also be focused on T-cells, particularly CD8⁺ effector T-cells and regulatory T- (Treg-) cells. Thereby, CD8⁺ T-cells were significantly negatively correlated with macrophage M0 (Pearson correlation value = −0.58) and significantly positively correlated with activated dendritic cells (Pearson correlation value = 0.55); Treg cells were significantly negatively correlated with CD8⁺ T-cells (Pearson correlation value = −0.49).

The Expression Pattern and Correlation Relationship of ICGs in OSCC Samples

The expression level of 88 ICGs in OSCC samples is shown in **Figure 3**. **Supplementary Table S3** presents the expression level of 88 ICGs in oral tumor samples compared with healthy control oral samples. Among the 88 ICGs, 30 genes were found to be with the value of *p* < 0.05 and thus regarded as DEGs. **Supplementary Table S4** presents the differential expression information of these 30 DEGs in oral tumor samples compared with healthy control oral samples. **Figure 4A** shows that 88 ICGs were mainly positively correlated. Interestingly, the aggregation effect was obviously observed, which indicated that the relationship between ICGs is mainly synergistic. It can be clearly observed that all of the other ICGs were

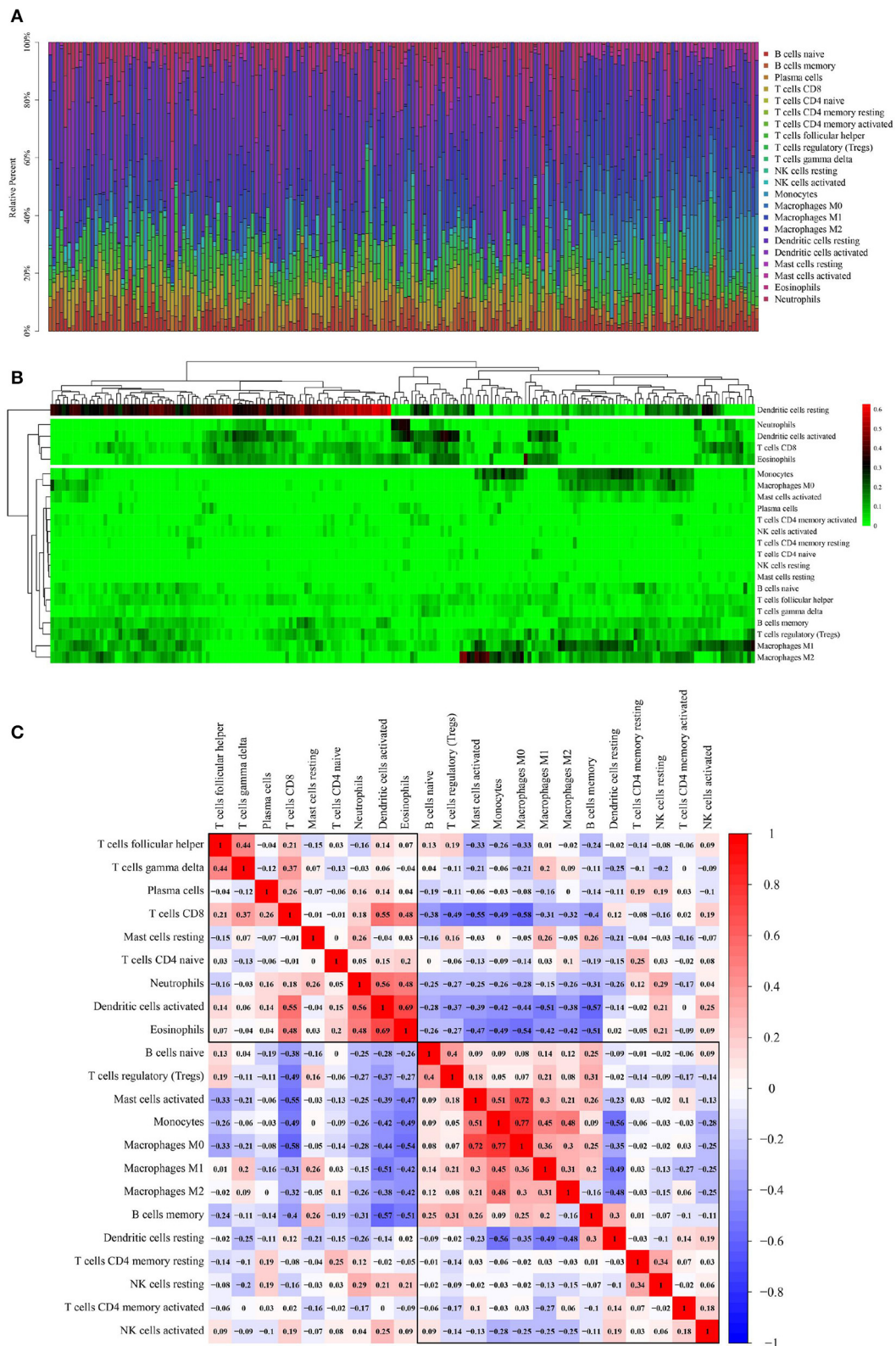


FIGURE 2 | Performance of CIBERSORT ascertained TILs in OSCC. **(A)** The distribution of 22 TILs in OSCC samples. X-axis represents the name of 115 samples, and y-axis represents the composition ratio of the cells in each sample. Different colors represent different types of cells. The longer column of each cell in a certain sample indicates that the proportion of this type of cell is higher in this sample. **(B)** A heat map of the 22 TILs proportions in 115 OSCC samples. Each column (Continued)

FIGURE 2 | represents a sample, and each row represents one type of immune cell population. The levels of the immune cell populations are shown in different colors, which transition from green to red with increasing proportions. Y-axis represents the expression levels of TIICs in each sample. In the color bar, green represents a low expression of TIICs in samples; red represents a high expression of TIICs in samples; and black represents that the TIICs were not expressed in the samples, meaning the expression level was 0. **(C)** A correlation matrix of 22 immune cell proportions and immune/stromal score in OSCC. Variables have been ordered by average linkage clustering. For comparison, the immune/stromal score has been rescaled to range between 0 and 1 separately in each study. The correlation between TIICs in the pathogenesis of OSCC. Both the x- and y-axis represent the 22 types of TIICs. The color bar shows the correlation value of TIICs. Blue indicated that the TIICs were negatively correlated, and red indicated that the TIICs were positively correlated. The darker color indicated that the correlation showed a higher significance. The diagonal line drawn from the coordinate (0,22) to the coordinate (22,0) has a correlation of 1.

positively correlated with CD274 except for four CD274-negatively correlated genes [i.e., DLX3 (Pearson correlation coefficient value $r = -0.11$), HHLA2 ($r = -0.09$), TNFRSF18 ($r = -0.14$), and VTCN1 ($r = -0.02$)]. This observation suggested that most ICGs play a coinhibitory role in tumor immunology and work synergistically with CD274, while the minority of ICGs play a costimulatory role in tumor immunology and work antagonistically with CD274. Because the correlation coefficient value $|r|$ between CD274 and its negatively correlated ICGs were very small even <0.2 showing a poor correlation, thus only ICGs, which were positively correlated with CD274, were included in subsequent analysis.

Identification of the Prognostic ICG That Is Significantly Positively Correlated With CD274

Among the 88 ICGs, 73 ICGs were selected based on the selection criteria: a positive correlation with CD274 and the value of $p < 0.05$. Based on the expression level of these 73 ICGs in the TCGA-OSCC data set and the prognosis of patients, **Figure 4B** shows the relationship between 73 ICGs and the prognosis using forest plots. The eight prognostic ICGs (CTLA4, ICOS, TNFRSF4, CD27, BTLA, ADORA2A, CD40LG, and CD28) with the value of $p < 0.05$ were marked with a five-pointed red star in **Figure 4B**. **Supplementary Table S5** listed the hazard ratio- (HR-) related parameters (i.e., HR, HR with a lower/higher 95% confidence index) and the values of p of eight ICGs. Subsequent analysis was focused on these eight ICGs.

Identification of Significantly Enriched Functional Terms of Eight ICGs

Figure 4C shows that the eight prognostic ICGs were significantly enriched in several TIICs related BPs, for example, T-cell-related BPs (e.g., T-cell proliferation, positive regulation of T-cell activation, T-cell costimulation, the regulation of T-cell activation, and T-cell activation), B-cell-related BPs (e.g., regulation of B-cell activation, and B-cell activation), lymphocyte-related BPs (e.g., the regulation of lymphocyte proliferation, lymphocyte costimulation, a positive regulation of lymphocyte activation, and the regulation of lymphocyte activation), and leukocyte-related BPs (e.g., a positive regulation of leukocyte cell-cell adhesion, leukocyte cell-cell adhesion, a positive regulation of leukocyte activation, the regulation of leukocyte cell-cell adhesion, and the regulation of leukocyte activation). In addition, the functional enrichment analysis results also revealed the significant KEGG pathways enriched by the eight prognostic ICGs, including the cytokine-cytokine

receptor interaction, an immune network for IgA production, cell adhesion molecules, and a T-cell receptor signaling pathway (**Figure 4D**).

Identification of TMB and Its Significantly Related ICGs From Eight ICGs

The correlation between TMB and eight ICGs identified in the abovementioned analysis was evaluated by using the Spearman correlation analysis. **Supplementary Table S6** presents that the correlations between the expression of TMB and all of the eight ICGs were negative (< 0), and with a statistical significance ($p < 0.05$) except for ADORA2A. The results in **Supplementary Table S6** are also displayed in the radar chart (**Figure 5A**) and also in the scatter plots (**Figure 5B**). These results showed that a high expression of these seven ICGs (CTLA4, ICOS, TNFRSF4, CD27, BTLA, CD40LG, and CD28) showing the worse/unfavorable prognosis corresponds to a low expression of TMB. TMB has been demonstrated to be a reliable biomarker for predicting the clinical efficacy of patients to PDL1 inhibitors (25). A low TMB predicts a poor response to PDL1 inhibitor therapy. The results in **Supplementary Table S6**, **Figures 5A,B** indicate that the patients with OSCC having a high expression of these seven ICGs are not suitable for the immune treatment of PDL1 inhibitor drugs and vice versa.

Identification of ICGs Correlated With CD274 and Significantly Related to Clinical Features

The box plots in **Figure 6** show the relationship between the clinical features and eight ICGs positively correlated with CD274. Regarding the expression level, the eight ICGs correlated with CD274 were significantly divided into a high- and a low-expression group. The high-expression group consisted of CTLA4, ICOS, TNFRSF4, CD27, and the low-expression group consisted of BTLA, ADORA2A, CD40LG, and CD28. As shown in **Figure 6A**, in the significance test, no significance was found between any of the eight ICGs with N (nodes) (Kruskal-Wallis test, $p > 0.05$). As shown in **Figure 6B**, a significance (Kruskal-Wallis test, $p = 0.028$) was found between CD40LG and the T stage. As shown in **Figure 6C**, a significance (Kruskal-Wallis test, $p = 0.038$) was found between CD40LG and the clinical stage. Taken together, the majority of ICGs positively correlated with CD274 were not significantly correlated with clinical features.

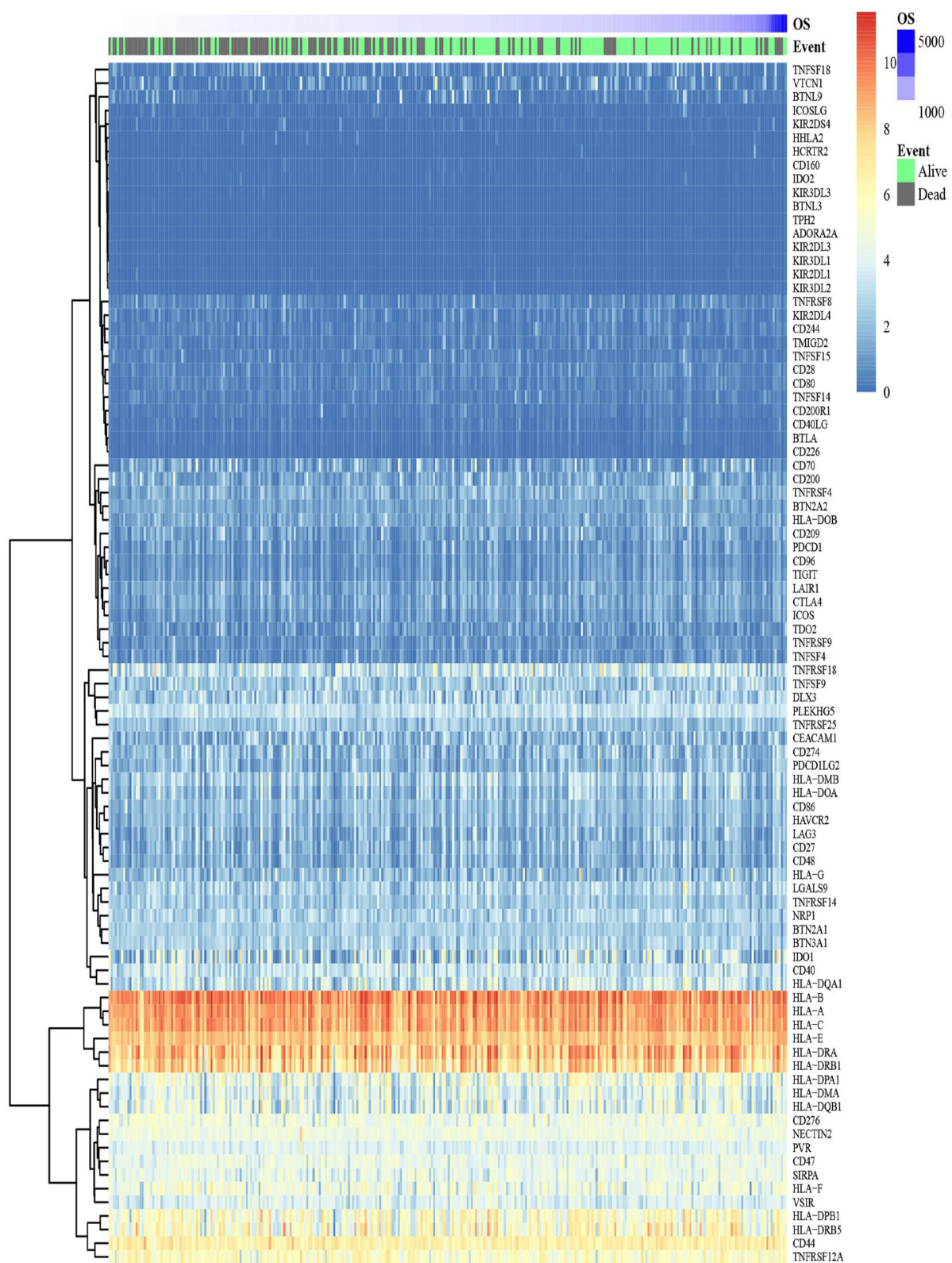
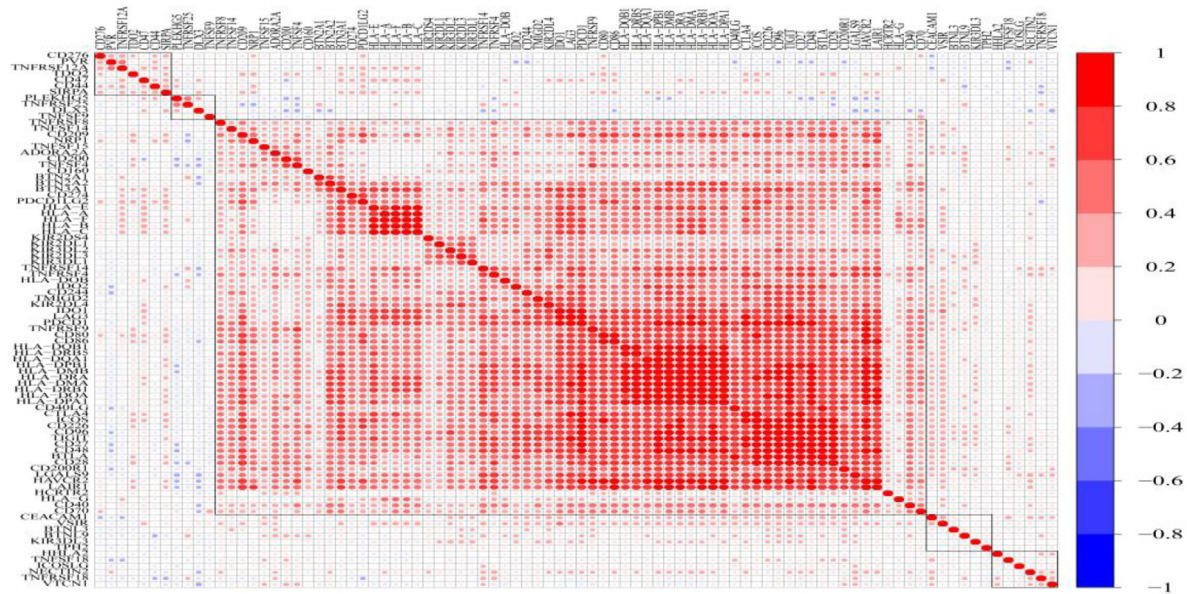
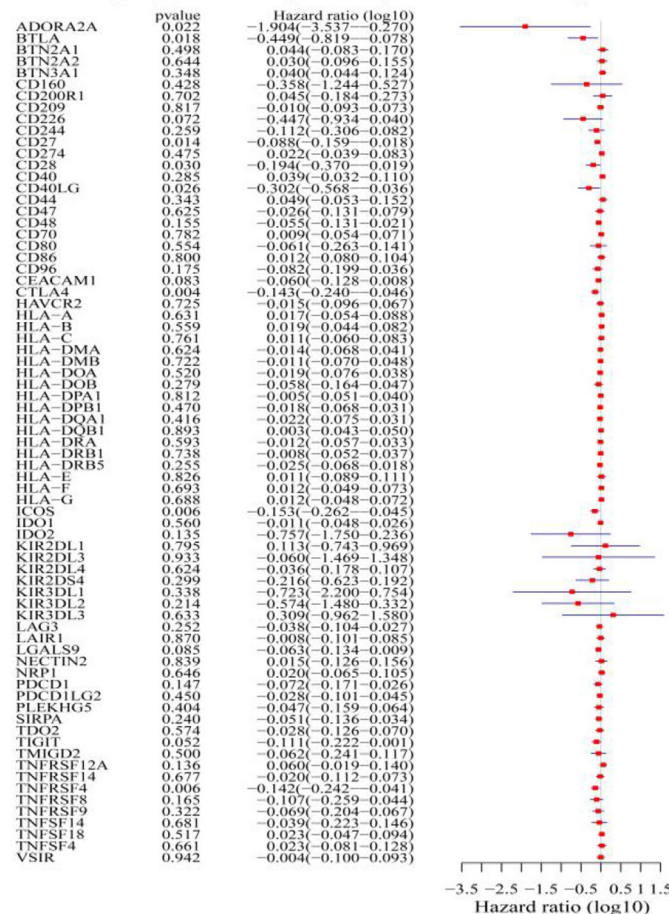


FIGURE 3 | A heat map shows the expression pattern of ICGs in the TCGA-OSCC data set. The color bar refers to the gene expression levels. Red indicates relatively higher gene expression levels, and blue indicates relatively low gene expression levels.

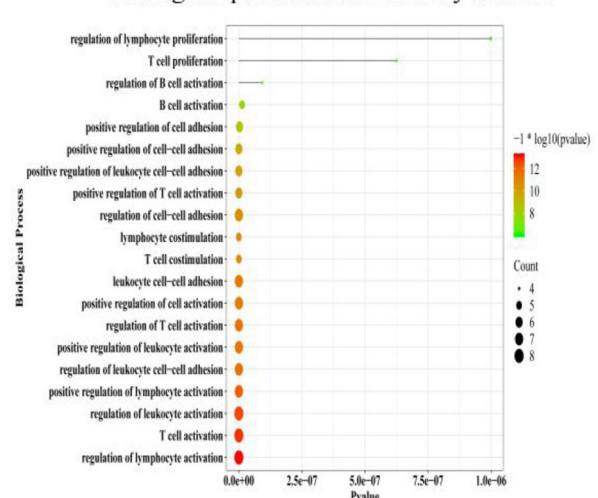
A Correlation between 95 ICGs



B Forest plot_73 CD274-positively correlated ICGs_OS



C Biological processes enriched by 8 ICGs



D KEGG pathways enriched by 8 ICGs

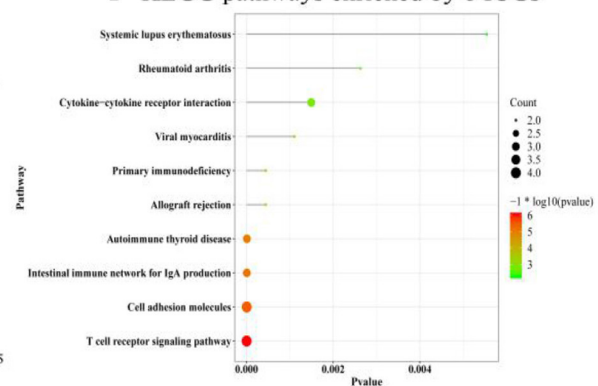


FIGURE 4 | The correlation between 95 ICGs and the prognostic values of 73 CD274-correlated ICGs. **(A)** The correlation between 95 ICGs. The index of the color bar indicates that the positively correlated ICGs (index > 0) and its negative correlation (index < 0). **(B)** The forest plot of log 10 hazard ratios (HRs) with 95% CIs shows the relationship between 73 CD274-positively correlated ICGs and overall survival (OS). The vertical line represents a HR of 0. **(C)** The Gene Ontology analysis (Continued)

FIGURE 4 | identifies the top 20 significant biological processes (BPs) enriched by the 8 prognostic ICGs positively correlated with CD274 (CTLA4, CD28, CD40LG, ADORA2A, B- and T-lymphocyte attenuator (BTLA), CD27, TNFRSF4, and ICOS). **(D)** Functional enrichment analysis identifies the top 20 significant KEGG signaling pathways enriched by the 8 prognostic ICGs positively correlated with CD274 (CTLA4, CD28, CD40LG, ADORA2A, BTLA, CD27, TNFRSF4, and ICOS).

A Correlation Between CD274-Related ICGs and Adaptive Immune Resistance Pathway Genes

Correlations between ICGs and 10 adaptive immune resistance pathway genes (CD68, NOS2, PRF1, GZMA, GZMB, GZMH, IFNG, CD8A, CD38, and CCR5) were analyzed. The heat map in **Figures 7A,B** show the correlation coefficients and $-\log_{10}$ (p -value) between 10 adaptive immune resistance pathway genes and 88 ICGs, respectively. In **Figures 7C,D**, a heat map is used to particularly show the correlation coefficients and $-\log_{10}$ (p -value) between 10 adaptive immune resistance pathway genes and CD274-related ICGs. As shown in **Figure 7**, the majority of adaptive immune resistance pathway genes (e.g., PRF1, GZMA, GZMB, GZMH, IFNG, CD8A, and CCR5) were positively correlated with the expression of the CD274-related ICGs and the majority of ICGs, whereas the three adaptive immune resistance pathway genes (e.g., NOS2, CD38, and CD68) were negatively correlated to the expression of nine CD274-related ICGs and the majority of ICGs. As observed in **Figures 7A,B**, almost all of the correlations were significant by performing the significance test of correlation coefficients (log-rank $p < 0.01$).

The Prognostic Values of the OSCC Subtypes Defined by ICGs Positively Correlated With CD274

Figure 8, **Supplementary Table S7** show a relationship between the OS rate and CD274-related ICGs (CD274 and its positively correlated ICGs). Although there is no significant relationship between CD274 and the OS of OSCC [CD274 ($p = 0.57 > 0.05$)], the significant prognostic values of ICGs positively correlated with CD274 were observed, for example, BTLA ($p = 0.0069 < 0.05$), CD27 ($p = 0.0056 < 0.05$), CTLA4 ($p = 0.0062 < 0.05$), CD40LG ($p = 0.046 < 0.05$), CD28 ($p = 0.049 < 0.05$), ICOS ($p = 0.041 < 0.05$), and TNFRSF4 ($p = 0.015 < 0.05$).

The Prognostic Values of the OSCC Subtypes Defined by CD274 and Its Positively Correlated ICGs

Oral and squamous cell carcinoma samples were divided into four combinations based on the median value of gene expression levels for eight pairs of genes (CD274-CTLA4, CD274-ICOS, CD274-TNFRSF4, CD274-CD27, CD274-BTLA, CD274-ADORA2A, CD274-CD40LG, and CD274-CD28). **Supplementary Table S8** presents that among these eight pairs of ICGs, only three pairs were found to be with significant prognostic values [CD274-BTLA ($p = 0.019 < 0.05$), CD274-CD27 ($p = 0.015 < 0.05$), and CD274-CTLA4 ($p = 0.0052 < 0.05$)]. **Figure 9A** shows the significance between the 5-year OS rate and the four high- and low-expression combinations

for three pairs of genes (CD274-BTLA, CD274-CD27, and CD274-CTLA4).

The Prognostic Values of the OSCC Subtypes Defined by CD8A and CD274-Related ICGs

Oral and squamous cell carcinoma samples were divided into four combinations based on the median value of gene expression levels for nine pairs of genes (CD8A-CD274, CD8A-CTLA4, CD8A-ICOS, CD8A-TNFRSF4, CD8A-CD27, CD8A-BTLA, CD8A-ADORA2A, CD8A-CD40LG, and CD8A-CD28). **Supplementary Table S8** presents that among these nine pairs of ICGs, only four pairs were found to be with significant prognostic values [CD8A-BTLA ($p = 0.019 < 0.05$), CD8A-CD27 ($p = 0.025 < 0.05$), CD274-CTLA4 ($p = 0.032 < 0.05$), and CD8A-TNFRSF4 ($p = 0.031 < 0.05$)]. To match the results in **Figure 9A**, only three pairs of genes (CD8A-BTLA, CD8A-CD27, and CD8A-CTLA4) were listed in **Figure 9B** to show the significance between the 5-year OS rate and the four high- and low-expression combinations corresponding to these three pairs of genes.

Validation of the Prognostic Values of the Immune Subtypes

Four independent cohorts (i.e., GSE41613, GSE42743, GSE75538, and GSE85446) in the GEO database were used to verify the prediction accuracy of the prognostic values of the immune subtypes identified by the survival analysis based on the TCGA data. **Supplementary Figures S1–S4** show the survival analysis results based on the four independent GEO data sets. These validation results were unfortunately not as anticipated and displayed insignificant prognostic values ($p > 0.05$) for the relationship between the OS and majority of immune subtype combinations identified by the TCGA-OSCC data set. This finding may be attributed to the small sample size of these GEO data sets as compared with the TCGA-OSCC data set [GSE41613: $n = 42$ samples, GSE42743 ($n = 52$), GSE75538 ($n = 8$), and GSE85446 ($n = 32$), and TCGA-OSCC data set ($n = 295$)]. As OSCC-related mRNA sequencing experiments are typically based on larger sample sizes, a sufficient documentation of the OS information should be encouraged in future research. Most of the existing GEO data sets lacked the survival information, and a few GEO data sets with the survival information had small sample sizes. Furthermore, other types of prognostic outcomes such as disease-free survival, metastasis-free survival, and progression-free survival were also lacking.

DISCUSSION

The present research aimed to characterize the prognostic subtypes of OSCC based on CD274 and its positively correlated

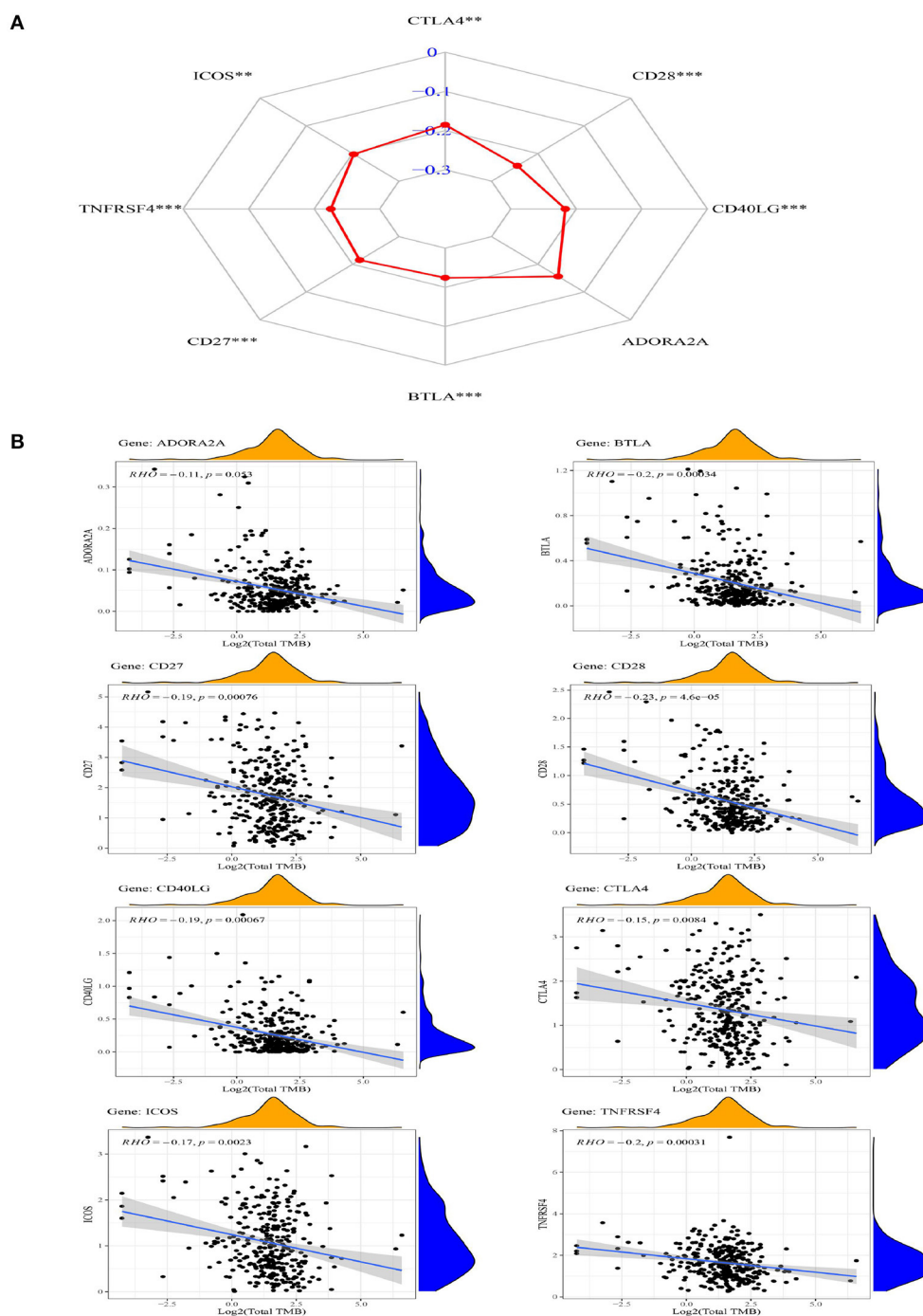


FIGURE 5 | The radar chart (A) and scatter plot (B) showing the relationship between the eight ICGs positively correlated with CD274 (CTLA4, CD28, CD40LG, ADORA2A, BTLA, CD27, TNFRSF4, and ICOS) and TMB.

ICGs. Using the mRNA expression data from TCGA and an independent cohort GEO data set, it was shown that eight ICGs (CTLA4, ICOS, TNFRSF4, CD27, BTLA, ADORA2A, CD40LG, and CD28) work synergistically and are positively correlated with CD274. An important finding was that although both CD8A and CD274 were not significantly related to OS, six subtypes with favorable survival were identified

based on the three ICGs positively correlated with CD274 (BTLA, CD27, and CTLA4). The six subtypes showing a better survival were CD274_low/BTLA_low, CD274_low/CD27_low, CD274_low/CTLA4_low, CD8A_high/BTLA_low, CD8A_high/CD27_low, and CD8A_high/CTLA4_low.

Based on the synergistic correlation and prognosis analysis, the three combination strategies of ICG inhibitors drugs may be

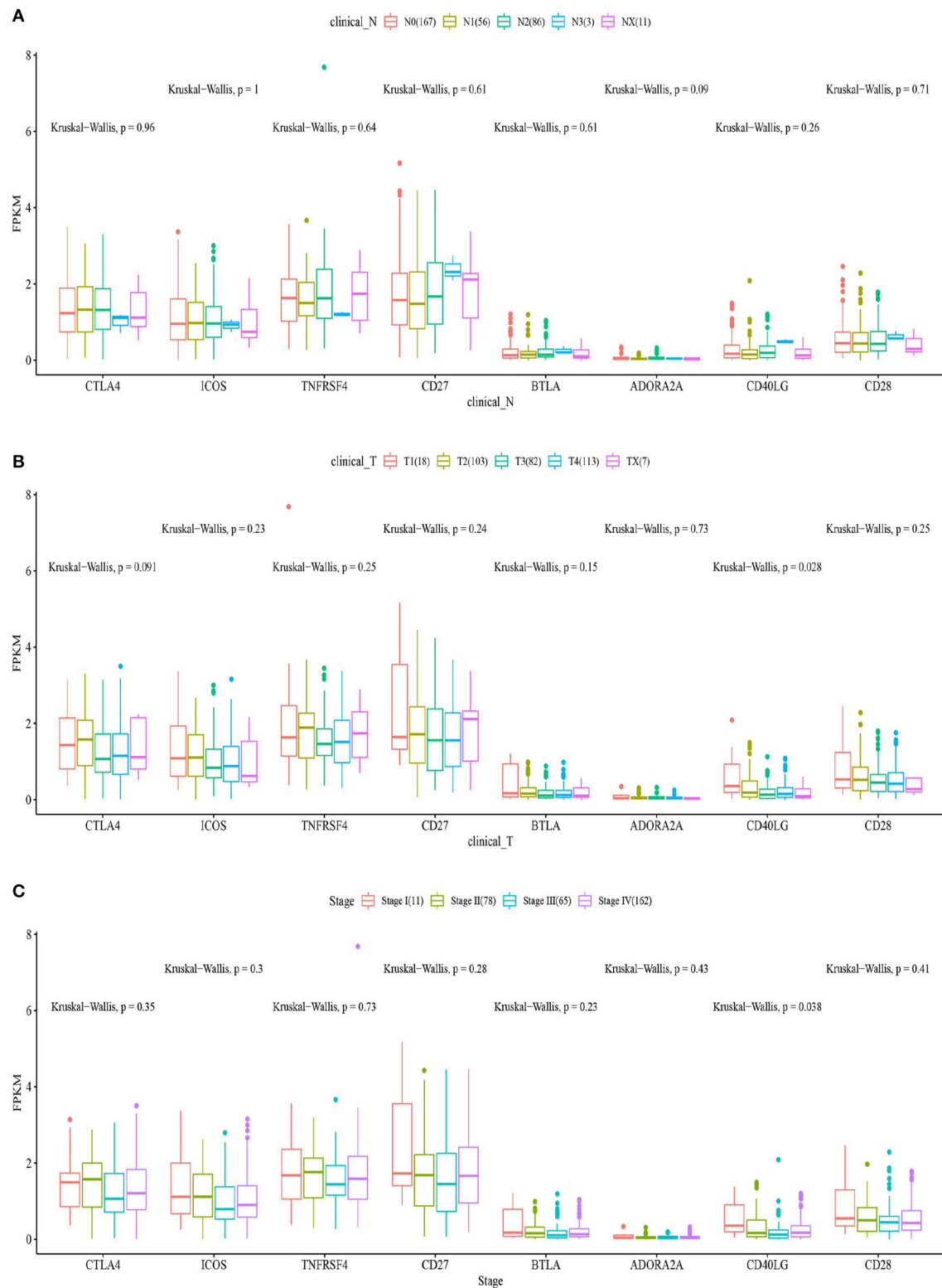


FIGURE 6 | The box plots show the eight ICGs positively correlated with CD274 and clinical features including N **(A)**, T **(B)**, and stages **(C)**.

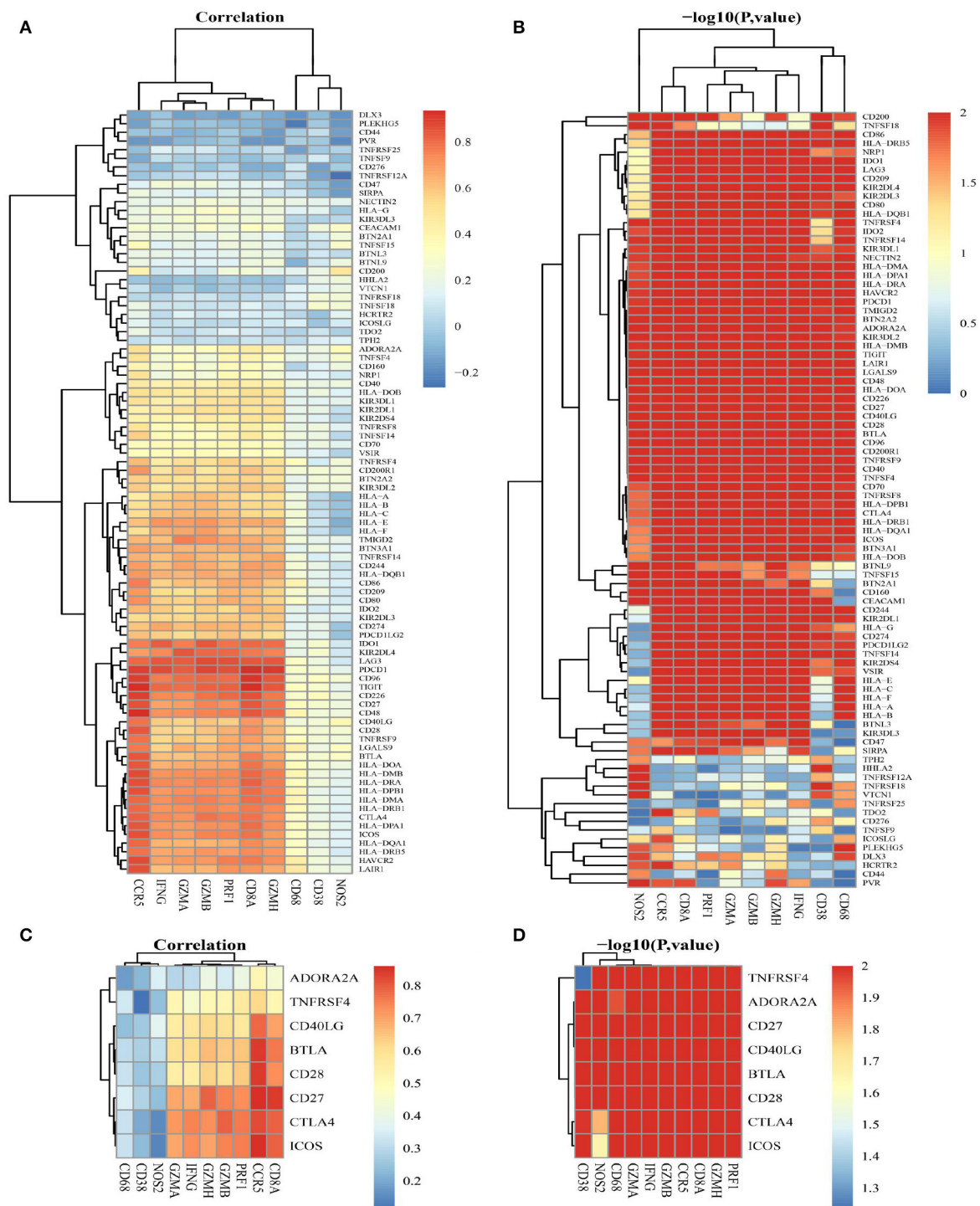
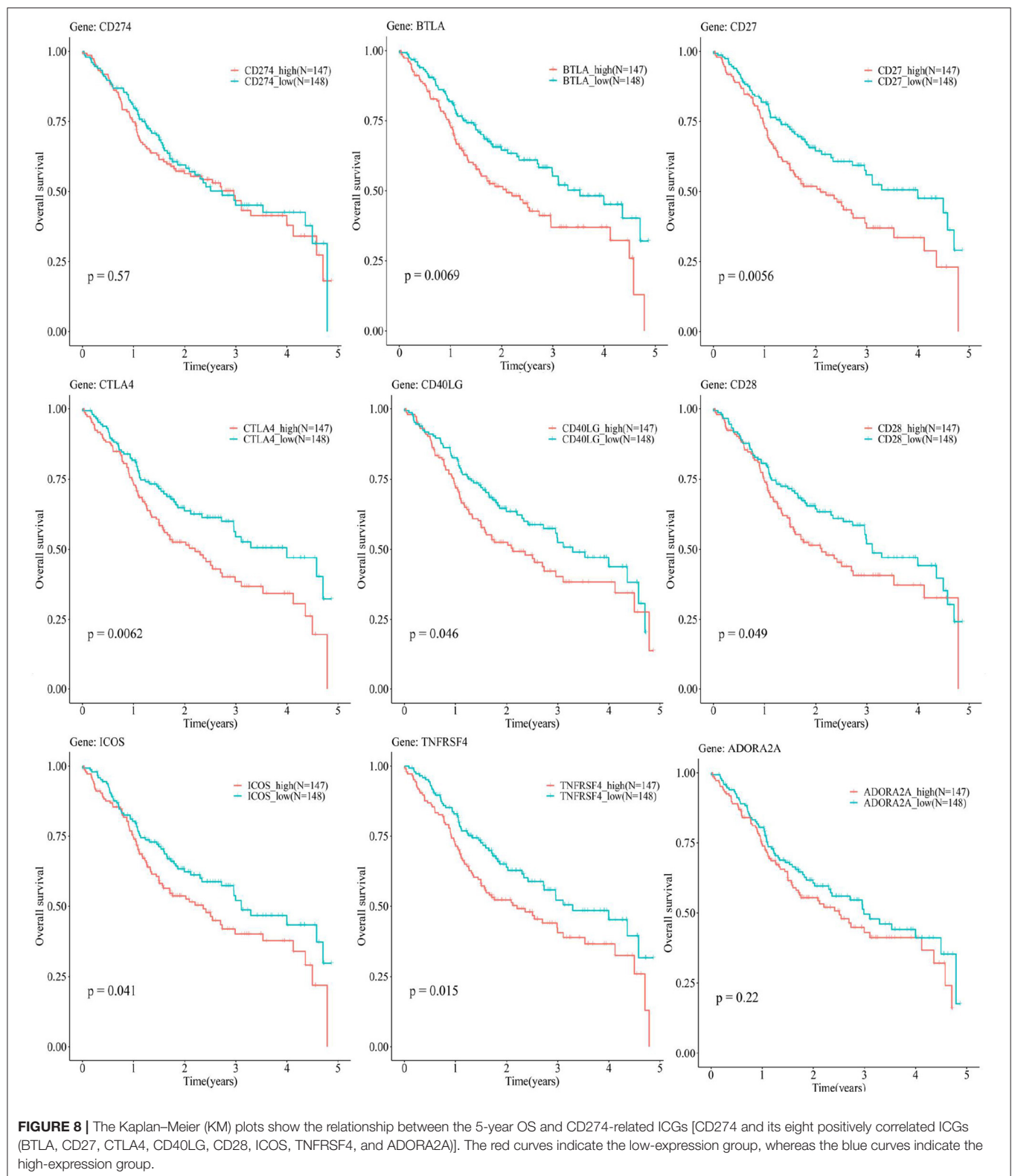
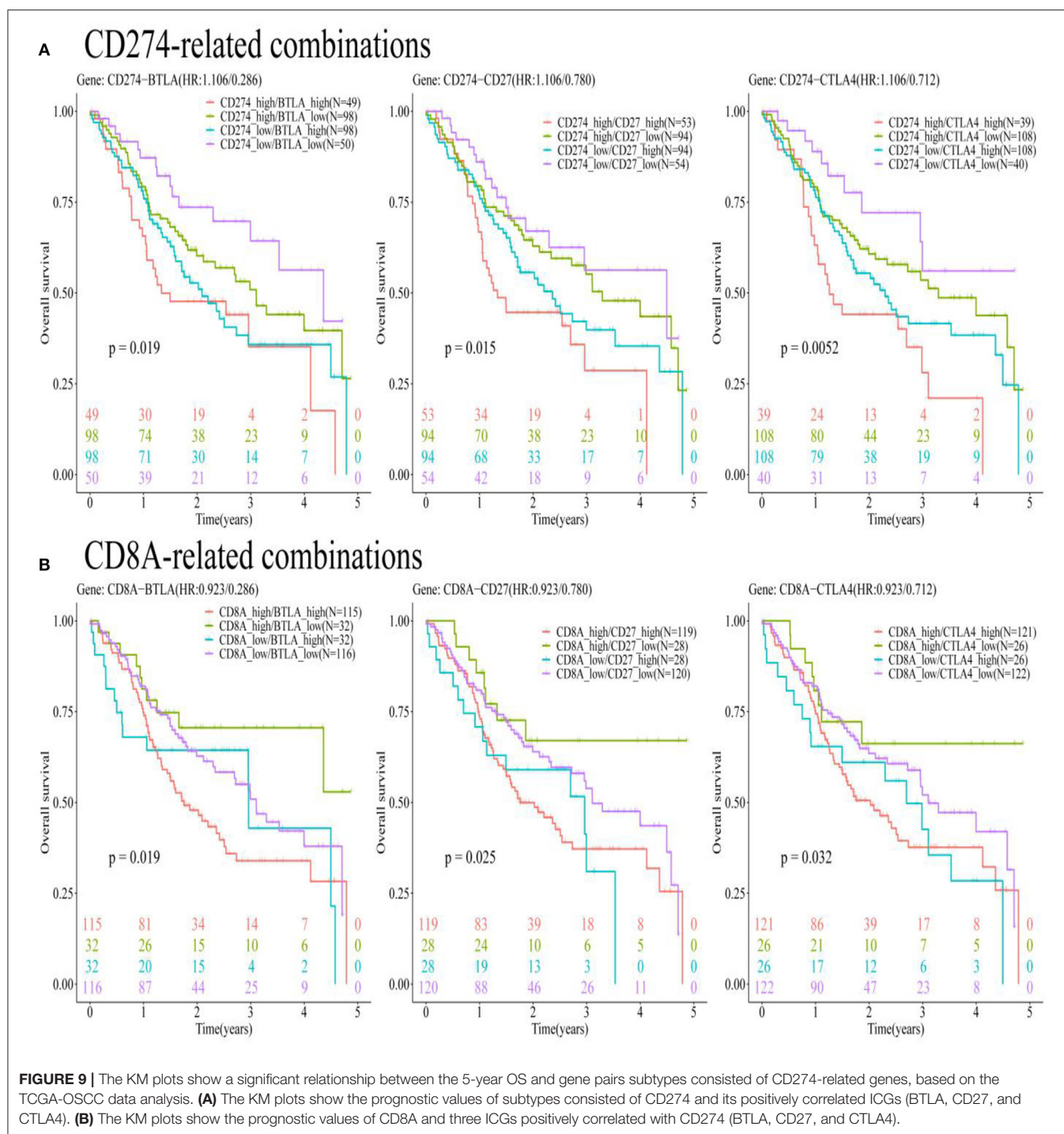


FIGURE 7 | The correlation between adaptive immune resistance pathway genes and ICGs. **(A)** A heat map shows the correlation coefficients between 10 adaptive immune resistance pathway genes and 88 ICGs. **(B)** A heat map shows $-\log_{10} p$ for the correlation between adaptive immune resistance pathway genes and 88 ICGs. **(C)** A heat map shows the correlation coefficients between adaptive immune resistance pathway genes and CD274-related ICGs (ADORA2A, TNFRSF4, CD40LG, BTLA, CD28, CD27, CTLA4, and ICOS). **(D)** A heat map shows $-\log_{10} p$ for the correlation between adaptive immune resistance pathway genes and CD274-related ICGs (ADORA2A, TNFRSF4, CD40LG, BTLA, CD28, CD27, CTLA4, and ICOS).



suggested, including, an inhibitor of PDL1+ inhibitor of BTLA, inhibitor of PDL1+ inhibitor of CD27, and inhibitor of PDL1+ inhibitor of CTLA. Much previous evidence supports the findings

of the current study. A high expression of BTLA in different types of cancers (e.g., colorectal cancer, melanoma cancer, and lung cancer) was found to inhibit the expression and function of



T-cells (51–53). Patients with lung cancer negative for both BTLA and PDL1 showed a better relapse-free survival (RFS) compared with patients, positive for either BTLA or PDL1 (51). Although BTLA and PDL1 employ distinct phosphatases to suppress T-cell signaling, both of them dampen the TCR and CD28 signaling pathways equally, and thus the inhibitors of both BTLA and PDL1 might be regarded as a combination of immunotherapeutic agents for cancer treatment (54, 55). However, the data regarding BTLA in oral cancer were still lacking. The TNFR superfamily

member CD27 showed a synergistic correlation with PDL1, and a low expression of CD27 indicated a significantly better survival outcome as compared with its high expression, and a low expression of both PDL1 and CD27 indicated the best survival. Contradictory results are reported in previous studies, showing CD27 as a costimulatory ICG, which plays critical roles in the activation, proliferation, and survival of T-cells, and that the blockade of PDL1 and agonist of CD27 activates CD8+ T-cell-driven antitumor immunity (56–58). Tumor heterogeneity and

different cancer types may underlie these contradictory findings. The present research showed that a low expression of CTLA4 indicated a better prognosis. Moreover, a low expression of both PDL1 and CTLA4 indicated an improved prognosis. This is in line with available research. The inhibitor of CTLA4-ipilimumab has been approved by FDA for treating melanoma (59). The mechanistic aspects of PDL1 and CTLA4 vary in immunology and PDL1 plays a suppressing role at the later stage of immune response, whereas CTLA4 plays an inhibiting role at the early stage of immune response (60). However, anti-PDL1 and anti-CTLA4 treatments have additive and synergistic effects on cancer treatment, based on the observation of the superior efficacy obtained by using the combination of CTLA4 blockade and PDL1 blockade in patients with melanoma compared with monotherapy (61).

In addition, the subtyping findings based on CD8A suggested that the coevaluation of the CD8A expression level and three ICGs positively correlated with CD274 (BTLA, CD27, and CTLA4) may enable a better evaluation of the immunological state of OSCC. The current study showed that patients with OSCC having the best survival had an increased CD8A infiltration and a low expression of one of the three ICGs positively correlated with CD274 (BTLA, CD27, and CTLA4). These results are reasonable considering CD8A is a surface biomarker of effector T-cells, and CD8+ T-cells in the TME indicate a good prognosis in many cancer types (62). The low expression of three ICGs positively correlated with CD274 (BTLA, CD27, and CTLA4) indicated a better prognosis compared to their high expression, indicating their coinhibitory roles in the tumor immunology of oral cancer. A previous study using the bioinformatic analysis also provided similar results, showing that patients with pancreatic adenocarcinoma (PDAC) with the best survival had increased CD8A infiltration without the expression of CD274 (63). Although the current study did not find a statistical significance of the relationship between CD8A/CD274 immunotypes and prognosis, significant prognostic values of three ICGs positively correlated with CD8A/CD274 (BTLA, CD27, and CTLA4) were found. A high expression of BTLA indicated a poor prognosis of OSCC, which is in accordance with the results of a previous study on ovarian cancer (64). A previous examination regarding melanoma showed that BTLA+ CD8+ tumor-infiltrating lymphocytes (TILs) showed a superior response and better survival compared to BTLA- CD8+ TILs (65) as BTLA plays a costimulatory role on activating CD8+ T-cells in melanoma. Such results were contradictory with the current research, which showed that the BTLA- CD8+ subset had the best survival in oral cancer, and might be explained by varying the roles of BTLA in different cancers. Previous studies have shown a costimulatory role of CD27 in inducing a potent proliferation of CD8+ T-cells, with a significant production of Th1 cytokines (IFN α , TNF α , and IL-2) and Th2 cytokines (mainly IL-13) by T-cells (7). This is contradictory with the results in the current computational prediction regarding oral cancer, which may be attributed to CD27 acting as either a costimulatory or coinhibitory receptor in different cancers and different circumstances (66). Considering the combination of CTLA4-CD8A, the present research showed

that patients with OSCC with high CD8A T-cell infiltration and a low expression of CTLA4 had the best prognosis. Previous research showed that the administration of CTLA4 blockade could increase the expansion and enhance the effector function of memory CD8+ T-cells, thereby contributing to the great accumulation of functional memory CD8+ T-cells (67). CTLA4+ tumor-infiltrating cells have been found to be an independent prognostic factor in OSCC, showing that its high infiltration indicated worse recurrence-free survival and metastasis-free survival (68).

It is important to clearly state the strengths and limitations of the current study. The main strength of the present study is that a series of comprehensive bioinformatics analyses was performed, including the analysis of TIICs, correlation analysis, TMB analysis, clinical feature relationship analysis, and survival analysis for identifying prognostic immune subtypes. The first limitation is the lack of experimental validation of the estimated synergistic effects of the administration of a PDL1 inhibitor and its positively correlated ICGs in OSCC. The second limitation is the lack of experimental validation of the prognostic values of the PDL1-based immune subtypes in OSCC. Both these aspects suggest the experimental design direction for future research. Another important limitation is the small sample size of the oral cancer-related data sets with the survival information for validation, which could be a reason for the insignificant results concerning the value of the immune subtypes. Although the findings of the analysis based on the TCGA data were statistically significant, the magnitude of their potential clinical effects must be recognized and could only be evaluated in clinical settings.

It is noteworthy to highlight the potential implications and the clinical transfer values of the current study. Firstly, the combination of PDL1-based ICG inhibitors might play additive and synergistic roles in the immune therapy of OSCC as compared to using either one of them alone. Thereby, the combinations identified in the current study might indicate novel therapeutic strategies for oral cancer treatment. Secondly, the immune subtypes identified in the current study could be used for predicting the OS outcomes of patients with oral cancer, and chairside testing based on these subsets could be developed as a useful prognostic prediction tool. Most importantly, the prognostic immune subtypes identified in the current study can have clinical implications for personalized immunotherapy. Patients belonging to specific subtypes should be administered with suitable ICG inhibitor agents with maximal efficacy. For example, considering the PDL1+CTLA4+ subset, this subgroup of patients with OSCC can receive a combination of a PDL1 inhibitor and CTLA4 inhibitor. However, the PDL1 and CTLA4 inhibitor agents will be not useful for the PDL1-CTLA4- subset. Therefore, immune subtypes as identified in this study can guide a refined patient selection, and enable personalized immune therapy strategies to significantly improve the OS in OSCC.

CONCLUSION

The current study identified several ICGs positively correlated with CD274 (BTLA, CD27, and CTLA4) comprising

immune subtypes indicating favorable OS outcomes, including CD274_low/BTLA_low, CD274_low/CD27_low, CD274_low/CTLA4_low, CD8A_high/BTLA_low, CD8A_high/CD27_low, and CD8A_high/CTLA4_low. These findings suggest that the three combinations of ICG inhibitors might play synergistic and additive effects in treating and improving the prognosis of OSCC, i.e., the combination of a CD274 inhibitor and BTLA inhibitor, the combination of a CD274 inhibitor and CD27 inhibitor, and the combination of a CD274 inhibitor and CTLA4 inhibitor. The combination of immune subtypes and suggested drug agents might provide precise immune strategies for application in personalized oral cancer treatment.

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

YY and HT contributed to the conceptualization, methodology, formal analysis, and writing—original draft. DF, PM, AA, and YD contributed to the data curation, formal analysis, methodology, resources, software, and visualization. BL, VS,

RZ, and DZ contributed to the formal analysis, methodology, writing—review and editing. YY, SL, and GS contributed to the project administration, supervision, writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

We appreciate the funding provided by the Science Research Cultivation Program of Stomatological Hospital, Southern Medical University (No. PY2020004) to support the postdoctoral research of the senior author Dr. SL (simin.li.dentist@gmail.com).

ACKNOWLEDGMENTS

We wish to acknowledge the bioinformatician Mrs. Xiangqiong Liu (dr.xqliu.bioinformatics@gmail.com) from China Tibetology Research Center, for her kind assistance on guiding us on the study design and bioinformatics analysis.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.759605/full#supplementary-material>

REFERENCES

- Bugshan A, Farooq I. Oral squamous cell carcinoma: metastasis, potentially associated malignant disorders, etiology and recent advancements in diagnosis. *F1000Research*. (2020) 9:229. doi: 10.12688/f1000research.22941.1
- Taghavi N, Yazdi I. Prognostic factors of survival rate in oral squamous cell carcinoma: clinical, histologic, genetic and molecular concepts. *Arch Iran Med*. (2015) 18:314–9.
- Le Campion A, Ribeiro CMB, Luiz RR, Da Silva Júnior FF, Barros HCS, Dos Santos KCB, et al. Low survival rates of oral and oropharyngeal squamous cell carcinoma. *Int J Dent*. (2017) 2017:5815493. doi: 10.1155/2017/5815493
- Dine J, Gordon R, Shames Y, Kasler MK, Barton-Burke M. Immune checkpoint inhibitors: an innovation in immunotherapy for the treatment and management of patients with cancer. *Asia Pac J Oncol Nurs*. (2017) 4:127–35. doi: 10.4103/apjon.apjon_4_17
- Qin S, Xu L, Yi M, Yu S, Wu K, Luo S. Novel immune checkpoint targets: moving beyond PD-1 and CTLA-4. *Mol Cancer*. (2019) 18:155. doi: 10.1186/s12943-019-1091-2
- Jeong S, Park SH. Co-stimulatory receptors in cancers and their implications for cancer immunotherapy. *Immune Netw*. (2020) 20:e3. doi: 10.4110/in.2020.20.e3
- Ramakrishna V, Sundarapandian K, Zhao B, Bylesjo M, Marsh HC, Keler T. Characterization of the human T cell response to *in vitro* CD27 costimulation with varilumab. *J Immunother Cancer*. (2015) 3:37. doi: 10.1186/s40425-015-0080-2
- Jiang Y, Chen M, Nie H, Yuan Y. PD-1 and PD-L1 in cancer immunotherapy: clinical implications and future considerations. *Hum Vaccin Immunother*. (2019) 15:1111–22. doi: 10.1080/21645515.2019.1571892
- Wu Y, Chen W, Xu ZP, Gu W. PD-L1 distribution and perspective for cancer immunotherapy—blockade, knockdown, or inhibition. *Front Immunol*. (2019) 10:2022. doi: 10.3389/fimmu.2019.02022
- Qiao XW, Jiang J, Pang X, Huang MC, Tang YJ, Liang XH, et al. The evolving landscape of PD-1/PD-L1 pathway in head and neck cancer. *Front Immunol*. (2020) 11:1721. doi: 10.3389/fimmu.2020.01721
- Wu X, Gu Z, Chen Y, Chen B, Chen W, Weng L, et al. Application of PD-1 blockade in cancer immunotherapy. *Comput Struct Biotechnol J*. (2019) 17:661–74. doi: 10.1016/j.csbj.2019.03.006
- Van Der Leun AM, Thommen DS, Schumacher TN. CD8(+) T cell states in human cancer: insights from single-cell analysis. *Nat Rev Cancer*. (2020) 20:218–32. doi: 10.1038/s41568-019-0235-4
- Barbari C, Fontaine T, Parajuli P, Lamichhane N, Jakubski S, Lamichhane P, et al. Immunotherapies and combination strategies for immuno-oncology. *Int J Mol Sci*. (2020) 21:5009. doi: 10.3390/ijms21145009
- Rotte A. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. *J Exp Clin Cancer Res*. (2019) 38:255. doi: 10.1186/s13046-019-1259-z
- Bu X, Yao Y, Li X. Immune checkpoint blockade in breast cancer therapy. *Adv Exp Med Biol*. (2017) 1026:383–402. doi: 10.1007/978-981-10-6020-5_18
- Chae YK, Arya A, Iams W, Cruz MR, Chandra S, Choi J, et al. Current landscape and future of dual anti-CTLA4 and PD-1/PD-L1 blockade immunotherapy in cancer; lessons learned from clinical trials with melanoma and non-small cell lung cancer (NSCLC). *J Immunother Cancer*. (2018) 6:39. doi: 10.1186/s40425-018-0349-3
- Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol*. (2015) 19:A68–77. doi: 10.5114/wo.2014.47136
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. (2002) 30:207–10. doi: 10.1093/nar/30.1.207
- Lan Q, Wang P, Tian S, Dong W. Mining TCGA database for genes of prognostic value in gastric cancer microenvironment. *J Cell Mol Med*. (2020) 24:11120–32. doi: 10.1111/jcmm.15595
- Ren N, Liang B, Li Y. Identification of prognosis-related genes in the tumor microenvironment of stomach adenocarcinoma by TCGA and GEO data sets. *Biosci Rep*. (2020) 40:BSR20200980. doi: 10.1042/BSR20200980

21. Xiao GF, Yan X, Chen Z, Zhang RJ, Liu TZ, Hu WL. Identification of a novel immune-related prognostic biomarker and small-molecule drugs in Clear Cell Renal Cell Carcinoma (ccRCC) by a merged microarray-acquired data set and TCGA database. *Front Genetics*. (2020) 11:810. doi: 10.3389/fgen.2020.00810
22. Zhao J, Guo C, Ma Z, Liu H, Yang C, Li S. Identification of a novel gene expression signature associated with the overall survival in patients with lung adenocarcinoma: a comprehensive analysis based on TCGA and GEO databases. *Lung Cancer*. (2020) 149:90–6. doi: 10.1016/j.lungcan.2020.09.014
23. Huang J, Liang B, Wang T. FOXD1 expression in head and neck squamous carcinoma: a study based on TCGA. GEO and meta-analysis. *Biosci Rep*. (2021) 41:BSR20210158. doi: 10.1042/BSR20210158
24. Wu Y, Deng J, Lai S, You Y, Wu J. A risk score model with five long non-coding RNAs for predicting prognosis in gastric cancer: an integrated analysis to combine TCGA and GEO data sets. *PeerJ*. (2021) 9:e10556. doi: 10.7717/peerj.10556
25. Clough E, Barrett T. The gene expression omnibus database. In: *Statistical Genomics*. New York, NY: Humana Press (2016). p. 93–110. doi: 10.1007/978-1-4939-3578-9_5
26. Ling B, Ye G, Zhao Q, Jiang Y, Liang L, Tang Q. Identification of an immunologic signature of lung adenocarcinomas based on genome-wide immune expression profiles. *Front Mol Biosci*. (2020) 7:603701. doi: 10.3389/fmolb.2020.603701
27. Long S, Li M, Liu J, Yang Y, Li G. Identification of immunologic subtype and prognosis of GBM based on TNFSF14 and immune checkpoint gene expression profiling. *Aging*. (2020) 12:7112. doi: 10.18632/aging.103065
28. Xu D, Liu X, Wang Y, Zhou K, Wu J, Cheng Chen J, et al. Identification of immune subtypes and prognosis of hepatocellular carcinoma based on immune checkpoint gene expression profile. *Biomed Pharmacother*. (2020) 126:109903. doi: 10.1016/j.biopha.2020.109903
29. Hu FF, Liu CJ, Liu LL, Zhang Q, Guo AY. Expression profile of immune checkpoint genes and their roles in predicting immunotherapy response. *Brief Bioinform*. (2021) 22:bbaa176. doi: 10.1093/bib/bbaa176
30. Lin NC, Hsien SI, Hsu JT, Chen MYC. Impact on patients with oral squamous cell carcinoma in different anatomical subsites: a single-center study in Taiwan. *Sci Rep*. (2021) 11:15446. doi: 10.1038/s41598-021-95007-5
31. Amezquita RA, Lun AT, Becht E, Carey VJ, Carp LN, Geistlinger L, et al. Orchestrating single-cell analysis with Bioconductor. *Nat Methods*. (2020) 17:137–45. doi: 10.1038/s41592-019-0654-x
32. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol*. (2018) 1711:243–59. doi: 10.1007/978-1-4939-7493-1_12
33. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med*. (2018) 18:91–3. doi: 10.1016/j.tjem.2018.08.001
34. Chen Y, Lun AT, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research*. (2016) 5:1438. doi: 10.12688/f1000research.8987.2
35. Li D, Yin Y, He M, Wang J. Identification of potential biomarkers associated with prognosis in gastric cancer via bioinformatics analysis. *Med Sci Monit*. (2021) 27:e929104. doi: 10.12659/MSM.929104
36. Kolde R, Kolde MR. Package 'pheatmap'. *R package*. (2015) 1:790.
37. Wei T, Simko V, Levy M, Xie Y, Jin Y, Zemla J. *Corrplot: Visualization of a Correlation Matrix*. *R package version 0.73*. (2013) 230, 11–15. Available Online at: <https://cran.r-project.org/web/packages/corrplot/corrplot.pdf>
38. Therneau TM, Lumley T. *Package Survival: Survival analysis*. Cham: CRAN (2014). Available Online at: <https://cran.r-project.org/web/packages/survival/survival.pdf>
39. Gordon M, Lumley T, Gordon MM. *Package 'forestplot'*. *Advanced Forest Plot Using 'grid'graphics*. The Comprehensive R Archive Network. Vienna (2019).
40. Carlson M. org. *Hs. eg. db: Genome Wide Annotation for Human*. *R package version 3.8.2*. (2019). Available online at: <https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html>
41. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*. (2021) 2:100141. doi: 10.1016/j.xinn.2021.100141
42. Villanueva RAM, Chen ZB. *ggplot2: Elegant Graphics for Data Analysis, 2nd Edn. Measurement: Interdisciplinary Research and Perspectives*. (2019) 17, 160–167. doi: 10.1080/15366367.2019.1565254
43. Kumar N, Nayak SB, Adiga M, Aithal AP. Determination of spearman correlation coefficient. *Dermatol Res Pract*. (2018) 2018:1–6. doi: 10.1155/2018/4512840
44. Nakazawa M, Nakazawa MM. *Package 'fmsb'*. (2019). Available online at: <https://cran.r-project.org/web/packages/fmsb/fmsb.pdf> (accessed December 13, 2021).
45. Attali D, Baker C. *ggExtra: Add Marginal Histograms to "ggplot2". And more "ggplot2" Enhancements*. (2018). Available online at: <https://cran.r-project.org/web/packages/ggExtra/readme/README.html>
46. Kassambara A, and Kassambara MA. *Package 'ggpubr'*. R package version 0.3.5. (2020). Available Online at: <https://cran.r-project.org/web/packages/ggpubr/ggpubr.pdf>
47. Shakur AH, Huang S, Qian X, Chang X. SURVFIT: doubly sparse rule learning for survival data. *J Biomed Informat*. (2021) 117:103691. doi: 10.1016/j.jbi.2021.103691
48. Fox J, and Weisberg S. *Cox Proportional-Hazards Regression for Survival Data. Appendix to an R and S-PLUS companion to applied regression*. Comprehensive R Archive Network (2002) 34, 12–17.
49. Lohavanichbutr P, Méndez E, Holsinger FC, Rue TC, Zhang Y, Houck J, et al. A 13-gene signature prognostic of HPV-negative OSCC: discovery and external validation. *Clin Cancer Res*. (2013) 19:1197–203. doi: 10.1158/1078-0432.CCR-12-2647
50. Krishnan NM, Dhas K, Nair J, Palve V, Bagwan J, Siddappa G, et al. A minimal DNA methylation signature in oral tongue squamous cell carcinoma links altered methylation with tumor attributes. *Mol Cancer Res*. (2016) 14:805–19. doi: 10.1158/1541-7786.MCR-15-0395
51. Li X, Xu Z, Cui G, Yu L, Zhang X. BTLA expression in Stage I-III non-small-cell lung cancer and its correlation with PD-1/PD-L1 and clinical outcomes. *Oncotargets Ther*. (2020) 13:215–24. doi: 10.2147/OTT.S232234
52. Song J, Wu L. Friend or foe: prognostic and immunotherapy roles of BTLA in colorectal cancer. *Front Mol Biosci*. (2020) 7:148. doi: 10.3389/fmolb.2020.00148
53. Ning Z, Liu K, Xiong H. Roles of BTLA in immunity and immune disorders. *Front Immunol*. (2021) 12:654960. doi: 10.3389/fimmu.2021.654960
54. Celis-Gutierrez J, Blattmann P, Zhai Y, Jarmuzynski N, Ruminski K, Grégoire C, et al. Quantitative interactomics in primary T cells provides a rationale for concomitant PD-1 and BTLA coinhibitor blockade in cancer immunotherapy. *Cell Rep*. (2019) 27:3315–30.e3317. doi: 10.1016/j.celrep.2019.05.041
55. Xu X, Hou B, Fulzele A, Masubuchi T, Zhao Y, Wu Z, et al. PD-1 and BTLA regulate T cell signaling differentially and only partially through SHP1 and SHP2. *J Cell Biol*. (2020) 219:e201905085. doi: 10.1083/jcb.201905085
56. Thomas LJ, He L-Z, Wasiuk A, Gergel LE, Boyer JM, Round SM, et al. Synergistic anti-tumor activity of PD-1 signaling blockade and CD27 costimulation correlates with enhanced ratio of effector to regulatory T cells at the tumor site. *Cancer Res*. (2015) 75:253. doi: 10.1158/1538-7445.AM2015-253
57. Buchan SL, Fallatah M, Thirdborough SM, Taraban VY, Rogel A, Thomas LJ, et al. PD-1 blockade and CD27 Stimulation activate distinct transcriptional programs that synergize for CD8(+) T-Cell-driven antitumor immunity. *Clin Cancer Res*. (2018) 24:2383–94. doi: 10.1158/1078-0432.CCR-17-3057
58. Vitale LA, He LZ, Thomas LJ, Wasiuk A, O'Neill T, Widger J, et al. Development of CDX-527: a bispecific antibody to combine PD-1 blockade and CD27 costimulation for cancer immunotherapy. *Cancer Immunol Immunother*. (2020) 69:2125–37. doi: 10.1007/s00262-020-02610-y
59. Savoia P, Astrua C, Fava P. Ipilimumab (Anti-Ctla-4 Mab) in the treatment of metastatic melanoma: effectiveness and toxicity management. *Human Vaccines Immunotherap*. (2016) 12:1092–101. doi: 10.1080/21645515.2015.1129478
60. Buchbinder EI, Desai A. CTLA-4 and PD-1 pathways: similarities, differences, and implications of their inhibition. *Am J Clin Oncol*. (2016) 39:98–106. doi: 10.1097/COC.0000000000000239
61. Callahan MK, Wolchok JD. At the bedside: CTLA-4- and PD-1-blocking antibodies in cancer immunotherapy. *J Leukoc Biol*. (2013) 94:41–53. doi: 10.1189/jlb.1212631
62. Maimela NR, Liu S, Zhang Y. Fates of CD8+ T cells in tumor microenvironment. *Comput Struct Biotechnol J*. (2019) 17:1–13. doi: 10.1016/j.csbj.2018.11.004

63. Danilova L, Ho WJ, Zhu Q, Vithayathil T, De Jesus-Acosta A, Azad NS, et al. Programmed Cell Death Ligand-1 (PD-L1) and CD8 expression profiling identify an immunologic subtype of pancreatic ductal adenocarcinomas with favorable survival. *Cancer Immunol Res.* (2019) 7:886–95. doi: 10.1158/2326-6066.CIR-18-0822
64. Chen YL, Lin HW, Chien CL, Lai YL, Sun WZ, Chen CA, et al. BTLA blockade enhances Cancer therapy by inhibiting IL-6/IL-10-induced CD19(high) B lymphocytes. *J Immunother Cancer.* (2019) 7:313. doi: 10.1186/s40425-019-0744-4
65. Ritthipichai K, Haymaker CL, Martinez M, Aschenbrenner A, Yi X, Zhang M, et al. Multifaceted role of BTLA in the control of CD8(+) T-cell Fate after antigen encounter. *Clin Cancer Res.* (2017) 23:6151–6164. doi: 10.1158/1078-0432.CCR-16-1217
66. O'Neill RE, Cao X. Co-stimulatory and co-inhibitory pathways in cancer immunotherapy. *Adv Cancer Res.* (2019) 143:145–94. doi: 10.1016/bs.acr.2019.03.003
67. Pedicord VA, Montalvo W, Leiner IM, Allison JP. Single dose of anti-CTLA-4 enhances CD8+ T-cell memory formation, function, and maintenance. *Proc Natl Acad Sci USA.* (2011) 108:266–71. doi: 10.1073/pnas.1016791108
68. Koike K, Dehari H, Ogi K, Shimizu S, Nishiyama K, Sonoda T, et al. Prognostic value of FoxP3 and CTLA-4 expression in patients with oral squamous

cell carcinoma. *PLoS One.* (2020) 15:e0237465. doi: 10.1371/journal.pone.0237465

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

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

Copyright © 2022 Yu, Tang, Franceschi, Mujagond, Acharya, Deng, Lethaus, Savkovic, Zimmerer, Ziebolz, Li and Schmalz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Transcriptome-Based Molecular Networks Uncovered Interplay Between Druggable Genes of CD8⁺ T Cells and Changes in Immune Cell Landscape in Patients With Pulmonary Tuberculosis

OPEN ACCESS

Edited by:

D. Thirumal Kumar,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Rajasekhar Chikati,
Yogivemana University, India
Nuzhath Fatima,
Jazan University, Saudi Arabia

*Correspondence:

Babajan Babanaganapalli
bbabajan@kau.edu.sa
Noor Ahmad Shaik
nshaik@kau.edu.sa
Venkatesh Vaidyanathan
v.vaidyanathan@auckland.ac.nz

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 10 November 2021

Accepted: 20 December 2021

Published: 07 February 2022

Citation:

Alsulaimany FA, Zaberemawi NMO,
Almukadi H, Parambath SV, Shetty PJ,
Vaidyanathan V, Elango R,
Babanaganapalli B and Shaik NA
(2022) Transcriptome-Based
Molecular Networks Uncovered
Interplay Between Druggable Genes
of CD8⁺ T Cells and Changes in
Immune Cell Landscape in Patients
With Pulmonary Tuberculosis.
Front. Med. 8:812857.
doi: 10.3389/fmed.2021.812857

Faten Ahmad Alsulaimany¹, Nidal M. Omer Zaberemawi¹, Haifa Almukadi²,
Snijesh V. Parambath³, Preetha Jayasheela Shetty⁴, Venkatesh Vaidyanathan^{5*},
Ramu Elango^{6,7}, Babajan Babanaganapalli^{6,7*} and Noor Ahmad Shaik^{6,7*}

¹ Department of Biology, Faculty of Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, ² Department of Pharmacology and Toxicology, Faculty of Pharmacy, King Abdulaziz University, Jeddah, Saudi Arabia, ³ Division of Molecular Medicine, St. John's Research Institute, Bangalore, India, ⁴ Department of Biomedical Sciences, College of Medicine, Gulf Medical University, Ajman, United Arab Emirates, ⁵ Auckland Cancer Society Research Centre (ACSRC), Faculty of Medical and Health Sciences (FM&HS), The University of Auckland, Auckland, New Zealand, ⁶ Princess Al-Jawhara Al-Brahim Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia, ⁷ Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

Background: Tuberculosis (TB) is a major infectious disease, where incomplete information about host genetics and immune responses is hindering the development of transformative therapies. This study characterized the immune cell landscape and blood transcriptomic profile of patients with pulmonary TB (PTB) to identify the potential therapeutic biomarkers.

Methods: The blood transcriptome profile of patients with PTB and controls were used for fractionating immune cell populations with the CIBERSORT algorithm and then to identify differentially expressed genes (DEGs) with R/Bioconductor packages. Later, systems biology investigations (such as semantic similarity, gene correlation, and graph theory parameters) were implemented to prioritize druggable genes contributing to the immune cell alterations in patients with TB. Finally, real time-PCR (RT-PCR) was used to confirm gene expression levels.

Results: Patients with PTB had higher levels of four immune subpopulations like CD8⁺ T cells ($P = 1.9 \times 10^{-8}$), natural killer (NK) cells resting ($P = 6.3 \times 10^{-5}$), monocytes ($P = 6.4 \times 10^{-6}$), and neutrophils ($P = 1.6 \times 10^{-7}$). The functional enrichment of 624 DEGs identified in the blood transcriptome of patients with PTB revealed major dysregulation of T cell-related ontologies and pathways ($q \leq 0.05$). Of the 96 DEGs shared between transcriptome and immune cell types, 39 overlapped with TB meta-profiling genetic signatures, and their semantic similarity analysis with the remaining 57 genes, yielded 45 new candidate TB markers. This study identified 9 CD8⁺ T cell-associated genes (*ITK*, *CD2*, *CD6*, *CD247*, *ZAP70*, *CD3D*, *SH2D1A*, *CD3E*,

and *IL7R*) as potential therapeutic targets of PTB by combining computational druggability and co-expression ($r^2 \geq |0.7|$) approaches.

Conclusion: The changes in immune cell proportion and the downregulation of T cell-related genes may provide new insights in developing therapeutic compounds against chronic TB.

Keywords: *Mycobacterium tuberculosis*, gene express profile, drug target, CD8⁺T cells, immune pathways

INTRODUCTION

Tuberculosis (TB) is a chronic infectious lung disease caused by pathogenic *Mycobacterium tuberculosis* (*MTB*) belonging to the *Mycobacteriaceae* family. Despite the widespread use of antibiotics and live attenuated vaccine, TB remains to be the major cause of morbidity and death among all bacterial diseases (1). This is primarily due to the rapid emergence of drug-resistant *MTB* strains and the incomplete knowledge of complex host-pathogen interactions (2). In the initial stages of infection, *MTB* invades and replicates in the macrophages after reaching the alveolar air sacs of the lungs (3). Granulomas, hallmark of TB, are formed around the infected macrophages by the organized aggregation of immune cells (like T and B lymphocytes), multinucleated giant cells, dendritic cells, and fibroblasts. Granulomas also suppress the host immune responses, as dendritic cells and macrophages were unable to present antigen to lymphocytes (4). It is noteworthy to mention that mycobacteria can induce distinct host responses from asymptomatic conditions to severe pulmonary illness (5). However, underlying immune cell types and their association with the differentially expressed genes in TB and how they contribute to severe infection are not yet fully explored.

Over the past few decades, microarray-based genome-wide RNA profiling has evolved as a powerful approach to investigate the host transcriptional response (of ~19,000 genes) in infectious diseases (6). However, differences in the type of clinical samples, array platforms, and statistical approaches used, created a discordance in interpreting massive transcriptomics data. Advances in statistical modeling and bioinformatics approaches have accelerated the identification of disease-centric genes by employing gene networking methods based on graph topological parameters for many infectious diseases (7, 8). Moreover, the new bioinformatic methods like estimating relative subsets of RNA transcripts (CIBERSORT), Tumor Immune Estimation Resource (TIMER), and Estimating the Proportions of Immune and Cancer cells (EPIC) are developed to characterize immune cell composition using large-scale gene expression data (9, 10). These bioinformatic methods implement functional enrichment scores based on the presence of the query genes over reference gene sets. They perform variety of biological analyses including immune responses based on the defined gene sets. Exploring abnormal immune cell infiltration is critical for developing novel transformative therapies to combat diseases such as cancer, myocarditis, and TB (11, 12). Therefore, in order to characterize alterations in immune cell proportion landscape and transcriptomic profile, and to identify new molecular

therapeutic targets, this study applied statistical and knowledge-based systemic investigations (such as semantic similarity, gene correlation, and graph theory parameters) to the blood transcription data of patients with TB.

MATERIALS AND METHODS

Study Design and Global Expression Data

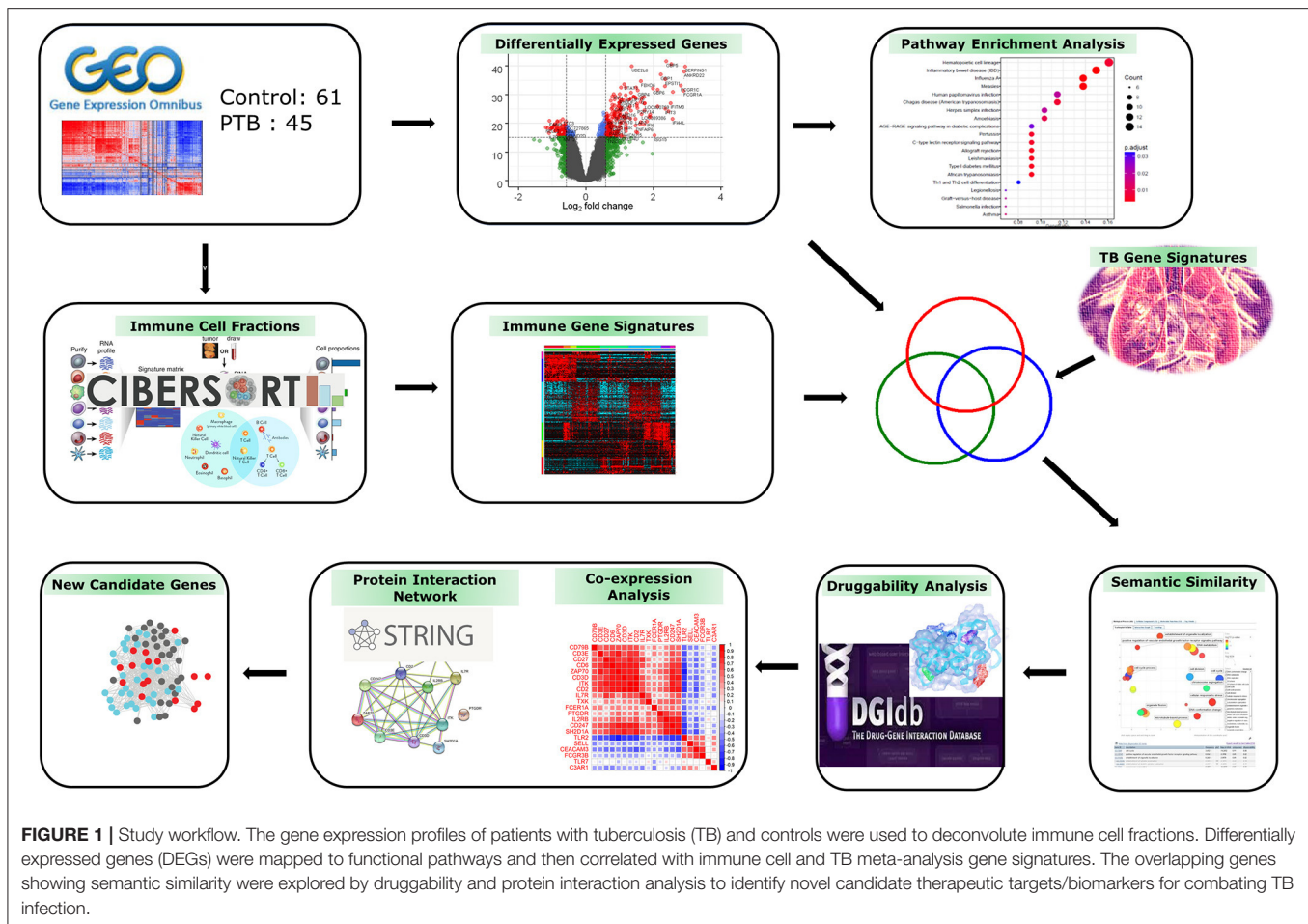
The genome-wide gene expression dataset (GSE83456) (13) was imported in raw format from the Gene Expression Omnibus (GEO) database (www.ncbi.nlm.nih.gov/geo). This dataset has expression profiles of 45 pulmonary TB (PTB) and 61 control blood samples generated on the Illumina Human HT-12 V4.0 expression bead chip (Illumina, Inc, USA). The detailed sample information is given in the **Supplementary Material Table 1**. **Figure 1** depicts the overall work design employed in the current research analysis.

Global Data Preprocessing and Screening of Differentially Expressed Genes

R/Bioconductor packages were used to analyze microarray gene expression data. Raw data was fed into the R package *limma* (14) for the standardization and noise reduction of the probe data, and the raw signal levels for each probe set were standardized. The Quantile method was used to normalize the microarray datasets. The t-statistic was used to detect statistically significant differentially expressed genes (DEGs) between the PTB and control samples. To eliminate false positives, the Benjamini and Hochberg false discovery rate (FDR) with $p < 0.05$ was used as a cut-off point for gene data. Thereafter, probes were matched to Entrez Gene IDs, and duplicates (with the highest fold change difference) and unmatched transcripts were filtered out. In the final stage, all the DEGs were classified as up- and downregulated genes based on the fold change threshold ($FC \geq |1.5|$).

Identification of Immune Cell Composition From Gene Expression Profiles

The fractions of 22 immune cell types in the PTB transcriptome profile were estimated using the CIBERSORT algorithm (15). This program employs linear support vector (SVR) regression to perform feature selection and to deconvolve the cell mixture from the gene expression profile. In this study, gene expression profiles of PTB and control samples were fed into the CIBERSORT algorithm where the algorithm converts the gene expression matrix into the immune cell-matrix and applies the filtering criteria of 1,000 permutations and significant p value set at ≤ 0.05 .



MTB Meta-Profiling Genetic Signatures

We obtained 380 genetic signatures identified from the modular analysis and meta-profiling of 16 publicly available gene expression datasets (16) (**Supplementary Material Table 2**). We compared the overlapping genes between these 380 TB genetic signatures, DEGs, and genes associated with immune cell populations for downstream analysis.

Identification of Immune Pathways From Gene Expression Profiles

The functional enrichment analysis of the DEGs was performed using g:Profiler (17), a webserver to interpret the function of gene lists (<https://biit.cs.ut.ee/gprofiler/gost>). This server matches a queried gene list to established functional data sources and uncovers gene ontologies as well as pathway terms that are significantly enriched at $q \leq 0.05$. Immune-related pathways were screened from the functional enrichment list. The DEGs which are contributing to immune-associated pathways were mapped to the known signature of TB and immune signature genes in the CIBERSORT to identify unreported genes in TB.

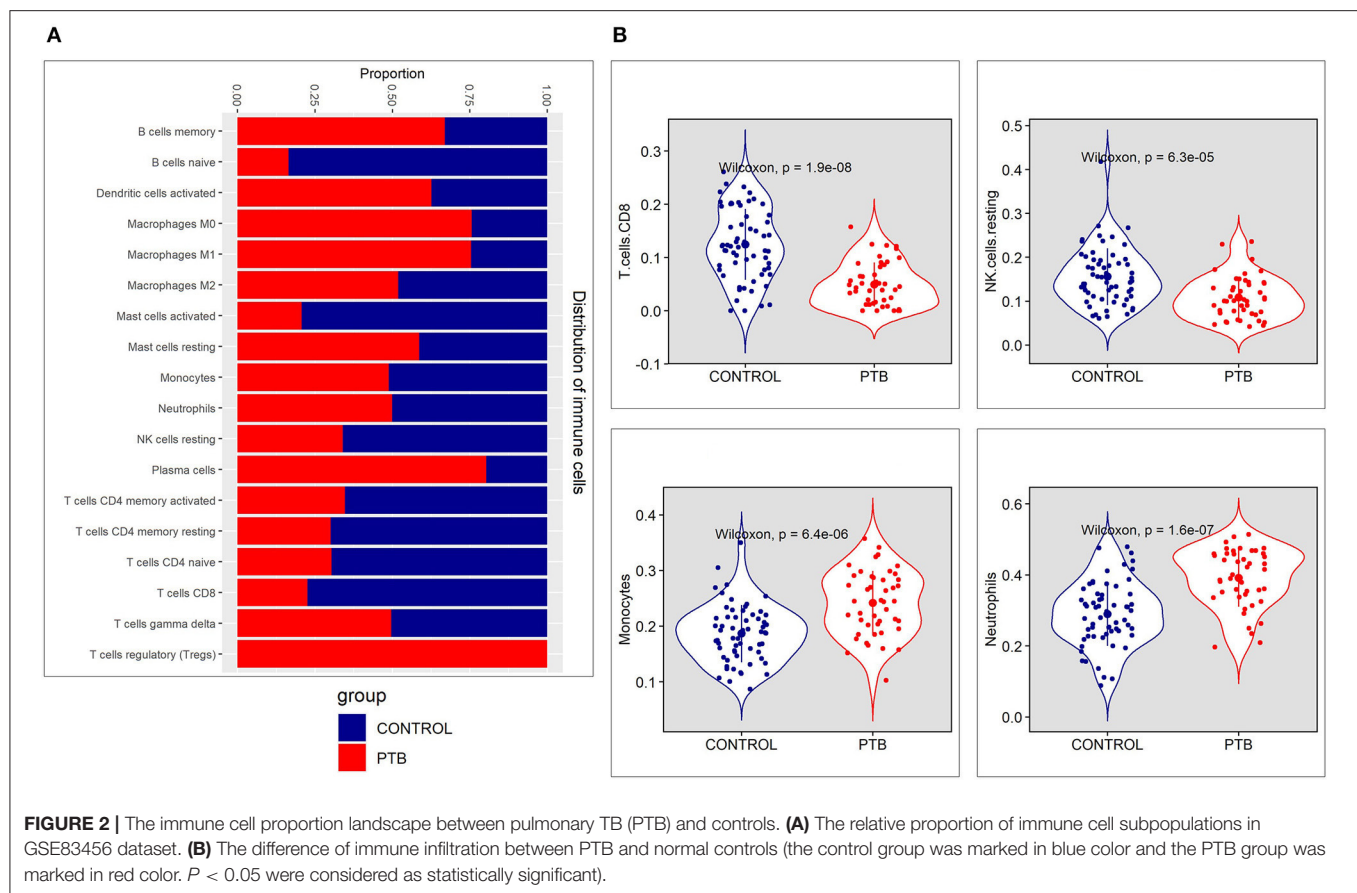
Identification of Semantic Similarity

Using encoded evidence in the Gene Ontology (GO) hierarchy, the functional similarity between unreported genes and known TB signatures is assessed. In this study, we used Wang's similarity metric to compare the biological process (BP) hierarchy. To quantify the semantic similarity between gene pairs, we used the R tool GoSemSim (18).

We employed Resnik's measure of Best-Match Average (BMA) method, which combines the semantic relationship scores of numerous GO terms and produces the average of all maximal similarities in each row and column because a gene can be annotated by many GO terms (19). Following that, gene pairs were selected based on a semantic score of ≤ 0.5 , with a larger score indicating a stronger relationship. The following formula was used to calculate the semantic similarity among gene pairs:

$$S_{GO}(A, B) = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{\sum_{t \in T_A} S_A(t) + \sum_{t \in T_B} S_B(t)} \quad (1)$$

where T_A designates the contribution of $t \in T_A$ term to the semantics of A based on the relative positions of t and A in the graph, and $S_A(t)$ implies the role of $t \in T_A$ term to the semantics of A.



Druggability Analysis

The Drug–Gene Interaction Database (DGIdb) (20) was used to assess the druggability of the genes. DGIdb is a central resource for drug–gene interaction data and the potential druggability of each query gene based on different databases. We included approved drugs, antineoplastic drugs, and Immunotherapeutic drug interactions filters and in advance filters, we selected 9 Disease-Agnostic sources databases, 43 gene categories, and 31 interaction types. We used drug target interaction with interactions score ≥ 0.03 to search the DEGs, which could act as potential drug target genes for MTB.

Correlation Among the Druggable Genes

The correlation between the druggable genes in PTB was investigated using Pearson's correlation method. The correlation (r) between each pair of gene matrices was ranked using Pearson's correlation coefficient (PCC). The formula used for computing the PCC existing between two genes is given below.

$$PCC(r) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2)$$

where \bar{x} and \bar{y} are the average of sample's gene expression signal in PTB of the two genes, respectively. The gene co-expression was confirmed using the Search Tool for the Retrieval of Interacting

Genes (STRING) (21), an online protein interaction database, with high confidence interaction score of ≥ 0.7 .

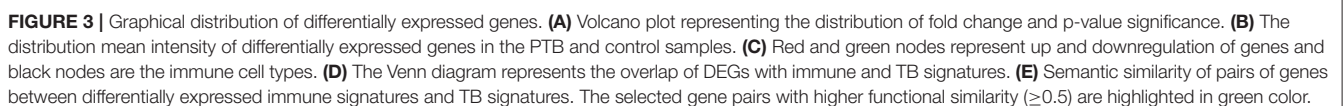
Real Time-PCR (RT-PCR) Validation of Druggable Genes

In order to verify our bioinformatics findings, we validated the expression of 9 druggable genes by the RT-PCR method. In brief, the RNA collected from THP-1 cell lines infected with the MTB strain (H37Rv) was collected after the post-incubation period as previously described (22). In brief, total RNA was reverse transcribed as complementary DNA (cDNA) and then amplified by the RT-PCR method using gene-specific oligonucleotide primers. The relative expression level of potentially druggable genes between the control and test cell lines was estimated by the $2^{-\Delta\Delta CT}$ formula after normalizing their expression levels with the GAPDH internal reference gene. A $p < 0.05$ under the standard two-tailed t -test was considered a significant value.

RESULTS

Immune Cell Proportion Analysis of PTB Gene Expression Profile

The immune cell proportion landscape of PTB is not yet fully revealed, particularly in low abundant cell subpopulations. In this study, the CIBERSORT algorithm has identified the



enrichment of genes associated with 10 types of adaptive immune cells like B cells naive, plasma cells, T follicular helper cells, CD8⁺ T cells, resting memory CD4⁺ T cells, T cells, CD4⁺ memory T cells activated, memory B cells, naive CD4⁺ T cells, regulatory T cells (Tregs), and Gamma-delta ($\gamma\delta$) T cells. On the other hand, DEGs associated with 12 innate immune cell type categories were NK cells resting, macrophages M2, monocytes, macrophages M1, macrophages M0, resting dendritic cells, eosinophils, dendritic cells activated, mast cells resting, NK cells activated, mast cells activated, and neutrophils were also found be enriched. The immune cell proportions of adaptive immune cells and innate immune cells are represented in **Supplementary Material Figure 1**.

The genes associated with four immune cells; NK cells activated, T follicular helper cells, dendritic cells resting, and eosinophils, were not significantly enriched in both groups. The proportion plot of the enriched immune cell types is represented in **Figure 2A**. We observed higher relative proportion of genes enriched for cell types like CD8⁺ T cells ($P = 1.9 \times 10^{-8}$), NK cells resting ($P = 6.3 \times 10^{-5}$), monocytes ($P = 6.4 \times 10^{-6}$), and neutrophils ($P = 1.6 \times 10^{-7}$) in the PTB samples compared to the control samples (**Figure 2B**, **Supplementary Material Figure 1**). Among the 4 cell types with a higher relative proportion of enriched genes from DEGs, the genes of CD8⁺ T cells and NK resting cells were found to be downregulated in PTB when compared to the control samples. On other hand, monocytes and neutrophil-associated genes were highly active in PTB when compared to control samples.

Identification of DEGs From Gene Expression Profile

The standardized gene expression data of “PTB vs. controls” was used to identify the differentially expressed genes. The volcano plot representing the distribution of fold change and the significant p -value is given in **Figure 3A**. The PTB vs. control group analysis revealed 624 DEGs (FC |1.5|, adj p -value of 0.05), with 393 upregulated and 231 downregulated genes. The top 10 DEGs obtained from PTB vs controls are given in **Table 1**. The mean distribution of intensity of differentially expressed genes in PTB and control samples is represented in **Figure 3B**.

Functional Enrichment Analysis of DEGs

The differentially expressed genes enriched using g:Profiler with the statistical significance of q value ≤ 0.05 , generated 309 ontologies of Biological Process (BP), 17 ontologies of Molecular Function (MF), 42 ontologies of Cellular Component (CC), and 85 terms in pathways (**Supplementary Material Table 3**). Overall, the enrichment analysis has shown the overlap with immune-related ontologies and pathways. We pooled immune-related pathways from enrichment terms to check how DEGs affect the immune system pathways. We observed the upregulation of pathways such as interferon signaling ($q = 9.16 \times 10^{-27}$), cytokine signaling in the immune system ($q = 2.34 \times 10^{-21}$), neutrophil degranulation ($q = 3.13 \times 10^{-11}$), viral genome replication ($q = 2.47 \times 10^{-7}$), and response to biotic stimulus, etc. (**Supplementary Material Figure S1**). On the other hand, pathways like T-cell antigen receptor

TABLE 1 | The top 10 differentially expressed gene list in pulmonary tuberculosis (PTB).

Symbol	FC	Gene name	Adj P value
<i>SERPING1</i>	7.75	Serpin family G member 1	2.08E–36
<i>ANKRD22</i>	7.58	Ankyrin repeat domain 22	9.83E–35
<i>FCGR1A</i>	7.51	Fc fragment of IgG receptor Ia	1.00E–28
<i>FCGR1C</i>	7.04	Fc fragment of IgG receptor Ic, pseudogene	3.79E–30
<i>FCGR1B</i>	6.02	Fc fragment of IgG receptor Ib	2.75E–28
<i>LRRN3</i>	–2.91	Leucine rich repeat neuronal 3	4.35E–13
<i>FCGBP</i>	–2.60	Fc fragment of IgG binding protein	1.03E–12
<i>NELL2</i>	–2.27	Neural EGFL like 2	6.95E–17
<i>GZMK</i>	–2.20	Granzyme K	2.19E–10
<i>CCR7</i>	–2.19	C-C motif chemokine receptor 7	1.15E–16

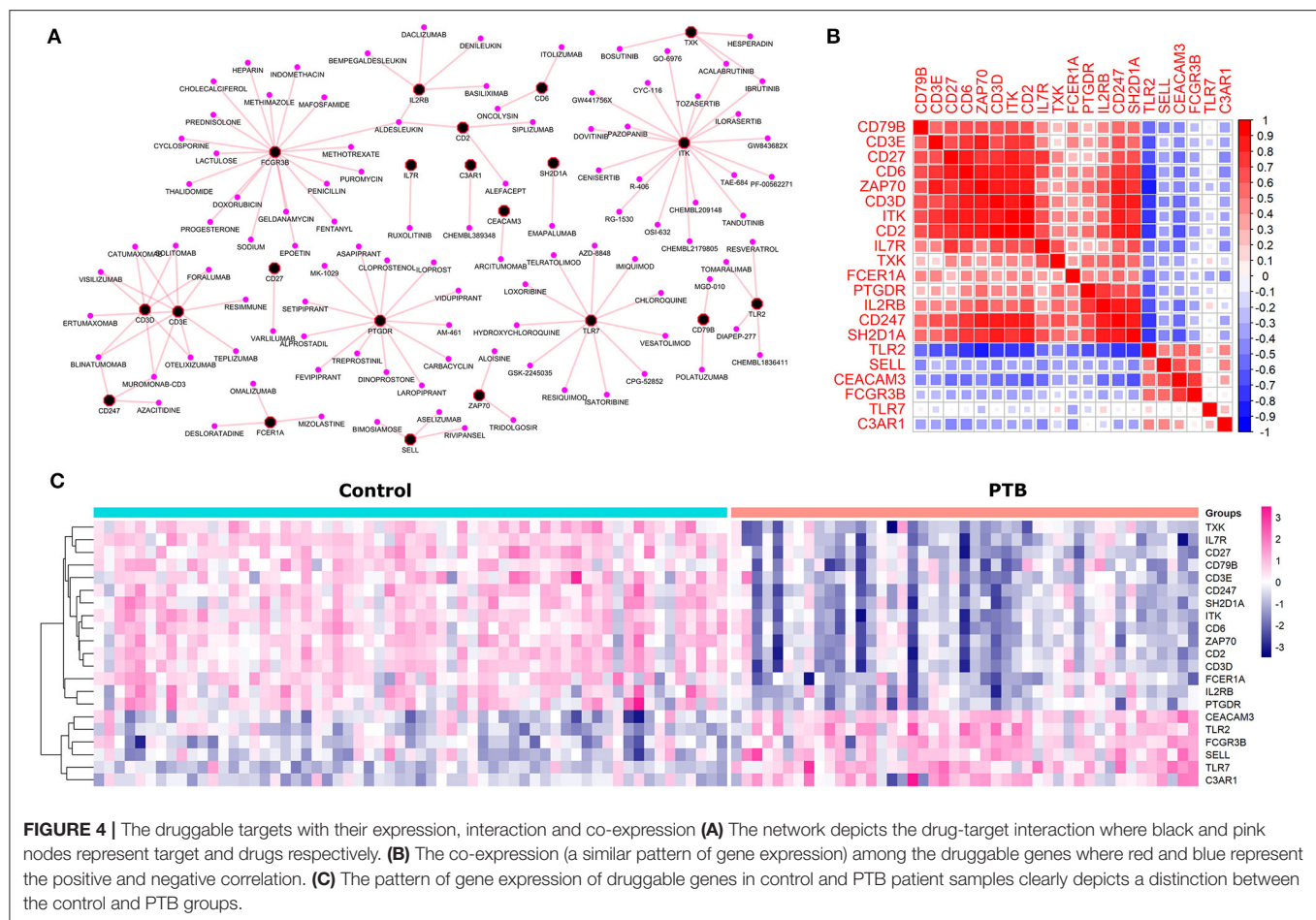
signaling, antigen receptor-mediated signaling, NF-kappa B signaling, T cell activation, T cell receptor signaling, leukocyte differentiation, leukocyte activation, alpha-beta T cell activation, and T cell differentiation, were downregulated (**Supplementary Material Figure S1**). Overall, our functional enrichment analysis points to a major downregulation of T cell-related ontologies and pathways.

Mapping DEGs to Immune Cell Proportions in PTB

Here we investigated the genes overlapping between the CIBERSORT signature and DEGs. There are about 96 DEGs (**Figure 3C**) contributing to different immune cell types (**Supplementary Material Table 4**). Interestingly, we found that 31.25% of DEGs were contributing to the immune cell type “CD8⁺ T cells.” We also observed that all those genes contributing to the “CD8⁺ T cells” were downregulated in PTB samples, as shown in **Figure 3C**. The findings from the mapping of DEGs to immune cell proportions are consistent with functional enrichment analysis, where T cell-related pathways have shown major dysregulation.

Comparison of DEGs, Immune Cell Signatures With TB Meta-Analysis Signatures

The differentially expressed immune signatures in the sample of patients with TB were compared with the known signatures of TB (**Supplementary Material Table 4**). Here, we observed 39 (40.6%) differentially expressed immune signatures overlapping with TB meta-analysis gene signatures and 57 novel genes (59.3%) contributing to the immune cell proportion (**Figure 3D**). Further, semantic similarity (functional association) of these 57 novel genes with 39 overlapping with TB signatures was performed to identify the most predominant genes. The semantic similarity score of ≥ 0.5 among the gene pairs was considered as a highly significant score implying a stronger association. The semantic similarity of 45/57 genes has shown a stronger functional association with overlapping with TB signatures (**Figure 3E**). Again, it is important to pinpoint that 20 out of 45



genes (44%) were contributing to the immune cell type “CD8+ T cells.”

Druggability and Co-expression Analysis

Druggability analysis was performed on the 45 genes that had shown higher functional similarity to the known TB signature. We found that 21 druggable genes (**Supplementary Material Figure S2**) (46%) with an interaction score ≥ 0.03 , were enriched against terms like an antibody, binder, inhibitor, antagonist, agonist, modulator, and activator. Of all the druggable genes, *ITK* and *FCGR3B* genes were observed to have the highest number of drug interactions (19 drugs) followed by *PTGDR* (13 drugs), *TLR7* (12 drugs), and *CD3E* (10 drugs). The drug-target interaction network is represented in **Figure 4A**. Next, we checked the association of druggable targets to the immune cell types. Among the 21 druggable targets, 48% were contributing to the immune cell type “CD8+ T cells” followed by monocytes (9%), naïve B cells (9%), and neutrophils (9%) (**Figure 4B**). Interestingly, the expression patterns of these 21 druggable genes have shown a clear distinction in PTB when compared to control samples (**Figure 4C**).

To check the correlation of these 21 druggable targets in patients with PTB, we performed Pearson’s correlation

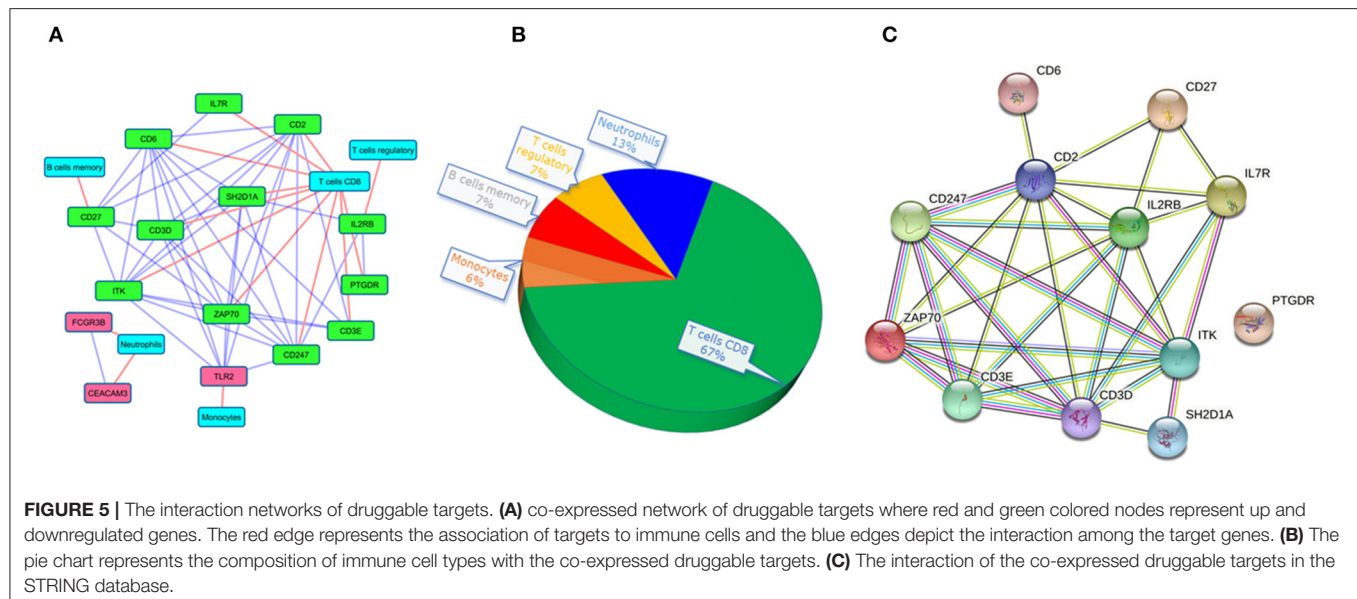
analysis (**Figure 4B**). The correlation analysis performed between druggable targets in PTB samples resulted in 15 genes (*ITK*, *CD2*, *CD6*, *CD247*, *CD27*, *CD3D*, *ZAP70*, *SH2D1A*, *IL2RB*, *CEACAM3*, *IL7R*, *TLR2*, *CD3E*, *PTGDR*, *FCGR3B*) with higher Pearson correlation coefficient ($r^2 \geq |0.7|$). The co-expressed genes are shown in **Table 2**. Among the 15 co-expressed genes, 12 and 3 were down and upregulated respectively (**Figure 5A**). Upregulated genes were contributing to the immune cell types “Monocytes” (*TLR2*) and Neutrophils (*FCGR3B* and *CEACAM3*). Again, 10 downregulated genes were contributing to “T cells CD8” and 1 each for “T cells regulatory” and “B cells memory” (**Figure 5B**).

Interestingly, we noticed a cluster of 12 co-expressed genes *ITK*, *CD2*, *CD6*, *ZAP70*, *CD247*, *CD3D*, *SH2D1A*, *CD27*, *CD3E*, *IL2RB*, *IL7R*, and *PTGDR*. To validate the co-expression among the 12 downregulated genes we queried them in the STRING database with a high confidence score of ≥ 0.7 . The STRING database identified strong interaction among the 11 genes except for *PTGDR* (**Figure 5C**). Co-expression and protein interaction network from the STRING database has shown the mutual influence of the 11 genes in the expression and functional activities. Nine genes (except *CD27*, *PTGDR*, and *IL2RB*) were contributing to “CD8+ T cells.” All the genes were predicted to be targeted by different drug molecules, most of which are

TABLE 2 | List of co-expressed genes and the druggable targets.

Gene	Immune cells	Regulation	Co-expressed genes	PCC range
<i>ITK</i>	CD8 ⁺ T cells	DOWN	<i>CD2, CD3D, ZAP70, CD247, SH2D1A, CD3E, TLR2, CD6, CD27</i>	0.72–0.91
<i>CD2</i>	CD8 ⁺ T cells	DOWN	<i>CD3D, SH2D1A, ZAP70, CD247, TLR2, CD3E, ITK, CD6, CD27</i>	0.74–0.91
<i>CD6</i>	CD8 ⁺ T cells	DOWN	<i>ITK, ZAP70, CD2, CD3D, CD247, SH2D1A, TLR2, CD3E, CD27</i>	0.74–0.89
<i>ZAP70</i>	CD8 ⁺ T cells	DOWN	<i>TLR2, CD247, CD3E, SH2D1A, CD6, CD2, ITK, CD3D, CD27</i>	0.76–0.88
<i>CD247</i>	CD8 ⁺ T cells	DOWN	<i>SH2D1A, IL2RB, TLR2, CD2, CD3D, ITK, ZAP70, CD6</i>	0.79–0.87
<i>CD3D</i>	CD8 ⁺ T cells	DOWN	<i>CD247, SH2D1A, ZAP70, TLR2, CD2, ITK, CD6, CD27</i>	0.72–0.90
<i>SH2D1A</i>	CD8 ⁺ T cells	DOWN	<i>IL2RB, TLR2, CD2, CD247, CD3D, ITK, CD6, ZAP70</i>	0.73–0.88
<i>TLR2</i>	Monocytes	UP	<i>ZAP70, CD247, CD6, CD2, SH2D1A, ITK, CD3D</i>	0.72–0.82
<i>CD27</i>	B cells memory	DOWN	<i>CD6, ITK, CD3D, CD2, ZAP70, IL7R</i>	0.71–0.85
<i>CD3E</i>	CD8 ⁺ T cells	DOWN	<i>ZAP70, ITK, CD6, CD2</i>	0.74–0.80
<i>IL2RB</i>	T cells regulatory	DOWN	<i>SH2D1A, CD247, PTGDR</i>	0.79–0.82
<i>CEACAM3</i>	Neutrophils	UP	<i>FCGR3B</i>	0.71
<i>IL7R</i>	CD8 ⁺ T cells	DOWN	<i>CD27</i>	0.71
<i>PTGDR</i>	CD8 ⁺ T cells	DOWN	<i>IL2RB</i>	0.79
<i>FCGR3B</i>	Neutrophils	UP	<i>CEACAM3</i>	0.71

#PCC, Pearson correlation coefficient.



monoclonal antibodies (Table 3). Hence, the integrated analysis has depicted predominant deregulation of “CD8⁺ T cells” as the key genetic signatures for active PTB.

RT-PCR Validation of Druggable Genes

The real-time PCR gene expression results showed that the relative expression levels of 9 potential druggable genes (*ITK*, *CD2*, *CD6*, *CD247*, *ZAP70*, *CD3D*, *SH2D1A*, *CD3E*, and *IL7R*) were consistent with the findings of microarray hybridization. All the genes were differentially expressed between treated and untreated cell lines ($p \leq 0.01$). These results confirm the dysregulated “CD8⁺ T cell signaling” plays important role in establishing TB infection (Supplementary Material Figure S3).

DISCUSSION

Host genetic factors are known to play an important role in regulating the initial TB infection and determining the disease progression in the lungs (37). Genome-wide association studies have underlined the relevance of numerous polymorphisms in immune response-related genes in contributing to susceptibility or resistance to TB (38). However, polymorphism studies were unable to provide full insight into the complex molecular crosstalk between thousands of host genes involved in innate and adaptive immune responses. In this context, high throughput transcriptome approaches have shown great promise in dissecting the host-pathogen interactions thereby helping to develop a novel vaccine and therapeutic targets for several

TABLE 3 | The list of drugs shows direct interaction with 9 genes associated with CD+T cell functioning.

Gene	Drug	Interaction type and directionality	Sources	PMIDs	Query Score	Interaction Score*
ITK	CHEMBL2179805	–	DTC	(23)	8.77	6.71
CD2	ALEFACEPT	Inhibitor (inhibitory)	TdgClinicalTrial	(24–26)	21.92	106.32
			ChemblInteractions TEND	(27, 28)		
			GuideToPharmacology	(29)		
	SIPLIZUMAB	Inhibitor (inhibitory)	TdgClinicalTrial	–	13.15	63.79
			ChemblInteractions TTD			
CD6	ITOLIZUMAB	Antibody (inhibitory)	GuideToPharmacology TTD	–	8.77	63.79
	ONCOLYSIN CD6	–	ChemblInteractions	–	4.38	31.9
CD247	MUROMONAB-CD3	–	TdgClinicalTrial	(30, 31)	9.86	15.95
ZAP70	TRIDOLGOSIR	–	DTC	(32)	2.92	21.26
	ALOISINE	Inhibitor (inhibitory)	GuideToPharmacology	–	1.46	10.63
CD3D	MUROMONAB-CD3	Inhibitor (inhibitory)	TdgClinicalTrial	(30, 31)	13.15	9.11
			ChemblInteractions			
	BLINATUMOMAB	Activator (activating)	TdgClinicalTrial	–	5.85	6.08
			ChemblInteractions			
SH2D1A	EMAPALUMAB	–	PharmGKB FDA	–	1.1	15.95
CD3E	MUROMONAB-CD3	Binder, inhibitor (inhibitory), antibody (inhibitory)	TdgClinicalTrial	(33, 34)	26.31	12.76
			ChemblInteractions TEND			
GuideToPharmacology TTD	OTELIXIZUMAB	Antibody (inhibitory)	TdgClinicalTrial	–	8.77	6.38
			ChemblInteractions			
			GuideToPharmacology			
IL7R	RUXOLITINIB	–	CGI	(35, 36)	1.1	15.95

*Drug molecules showing >5 interaction score is shown here.

infectious diseases (39–41). Therefore, we explored host immune system response through integrated systems biology approach based on immune cell subtyping and differential gene expression profiles of patients with PTB to normal controls.

The cellular and molecular background of TB-induced systemic immunological dysregulation is poorly understood. Therefore, we screened the DEGs in PTB and deciphered their contribution to immune cell proportion alterations. Traditionally, host transcriptomics studies have relied on whole blood to characterize TB gene signatures by aggregating transcriptomic signals from many different cell types but were unable to identify specific immune cell type signatures (42). To overcome these constraints, we used a powerful computational technique called CIBERSORT to define the range of immune cell states in the blood of patients with TB. This method relies on linear support vector regression (SVR), a machine learning approach to deconvolute the gene expression signatures, known as “signature matrix” for determining the relative fraction of immune cell proportions in blood or tissues (15). The CIBERSORT method has been widely used to infer immune cell types from transcriptomics data to predict outcomes of different cancers (9, 43, 44) and infectious diseases (45–47). In this study, the CIBERSORT output identified the downregulation of “CD8⁺ T cells” in patients with PTB.

The GO functional enrichment analysis of gene expression profile revealed the upregulation of interferon signaling, cytokine signaling in the immune system, neutrophil degranulation, and response to biotic stimulus pathways. The MTB infection of primary human macrophages is shown to induce type I IFN signaling and limit the expression of IL-1 β , which imparts immunity against the infection (48). Additionally, the downregulation of major pathways associated with T cells function like T-cell antigen receptor signaling, leukocyte differentiation, leukocyte activation, T cell activation, T cell differentiation, T cell receptor signaling, alpha-beta T cell activation, NF-kappa B signaling, and antigen receptor-mediated signaling pathway were noted in PTB samples.

Majority of the DEGs contributing to the immune cell type “CD8⁺ T cells” were clearly downregulated in the PTB samples indicating their potential roles in defense against TB. These cells are also known as killer or cytotoxic T lymphocytes, as they potentially destroy the infected cells by recruiting cytokines and other immune cells to the site of infection. The low abundance of blood CD8⁺ T lymphocytes may impair the effective immunity against pathogens, as they lack a sufficient cytotoxic T cells to recognize the MHC class I-restricted epitopes of MTB antigens, in the site of infection (49). A recent RNA transcriptome study used the positron emission tomography (PET) data collected

from recovered patients with TB, at 4th and 24th weeks has also reported that genes associated with the overexpression of B cells and down expression of T cells and platelets confirms our findings (50). Thus, the downregulation that contribute to immune cell type is concordant with the pathway enrichment analysis findings of lower expression of T cell-related ontologies and pathways.

The druggability potential of any protein is attributed to its binding specificity with small compounds following Lipinski's rule-of-five for drug likeliness (51). Numerous bioinformatic and empirical methods which consume less time and provide faster prescreening of druggability of candidate proteins than conventional methods have been developed (52, 53). There are a variety of computational methods available, which can predict druggability and protein-binding sites by using energy dynamics to geometrical topological estimations, and from flexible to rigid proteins (54, 55).

By applying druggability and co-expression features we identified 9 CD8⁺ T cells associated genes (*ITK*, *CD2*, *CD6*, *CD247*, *ZAP70*, *CD3D*, *SH2D1A*, *CD3E*, and *IL7R*) as potential therapeutic targets of PTB. However, it is pivotal to carefully prioritize the drug molecules based on their mode of action, whether activator or inhibitor based on the gene expression status. For example, over-expressed genes can be targeted by inhibitory molecules, and downregulated genes can be targeted by activator molecules (56). From the above 9 genes, the therapeutic potential of *ITK* and *IL7R* has been characterized by experimental methods. *ITK* is a tyrosine kinase expressed on T-cells, which regulates its T-cell development and function. Human lungs with *ITK* deficiency impair early protection against MTB *in vivo* (57). Improving *ITK* signaling pathways could become an alternative approach for combating MTB infection. One study reveals the role of *IL7R* on T-cell immunity in human TB (58). The authors reported that patients with TB had lower *IL7R* concentrations and lower *IL7R* expression in T cells than healthy controls, indicating that patients with TB have impaired T-cell sensitivity. In addition, due to post-transcriptional processes, patients with TB had reduced amounts of *IL7R* in T cells. *In vitro* experiments revealed that MTB-specific T lymphocytes from patients with TB have reduced *IL-7*-induced *STAT5* phosphorylation and *IL-7*-promoted cytokine production (59). The role of the remaining 7 genes (*CD2*, *CD6*, *CD247*, *ZAP70*, *CD3D*, *SH2D1A*, and *CD3E*) in T-cell signaling and modulation of host immune responses in mycobacterium infections is also supported (60–64).

Our results highlight the dysregulation of CD8⁺ T cells and the associated genes in PTB patients. These findings are exciting not just from the fact that CD8⁺ T cell-associated genes have the potential to act as potential therapeutic targets but prove that their role is not less important than CD4⁺ T cells in controlling MTB infection. We acknowledge that our strategy has some technical constraints. CIBERSORT was a convenient computational tool for determining infiltrating immune cell fractions, but it was still less precise than immunohistochemistry

or flow cytometry, which could lead to inaccuracies in immune cell fractions. However, to overcome this limitation to some extent, we have linked gene expression profiles of immune signatures followed by functional enrichment, semantic similarity, druggability, and co-expression among the identified key signatures.

CONCLUSION

In this study, by coupling computational deconvolution algorithms and high throughput blood transcriptomics data, we identified the difference in T-cell-related immune cell populations among patients with PTB. The functional enrichment of 624 DEGs (393 over-expressed and 231 under-expressed) identified in the blood transcriptome of PTB patients revealed the major dysregulation of T cell-related ontologies and pathways. By linking DEGs against immune cell populations and TB gene signature, this study identified 9 CD8⁺ T cells associated genes (*ITK*, *CD2*, *CD6*, *CD247*, *ZAP70*, *CD3D*, *SH2D1A*, *CD3E*, and *IL7R*) as potential therapeutic targets of PTB. The expression levels of these 9 genes in MTB infection in cell lines were assessed by RT-PCR-based expression assay, confirming the experimental validation. However, further *in vitro* and *in vivo* studies are needed to establish the role of these genes in PTB infection, progression, and treatment.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: <http://www.ncbi.nlm.nih.gov/geo>; GSE83456.

AUTHOR CONTRIBUTIONS

FA, BB, VV, and NS: conceptualization. FA, HA, SP, and BB: data curation. FA and BB: formal analysis. FA: funding acquisition and project administration. FA, NZ, SP, and BB: methodology. BB and SP: software and visualization. FA, NS, and BB: supervision. BB: validation. FA, NZ, HA, SP, PS, VV, RE, NS, and BB: writing original draft and editing. All authors contributed to the article and approved the submitted version.

FUNDING

The authors extend their appreciation to the Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number IFPRC-063-247-2020 and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.812857/full#supplementary-material>

REFERENCES

1. Makam P, Matsa R. “Big Three” infectious diseases: tuberculosis, malaria and HIV/AIDS. *Curr Top Med Chem.* (2021) 21:2779–99. doi: 10.2174/1568026621666210916170417
2. Chakaya J, Khan M, Ntoumi F, Akillu E, Fatima R, Mwaba P, et al. Global Tuberculosis Report 2020—reflections on the Global TB burden, treatment and prevention efforts. *Int J Infect Dis.* (2021). doi: 10.1016/j.ijid.2021.02.107
3. Maertzdorf J, Tonnies M, Lozza L, Schommer-Leitner S, Mollenkopf H, Bauer TT, et al. Mycobacterium tuberculosis Invasion of the Human Lung: First Contact. *Front Immunol.* (2018) 9:1346. doi: 10.3389/fimmu.2018.01346
4. Reece ST, Kaufmann SH. Floating between the poles of pathology and protection: can we pin down the granuloma in tuberculosis? *Curr Opin Microbiol.* (2012) 15:63–70. doi: 10.1016/j.mib.2011.10.006
5. Morris SK, Giroux RJP, Consunji-Araneta R, Stewart K, Baikie M, Kakkar F, et al. Epidemiology, clinical features and outcomes of incident tuberculosis in children in Canada in 2013–2016: results of a national surveillance study. *Arch Dis Child.* (2021). doi: 10.1136/archdischild-2021-322092
6. Von Both U, Kaforou M, Levin M, Newton SM. Understanding immune protection against tuberculosis using RNA expression profiling. *Vaccine.* (2015) 33:5289–93. doi: 10.1016/j.vaccine.2015.05.025
7. Banaganapalli B, Al-Rayes N, Awan ZA, Alsulaimany FA, Alamri AS, Elango R, et al. Multilevel systems biology analysis of lung transcriptomics data identifies key miRNAs and potential miRNA target genes for SARS-CoV-2 infection. *Comput Biol Med.* (2021) 135:104570. doi: 10.1016/j.combiomed.2021.104570
8. Banaganapalli B, Mansour H, Mohammed A, Alharthi AM, Aljuaid NM, Nasser KK, et al. Exploring celiac disease candidate pathways by global gene expression profiling and gene network cluster analysis. *Sci Rep.* (2020) 10:16290. doi: 10.1038/s41598-020-73288-6
9. Craven KE, Gokmen-Polar Y, Badve SS. CIBERSORT analysis of TCGA and METABRIC identifies subgroups with better outcomes in triple negative breast cancer. *Sci Rep.* (2021) 11:4691. doi: 10.1038/s41598-021-83913-7
10. Xie Z, Bailey A, Kuleshov MV, Clarke DJB, Evangelista JE, Jenkins SL, et al. Gene set knowledge discovery with enrichr. *Curr Protocol.* (2021) 1:e90. doi: 10.1002/cpz1.90
11. Kawada JI, Takeuchi S, Imai H, Okumura T, Horiba K, Suzuki T, et al. Immune cell infiltration landscapes in pediatric acute myocarditis analyzed by CIBERSORT. *J Cardiol.* (2021) 77:174–8. doi: 10.1016/j.jjcc.2020.08.004
12. Feng Z, Qu J, Liu X, Liang J, Li Y, Jiang J, et al. Integrated bioinformatics analysis of differentially expressed genes and immune cell infiltration characteristics in Esophageal Squamous cell carcinoma. *Sci Rep.* (2021) 11:16696. doi: 10.1038/s41598-021-96274-y
13. Blankley S, Graham CM, Turner J, Berry MP, Bloom CI, Xu Z, et al. The transcriptional signature of active tuberculosis reflects symptom status in extra-pulmonary and pulmonary tuberculosis. *PLoS ONE.* (2016) 11:e0162220. doi: 10.1371/journal.pone.0162220
14. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* (2015) 43:e47. doi: 10.1093/nar/gkv007
15. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods.* (2015) 12:453–7. doi: 10.1038/nmeth.3337
16. Blankley S, Graham CM, Levin J, Turner J, Berry MP, Bloom CI, et al. A 380-gene meta-signature of active tuberculosis compared with healthy controls. *Eur Respir J.* (2016) 47:1873–6. doi: 10.1183/13993003.02121-2015
17. Raudvere U, Kolberg L, Kuzmin I, Arak T, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* (2019) 47:W191–8. doi: 10.1093/nar/gkz369
18. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics.* (2010) 26:976–8. doi: 10.1093/bioinformatics/btq064
19. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics.* (2008) 9 Suppl 5:S4. doi: 10.1186/1471-2105-9-S5-S4
20. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. (2018). DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res* 46:D1068–73. doi: 10.1093/nar/gkx1143
21. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* (2021) 49:D605–12. doi: 10.1093/nar/gkab835
22. Zhang W, Niu C, Fu RY, Peng ZY. Mycobacterium tuberculosis H37Rv infection regulates alternative splicing in Macrophages. *Bioengineered.* (2018) 9:203–8. doi: 10.1080/21655979.2017.1387692
23. Zapf CW, Gerstenberger BS, Xing L, Limburg DC, Anderson DR, Caspers N, et al. Covalent inhibitors of interleukin-2 inducible T cell kinase (itk) with nanomolar potency in a whole-blood assay. *J Med Chem.* (2012) 55:10047–63. doi: 10.1021/jm301190s
24. Ortonne JP. Clinical response to alefacept: results of a phase 3 study of intramuscular administration of alefacept in patients with chronic plaque psoriasis. *J Eur Acad Dermatol Venereol.* (2003) 17 Suppl 2:12–6. doi: 10.1046/j.1468-3083.17.s2.3.x
25. Goedkoop AY, De Rie MA, Picavet DI, Kraan MC, Dinant HJ, Van Kuijk AW, et al. Alefacept therapy reduces the effector T-cell population in lesional psoriatic epidermis. *Arch Dermatol Res.* (2004) 295:465–73. doi: 10.1007/s00403-004-0450-y
26. Chamian F, Lin SL, Lee E, Kikuchi T, Gilleaudeau P, Sullivan-Whalen M, et al. Alefacept (anti-CD2) causes a selective reduction in circulating effector memory T cells (Tem) and relative preservation of central memory T cells (Tcm) in psoriasis. *J Transl Med.* (2007) 5:27. doi: 10.1186/1479-5876-5-27
27. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res.* (2002) 30:412–5. doi: 10.1093/nar/30.1.412
28. Da Silva AJ, Brickelmaier M, Majeau GR, Li Z, Su L, Hsu YM, et al. Alefacept, an immunomodulatory recombinant LFA-3/IgG1 fusion protein, induces CD16 signaling and CD2/CD16-dependent apoptosis of CD2(+) cells. *J Immunol.* (2002) 168:4462–71. doi: 10.4049/jimmunol.168.9.4462
29. Larsen R, Ryder LP, Svejgaard A, Gniadecki R. Changes in circulating lymphocyte subpopulations following administration of the leucocyte function-associated antigen-3 (LFA-3)/IgG1 fusion protein alefacept. *Clin Exp Immunol.* (2007) 149:23–30. doi: 10.1111/j.1365-2249.2007.03380.x
30. Todd PA, Brogden RN. Muromonab CD3. A review of its pharmacology and therapeutic potential. *Drugs.* (1989) 37:871–99. doi: 10.2165/00003495-198937060-00004
31. Wilde MI, Goa KL. Muromonab CD3: a reappraisal of its pharmacology and use as prophylaxis of solid organ transplant rejection. *Drugs.* (1996) 51:865–94. doi: 10.2165/00003495-199651050-00010
32. Chen JJ, Chen HL, Demetriou M. Lateral compartmentalization of T cell receptor versus CD45 by galectin-N-glycan binding and microfilaments coordinate basal and activation signaling. *J Biol Chem.* (2007) 282:35361–72. doi: 10.1074/jbc.M706923200
33. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov.* (2006) 5:821–34. doi: 10.1038/nrd2132
34. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? *Nat Rev Drug Discov.* (2006) 5:993–6. doi: 10.1038/nrd2199
35. Roberts KG, Morin RD, Zhang J, Hirst M, Zhao Y, Su X, et al. Genetic alterations activating kinase and cytokine receptor signaling in high-risk acute lymphoblastic leukemia. *Cancer Cell.* (2012) 22:153–66. doi: 10.1016/j.ccr.2012.06.005
36. Maude SL, Tasián SK, Vincent T, Hall JW, Sheen C, Roberts KG, et al. Targeting JAK1/2 and mTOR in murine xenograft models of Ph-like acute lymphoblastic leukemia. *Blood.* (2012) 120:3510–8. doi: 10.1182/blood-2012-03-415448
37. Chai Q, Lu Z, Liu CH. Host defense mechanisms against Mycobacterium tuberculosis. *Cell Mol Life Sci.* (2020) 77:1859–78. doi: 10.1007/s00018-019-03353-5
38. Van Tong H, Velavan TP, Thye T, Meyer CG. Human genetic factors in tuberculosis: an update. *Trop Med Int Health.* (2017) 22:1063–71. doi: 10.1111/tmi.12923
39. Domínguez Á, Muñoz E, López MC, Cordero M, Martínez JP, Viñas M. Transcriptomics as a tool to discover new antibacterial targets. *Biotechnol Lett.* (2017) 39:819–28. doi: 10.1007/s10529-017-2319-0
40. Guleria V, Jaiswal V. Comparative transcriptome analysis of different stages of Plasmodium falciparum to explore vaccine and drug candidates. *Genomics.* (2020) 112:796–804. doi: 10.1016/j.ygeno.2019.05.018

41. Le BL, Andreoletti G, Oskotsky T, Vallejo-Gracia A, Rosales R, Yu K, et al. Transcriptomics-based drug repositioning pipeline identifies therapeutic candidates for COVID-19. *Sci Rep.* (2021) 11:12310. doi: 10.1038/s41598-021-91625-1
42. Blankley S, Berry MP, Graham CM, Bloom CI, Lipman M, O'garra A. The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis. *Philos Trans R Soc Lond B Biol Sci.* (2014) 369:20130427. doi: 10.1098/rstb.2013.0427
43. Huang R, Mao M, Lu Y, Yu Q, Liao L. A novel immune-related genes prognosis biomarker for melanoma: associated with tumor microenvironment. *Aging.* (2020) 12:6966–80. doi: 10.18632/aging.103054
44. Wu Z, Chen H, Luo W, Zhang H, Li G, Zeng F, et al. The landscape of immune cells infiltrating in prostate cancer. *Front Oncol.* (2020) 10:517637. doi: 10.3389/fonc.2020.517637
45. Reyes M, Filbin MR, Bhattacharyya RP, Billman K, Eisenhaure T, Hung DT, et al. An immune-cell signature of bacterial sepsis. *Nat Med.* (2020) 26:333–40. doi: 10.1038/s41591-020-0752-4
46. Bossel Ben-Moshe N, Hen-Avivi S, Levitin N, Yehezkel D, Oosting M, Joosten L, et al. Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. *Nat Commun.* (2019) 10:3266. doi: 10.1038/s41467-019-11257-y
47. Wu YY, Wang SH, Wu CH, Yen LC, Lai HF, Ho CL, et al. In silico immune infiltration profiling combined with functional enrichment analysis reveals a potential role for naïve B cells as a trigger for severe immune responses in the lungs of COVID-19 patients. *PLoS ONE.* (2020) 15:e0242900. doi: 10.1371/journal.pone.0242900
48. Novikov A, Cardone M, Thompson R, Shenderov K, Kirschman KD, Mayer-Barber KD, et al. Mycobacterium tuberculosis triggers host type I IFN signaling to regulate IL-1 β production in human macrophages. *J Immunol.* (2011) 187:2540–7. doi: 10.4049/jimmunol.1100926
49. Cho S, Mehra V, Thoma-Urszyski S, Stenger S, Serbina N, Mazzaccaro RJ, et al. Antimicrobial activity of MHC class I-restricted CD8⁺ T cells in human tuberculosis. *Proc Natl Acad Sci USA.* (2000) 97:12210–5. doi: 10.1073/pnas.210391497
50. Oda T, Malherbe ST, Meier S, Maasdorp E, Kleynhans L, Du Plessis N, et al. The peripheral blood transcriptome is correlated with pet measures of lung inflammation during successful tuberculosis treatment. *Front Immunol.* (2020) 11:596173. doi: 10.3389/fimmu.2020.596173
51. Abi Hussein H, Geneix C, Petitjean M, Borrel A, Flatters D, Camproux AC. Global vision of druggability issues: applications and perspectives. *Drug Discov Today.* (2017) 22:404–15. doi: 10.1016/j.drudis.2016.11.021
52. Moumbock AFA, Li J, Mishra P, Gao M, Günther S. Current computational methods for predicting protein interactions of natural products. *Comput Struct Biotechnol J.* (2019) 17:1367–76. doi: 10.1016/j.csbj.2019.08.008
53. Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, et al. The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nat Protoc.* (2015) 10:733–55. doi: 10.1038/nprot.2015.043
54. Ozdemir ES, Halakou F, Nussinov R, Gursoy A, Keskin O. Methods for discovering and targeting druggable protein-protein interfaces and their application to repurposing. *Methods Mol Biol.* (2019) 1903:1–21. doi: 10.1007/978-1-4939-8955-3_1
55. Fauman EB, Rai BK, Huang ES. Structure-based druggability assessment—identifying suitable targets for small molecule therapeutics. *Curr Opin Chem Biol.* (2011) 15:463–8. doi: 10.1016/j.cbpa.2011.05.020
56. Gottesfeld JM, Neely L, Trauger JW, Baird EE, Dervan PB. Regulation of gene expression by small molecules. *Nature.* (1997) 387:202–5. doi: 10.1038/387202a0
57. Huang L, Ye K, McGee MC, Nidetz NF, Elmore JP, Limper CB, et al. Interleukin-2-inducible T-cell kinase deficiency impairs early pulmonary protection against mycobacterium tuberculosis infection. *Front Immunol.* (2019) 10:3103. doi: 10.3389/fimmu.2019.03103
58. Lundtoft C, Afum-Adjei Awuah A, Rimpler J, Harling K, Nausch N, Kohns M, et al. Aberrant plasma IL-7 and soluble IL-7 receptor levels indicate impaired T-cell response to IL-7 in human tuberculosis. *PLoS Pathog.* (2017) 13:e1006425. doi: 10.1371/journal.ppat.1006425
59. Rochman Y, Spolski R, Leonard WJ. New insights into the regulation of T cells by gamma(c) family cytokines. *Nat Rev Immunol.* (2009) 9:480–90. doi: 10.1038/nri2580
60. Mahon RN, Sande OJ, Rojas RE, Levine AD, Harding CV, Boom WH. Mycobacterium tuberculosis ManLAM inhibits T-cell-receptor signaling by interference with ZAP-70, Lck and LAT phosphorylation. *Cell Immunol.* (2012) 275:98–105. doi: 10.1016/j.cellimm.2012.02.009
61. Bughani U, Saha A, Kuriakose A, Nair R, Sadashivarao RB, Venkataraman R, et al. T cell activation and differentiation is modulated by a CD6 domain 1 antibody Itolizumab. *PLoS ONE.* (2017) 12:e0180088. doi: 10.1371/journal.pone.0180088
62. Cai Y, Dai Y, Wang Y, Yang Q, Guo J, Wei C, et al. Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis. *EBioMedicine.* (2020) 53:102686. doi: 10.1016/j.ebiom.2020.102686
63. Ju Y, Kim S-S, Lee K, Park S, Jin H. Association of the Genetic Polymorphisms for CD247 Gene and Tuberculosis Case. *Biomed Sci Lett.* (2020) 26:22–7. doi: 10.15616/BSL.2020.26.1.22
64. Gebremicael G, Kassa D, Quinten E, Alemayehu Y, Gebreegziabier A, Belay Y, et al. Host gene expression kinetics during treatment of tuberculosis in HIV-coinfected individuals is independent of highly active antiretroviral therapy. *J Infect Dis.* (2018) 218:1833–46. doi: 10.1093/infdis/jiy404

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

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

Copyright © 2022 Alsulaimany, Zabermaawi, Almukadi, Parambath, Shetty, Vaidyanathan, Elango, Babanaganapalli and Shaik. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Increased CDCA2 Level Was Related to Poor Prognosis in Hepatocellular Carcinoma and Associated With Up-Regulation of Immune Checkpoints

Mengying Tang¹, Mingchu Liao², Xiaohong Ai³ and Guicheng He^{2*}

¹ Department of Infectious Disease, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, China, ² Department of Medical Oncology, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, China, ³ Department of Radiation Oncology, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, China

OPEN ACCESS

Edited by:

Thirumal Kumar D.,
Meenakshi Academy of Higher
Education and Research, India

Reviewed by:

Miao Liu,
Harvard Medical School,
United States
Yu Wang,
Zhejiang Chinese Medical
University, China
Qin Sun,
Southwest Medical University, China

*Correspondence:

Guicheng He
313397526@qq.com

Specialty section:

This article was submitted to
Precision Medicine,
a section of the journal
Frontiers in Medicine

Received: 10 September 2021

Accepted: 22 November 2021

Published: 07 March 2022

Citation:

Tang M, Liao M, Ai X and He G (2022)
Increased CDCA2 Level Was Related
to Poor Prognosis in Hepatocellular
Carcinoma and Associated With
Up-Regulation of Immune
Checkpoints. *Front. Med.* 8:773724.
doi: 10.3389/fmed.2021.773724

Background: Cell division cycle-associated protein 2 (CDCA2) is a member of cell cycle-related proteins. CDCA2 plays a role in the regulation of protein phosphatase 1 (PP1) γ -dependent DNA damage response (DDR) and H3 phosphorylation. CDCA2 promotes the tumorigenesis and development of several types of cancers by promoting the proliferation of tumor cells. However, the relationship between CDCA2 expression and the clinicopathological characteristics of hepatocellular carcinoma (HCC) is unknown.

Methods: Gene expression information and clinical data were downloaded from The Cancer Genome Atlas (TCGA) database. The expression of CDCA2 and its correlation to clinical characteristics in HCC were analyzed. The expression level of CDCA2 was validated in HCC cell lines. The relationship between CDCA2 expression and the survival of patients with HCC was analyzed by using Kaplan–Meier method. The prognostic value of CDCA2 in HCC was estimated by Cox regression analysis. The expression difference of CDCA2 between HCC and normal tissues and its correlation to survival were verified in independent datasets. Gene set enrichment analysis (GSEA) was used to screen the CDCA2-related signaling pathways.

Results: Cell division cycle-associated protein 2 expression was upregulated in HCC tissues ($p < 0.001$) and increased CDCA2 was correlated to increased T stage, pathologic stage, histologic grade, and alpha-fetoprotein (AFP) level ($p < 0.001$). In addition, CDCA2 was overexpressed in HCC cell lines HepG2 and LM3. High CDCA2 expression level was associated with poor overall survival [hazard ratio (HR) = 1.69; 95% CI , 1.20–1.40, $p = 0.003$], disease specific survival ($HR = 1.73$; 95% CI , 1.11–2.71, $p = 0.016$), and progress free interval ($HR = 1.74$; 95% CI , 1.30–2.34, $p < 0.001$). Overexpression of CDCA2 and its correlation to poor survival in HCC were verified in Gene Expression Omnibus (GEO) datasets and Kaplan–Meier plotter database. Increased CDCA2 expression was associated with upregulation of PD-L1 (Spearman's coefficient = 0.207, $p < 0.001$), PD-L2 (Spearman coefficient's = 0.118, $p < 0.05$), and CTLA4 (Spearman's coefficient = 0.355, $p < 0.001$). GSEA showed that homologous

recombination pathway, insulin signaling pathway, mitogen-activated protein kinase (MAPK) pathway, mismatch repair pathway, mechanistic target of rapamycin (mTOR) pathway, Notch pathway, T cell receptor pathway, toll like receptor pathway, and WNT pathway were enriched in CDCA2 high expression phenotype.

Conclusion: Cell division cycle-associated protein 2 may serve as an independent biomarker for poor prognosis in HCC and increased CDCA2 expression was associated with upregulation of immune checkpoints.

Keywords: liver cancer, hepatocellular carcinoma, CDCA2, survival, prognosis

INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the most common malignant tumors in the world, with a high cancer-related mortality. Each year, there are about 840,000 new cases of HCC and about 780,000 HCC related-deaths worldwide. The prognosis of HCC is poor, with a survival interval of 6–20 months without treatment (1, 2). For patients with resectable disease, surgical resection is the recommended treatment. However, recurrence occurs in about 70% of patients (3). Systemic therapy is the standard treatment for patients with inoperable or recurrent disease, such as sorafenib, lenvatinib, and immune checkpoint inhibitor (4). However, the prognosis of these patients are poor, with a 5-year survival rate of <8% (5). Thus, it is an urgent need to find new biomarkers for the diagnosis, treatment, and prognosis.

Cell division cycle-associated protein 2 (CDCA2) is a member of cell cycle-related proteins. It is reported that CDCA2 plays a role in the regulation of protein phosphatase 1 (PP1) γ -dependent DNA damage response (DDR) by forming a complex with PP1 γ (6). In addition, CDCA2 regulates H3 phosphorylation in a PP1 dependent manner (7). CDCA2 promotes the tumorigenesis and development of prostate cancer, malignant melanoma, renal cancer, and other malignant tumors by promoting the proliferation of tumor cells (6, 8–10). CDCA2 participates in cell cycle regulation. It was reported that CDCA2 expression level affected the activation of DNA damage checkpoint. Cell cycle checkpoints are induced by DNA damage and cause cell cycle arrest (11, 12). Thus, CDCA2 plays an important role in the regulation of cell cycle progression. Previous studies have shown that CDCA2 is upregulated and associated with poor prognosis in some tumors, such as lung cancer (13), breast cancer (14), and pancreatic cancer (15). However, there are few reports about the correlation between CDCA2 expression and the clinicopathological characteristics of HCC.

To explore the expression pattern and the prognostic value of CDCA2 in HCC, we performed the current study.

METHODS

Datasets and Clinical Information

Cell division cycle-associated protein 2 expression data of normal liver tissue (50 cases) and HCC tissues (374 cases), and the clinical data of patients with HCC were downloaded from The Cancer Genome Atlas database (TCGA-LIHC). The expression

information of CDCA2 and patient information used in the current study were obtained from public database and therefore ethical approval was not required. R software (version 3.6.3) was used to perform the analysis. The difference of expression is visualized by dot graphs and box graphs.

RNA Extraction and Quantitative Real-Time PCR Analysis of CDCA2 Expression in HCC Cell Lines

Hepatocellular carcinoma cell lines HepG2 and SNU182 and normal liver cell line THLE-3 were purchased from American Type Culture Collection (ATCC) cell bank. Total RNA of the cell lines was extracted using the TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and reverse transcription was performed to obtain cDNA. Primer sequences of CDCA2 were shown as follows: forward, 5'-ATGACCGGCTGTCTGGAAT-3', and reverse, 5'-GCTGAGACCTTCCTTTCTGGT-3'. According to the instructions of manufacturer of the SYBR Green reagent (ABI, CA, USA), quantitative real-time PCR (qRT-PCR) was performed to examine the expression of CDCA2 mRNA.

Verification of CDCA2 Expression and Its Correlation With Survival by GEO Datasets and Kaplan–Meier Plotter

Microarray data and RNA sequencing data were downloaded from GEO database. The terms, such as “liver,” or “hepatocellular” and “cancer,” “carcinoma,” or “neoplasm” were used for the search. GSE27150, GSE54236, GSE56140, GSE64041, and GSE76427 were downloaded. GSE56140, GSE76427, and GSE64041 were used to validate the CDCA2 expression difference between normal tissues and HCC tissues. GSE27150, GSE54236, and GSE76427 were used to validate the relationship between CDCA2 expression and survival. Meta-analysis was performed to verify the hazard ratio (HR) of CDCA2 expression to survival. The combined value was calculated by HR and 95% CI. Heterogeneity between datasets was assessed by using the τ^2 and I^2 test. If $I^2 > 50\%$, the random-effects model was used, otherwise, the fixed-effects model was used. HCC data from Kaplan–Meier plotter database (<https://kmplot.com/analysis/>) were used to validate the relationship between CDCA2 expression and survival.

Gene Set Enrichment Analysis

Patients were classified as CDCA2-high group and CDCA2-low group, using the median expression level of CDCA2 as cutoff value. Gene set enrichment analysis (GSEA) was conducted to assess the potential mechanism of CDCA2 in HCC. The *c2.cp.kegg.v6.2.symbols.gmt* was used as reference gene set. Parameter of gene set permutation for each analysis was 1,000. The significance of enriched gene sets was estimated by nominal *p*-value and false discovery rate (FDR) *Q*-value.

Statistical Analysis

Statistical analyses were conducted by using R software (version 3.6.3). Results were considered as statistically significant if $p < 0.05$. First, the expression of CDCA2 in normal tissues and tumor tissues was compared by Wilcoxon rank sum test. The correlation between CDCA2 expression and clinicopathological characteristics was examined by logistic regression analysis. Then, the relationship between CDCA2 expression and survival in HCC was estimated by Kaplan–Meier method. The prognostic value of CDCA2 in HCC was estimated by the univariate and multivariate Cox regression analysis.

RESULTS

CDCA2 Was Overexpressed in HCC Tissues

The expression level of CDCA2 in 50 adjacent noncancer tissues and 374 HCC tissues was compared. It was shown that expression of CDCA2 was significantly higher in HCC tissues ($p < 0.001$) (Figure 1A). In fifty pairs of adjacent noncancerous and HCC tissues, CDCA2 expression was increased in HCC tissues in comparison with noncancerous tissues ($p < 0.001$) (Figure 1B). In short, CDCA2 was overexpressed in HCC tissues.

CDCA2 Was Upregulated in HCC Cell Lines

To verify the upregulation of CDCA2 expression in HCC, we compared the expression of CDCA2 mRNA in HCC cell lines (HepG2 and SNU182) and normal liver epithelial cell line (THLE-3). Results showed that CDCA2 mRNA was upregulated in both HepG2 and SNU182 cell lines (Figure 1L).

Correlations Between CDCA2 Expression Level and Clinicopathological Characteristics in Patients With HCC

Expression of CDCA2 in patients with HCC with different clinicopathological characteristics was analyzed. As shown in Figures 1C,D, the expression level of CDCA2 was increased as T stage (Kruskal–Wallis test, $p < 0.001$) and pathologic stage (Kruskal–Wallis test, $p < 0.001$) was increased. CDCA2 expression level in poorly differentiated groups (G3 and G4) was significantly higher than that in well differentiated groups (G1 and G2) ($p < 0.001$) (Figure 1E). Patients with high AFP level ($p < 0.001$) (Figure 1F) and low body mass index (BMI) ($p < 0.05$) (Figure 1G) also had higher CDCA2 expression level. Logistic regression analysis was performed to estimate the relationships between CDCA2 expression and clinicopathological characteristics of patients with HCC. It was revealed that an increased CDCA2 expression was significantly

related to age [for >60 years vs. ≤ 60 years, odds ratio (OR) = 0.505; 95% CI, 0.333–0.761, $p = 0.001$], T stage (for T2–T4 vs. T1, OR = 2.541; 95% CI, 1.678–3.875, $p < 0.001$), pathological stage (for stage II–IV vs. stage I, OR = 2.359; 95% CI, 1.541–3.636, $p < 0.001$), AFP level (for >400 ng/ml vs. <400 ng/ml, OR = 3.558; 95% CI, 1.969–6.667, $p < 0.001$) and histologic grade (for G3–4 vs. G1–2, OR = 3.375; 95% CI, 2.170–5.314, $p < 0.001$) (Table 1).

Comparison of Survival in CDCA2-High and CDCA2-Low Patients

The median expression level of CDCA2 was 1.008 and it was used as the cutoff value. Patients with CDCA2 expression level higher than the cutoff value were considered as CDCA2 high expression, otherwise they were considered as CDCA2 low expression. Kaplan–Meier method was used to compare the survival of patients with high and low expression of CDCA2 from TCGA database. Patients with high CDCA2 expression level had worse overall survival ($HR = 1.69$; 95% CI, 1.20–1.40, $p = 0.003$), disease specific survival ($HR = 1.73$; 95% CI, 1.11–2.71, $p = 0.016$), and progress free interval ($HR = 1.74$; 95% CI, 1.30–2.34, $p < 0.001$) (Figures 1H–J).

Verification of CDCA2 Overexpression in HCC by GEO Datasets

Cell division cycle-associated protein 2 expression in GSE56140, GSE76427, and GSE64041 was analyzed. The expression difference of CDCA2 between normal tissues and HCC tissues was compared. We found that CDCA2 expression was increased in HCC tissues (Figures 2A–C), which was consistent with the results of TCGA database.

Verification the Relationship Between CDCA2 Expression and Survival in HCC

The survivals of patients with high and low CDCA2 expression in GSE27150, GSE54236, and GSE76427 were compared. In GSE54236 cohort, patients with high CDCA2 expression showed significantly worse survival than patients with low CDCA2 expression ($p = 0.030$) (Figure 2E). In GSE27150 and GSE76427 cohorts, survivals were not significantly different between CDCA2 high and CDCA2 low expression patients (Figures 2D,F). To further confirm the correlation of CDCA2 expression with survival in patients from GEO datasets, meta-analysis was conducted. Meta-analysis result of the GSE27150, GSE54236, and GSE76427 showed that CDCA2 overexpression was not associated with poor survival [combined $HR = 1.07$ (95% CI: 0.56–2.04)] (Figure 2G). After analyzing the result, we found that the homogeneity between studies was poor, with $I^2 = 85\%$. We further analyzed the array data of the three GEO datasets and found that GSE27150 did not provide normalization information and normalization method about the data. The poor homogeneity was mainly due to GSE27150. We excluded GSE27150 and performed survival analysis using the GSE54236 and GSE76427 datasets. The result indicated that high expression of CDCA2 was associated with better survival ($p < 0.001$) (Figure 2H). The result was inconsistent with the previous results from TCGA. We analyzed the data of the two datasets and we found that 69.3% of patients in high CDCA2 group were

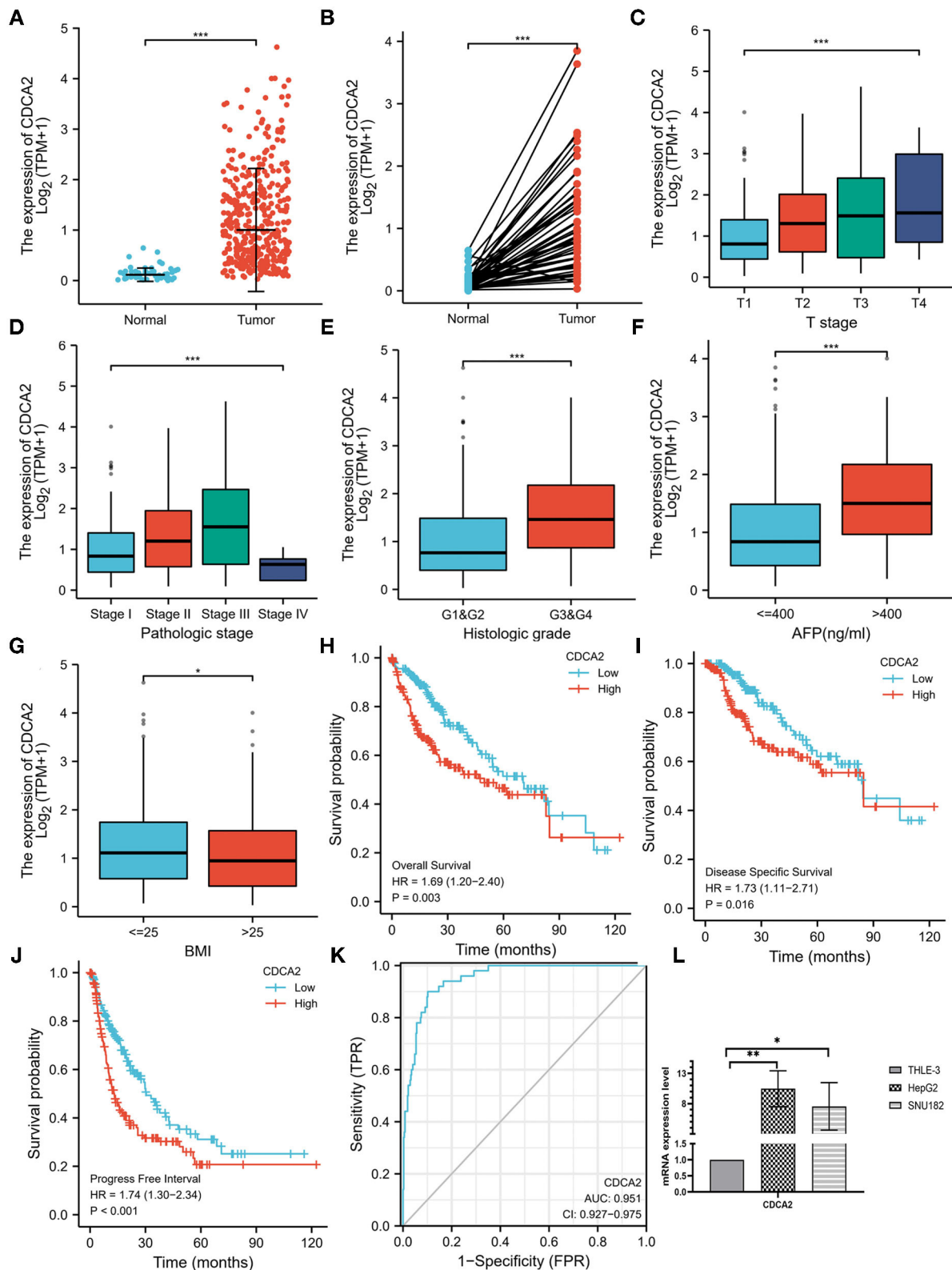


FIGURE 1 | Cell division cycle-associated protein 2 (CDCA2) expression in hepatocellular carcinoma (HCC) and the correlations between CDCA2 expression and clinicopathological characteristic. **(A)** Cell division cycle-associated protein 2 expression in HCC tissues and normal tissues (Wilcoxon rank sum test, *** $p < 0.001$). **(B)** Cell division cycle-associated protein 2 expression in HCC tissues and adjacent noncancerous tissues (Wilcoxon signed rank test, *** $p < 0.001$). **(C)** Expression (Continued)

FIGURE 1 | level of CDCA2 in patients with different T stages (Kruskal–Wallis test, $***p < 0.001$). **(D)** Expression level of CDCA2 in patients with different pathological stages (Kruskal–Wallis test, $***p < 0.001$). **(E)** Expression level of CDCA2 in patients with different histologic grades (Wilcoxon rank sum test, $***p < 0.001$). **(F)** Expression level of CDCA2 in patients with different AFP levels (Wilcoxon rank sum test, $***p < 0.001$). **(G)** Expression level of CDCA2 in patients with different body mass index (BMI) (Wilcoxon rank sum test, $***p < 0.05$). **(H)** Overall survivals of patients with high and low CDCA2 expression (log-rank test, $p = 0.003$). **(I)** Disease specific survivals of patients with high and low CDCA2 expression (log-rank test, $p = 0.016$). **(J)** Progress free intervals of patients with high and low CDCA2 expression (log-rank test, $p < 0.001$). **(K)** A receiver operating characteristic (ROC) curve and the area under the curve (AUC) of CDCA2 in HCC. **(L)** Cell division cycle-associated protein 2 mRNA was upregulated in both HepG2 and SNU182 cell lines (t -test, $*p < 0.05$, $**p < 0.01$).

TABLE 1 | Logistic regression analysis was performed to estimate the relationships between cell division cycle-associated protein 2 (CDCA2) expression and clinicopathological characteristics.

Characteristics	Total (N)	Odds Ratio (OR)	P-value
Age (>60 vs. ≤60)	373	0.505 (0.333–0.761)	0.001[†]
Gender (Male vs. Female)	374	0.843 (0.545–1.300)	0.439
T stage (T2–T4 vs. T1)	371	2.541 (1.678–3.875)	<0.001[†]
Pathologic stage (Stage II–IV vs. Stage I)	350	2.359 (1.541–3.636)	<0.001[†]
BMI (>25 vs. ≤25)	337	0.817 (0.532–1.253)	0.355
AFP(ng/ml) (>400 vs. ≤400)	280	3.558 (1.969–6.667)	<0.001[†]
Histologic grade (G3–4 vs. G1–2)	369	3.375 (2.170–5.314)	<0.001[†]

[†] $p < 0.05$.

lost to follow-up while only 11.4% of patients in low CDCA2 group were lost to follow-up. The unbalanced loss of follow up rate between the two groups may affect the survival rate, and the high loss of follow-up rate in the high CDCA2 group may make the calculated survival rate higher than the actual survival rate. In addition, meta-analysis of the GSE54236 and GSE76427 showed that homogeneity between the two datasets was good ($I^2 = 27\%$) and increased CDCA2 was associated with poor clinical outcome (combined $HR = 1.61$ (95% CI: 1.30–1.99). (Supplementary Figure S2). HCC data from Kaplan–Meier plotter database (<https://kmplot.com/analysis/>) were used to validate the relationship between CDCA2 expression and survival. It was indicated that patients with high CDCA2 expression showed poor overall survival ($HR = 1.94$; 95% CI, 1.36–2.76, $p < 0.001$) (Figure 2I) and progress free survival ($HR = 1.81$; 95% CI, 1.34–2.43, $p < 0.001$) (Figure 2L). In stage I–II subgroup and stage III–IV subgroup patients, CDCA2 overexpression was related to poor overall survival (for stage I–II subgroup, $HR = 1.87$; 95% CI, 1.08–3.21, $p = 0.022$; for stage III–IV subgroup, $HR = 2.16$; 95% CI, 1.20–3.90, $p = 0.0089$) (Figures 2J,K) and progress free survival (for stage I–II subgroup, $HR = 1.68$; 95% CI, 1.12–2.52, $p = 0.011$; for stage III–IV subgroup, $HR = 2.14$; 95% CI, 1.10–4.16, $p = 0.021$) (Figures 2M,N). The results were consistent with the results of TCGA database.

Diagnostic and Prognostic Values of CDCA2 in HCC

A receiver operating characteristic (ROC) curve was plotted and the area under the curve (AUC) was calculated to examine the diagnostic value of CDCA2 in HCC. The ROC showed a

sensitivity of 0.900 and a specificity of 0.898 and the AUC was 0.951 (Figure 1K). Univariate and multivariate analysis were used to estimate the correlation between clinicopathological characteristics and prognosis of HCC. Univariate analysis showed that CDCA2 expression ($HR = 1.694$; 95% CI, 1.196–2.401; $p = 0.003$), T stage ($HR = 2.126$; 95% CI, 1.481–3.052; $p < 0.001$), and pathologic stage ($HR = 2.090$; 95% CI, 1.429–3.055; $p < 0.001$) were related to poor survival (Table 2). Multivariate analysis showed that CDCA2 expression ($HR = 1.613$; 95% CI, 1.108–2.349; $p = 0.013$) was independently related to survival (Table 2). In summary, CDCA2 expression was an independent prognostic factor for HCC and increased CDCA2 expression was related to poor survival.

Overexpression of CDCA2 Was Related to Increased Expression of Immune Checkpoint

Spearman's correlation analysis was performed to estimate the relation of CDCA2 expression to immune checkpoint expression, such as PD-L1, PD-L2, and CTLA4. The results showed that overexpression of CDCA2 was associated with increased expression of PD-L1 (Spearman's coefficient = 0.207, $p < 0.001$), PD-L2 (Spearman's coefficient = 0.118, $p < 0.05$), and CTLA4 (Spearman's coefficient = 0.355, $p < 0.001$) (Figure 3).

Identification of CDCA2-Related Pathways

Patients were classified into CDCA2 high group and CDCA2 low group according to the median value of CDCA2 expression. CDCA2-related pathways were screened by GSEA. Results showed that homologous recombination pathway, insulin signaling pathway, mitogen-activated protein kinase (MAPK) pathway, mismatch repair pathway, mTOR pathway, Notch pathway, T cell receptor pathway, toll like receptor pathway, and WNT pathway were enriched in CDCA2 high expression phenotype (Table 3 and Figure 4).

DISCUSSION

Cell division cycle-associated protein 2 participates in cell cycle regulation. It was reported that CDCA2 expression level affected the activation of DNA damage checkpoint. Cell cycle checkpoints are induced by DNA damage and cause cell cycle arrest (11, 12). CDCA2 participates in chromatin remodeling by regulating histone H3 de-phosphorylation (7). Thus, CDCA2 plays an important role in the regulation of cell cycle progression. Previous studies have shown that CDCA2 is upregulated and associated with poor prognosis in some tumors, such as lung cancer (13), breast cancer (14), and pancreatic cancer (15).

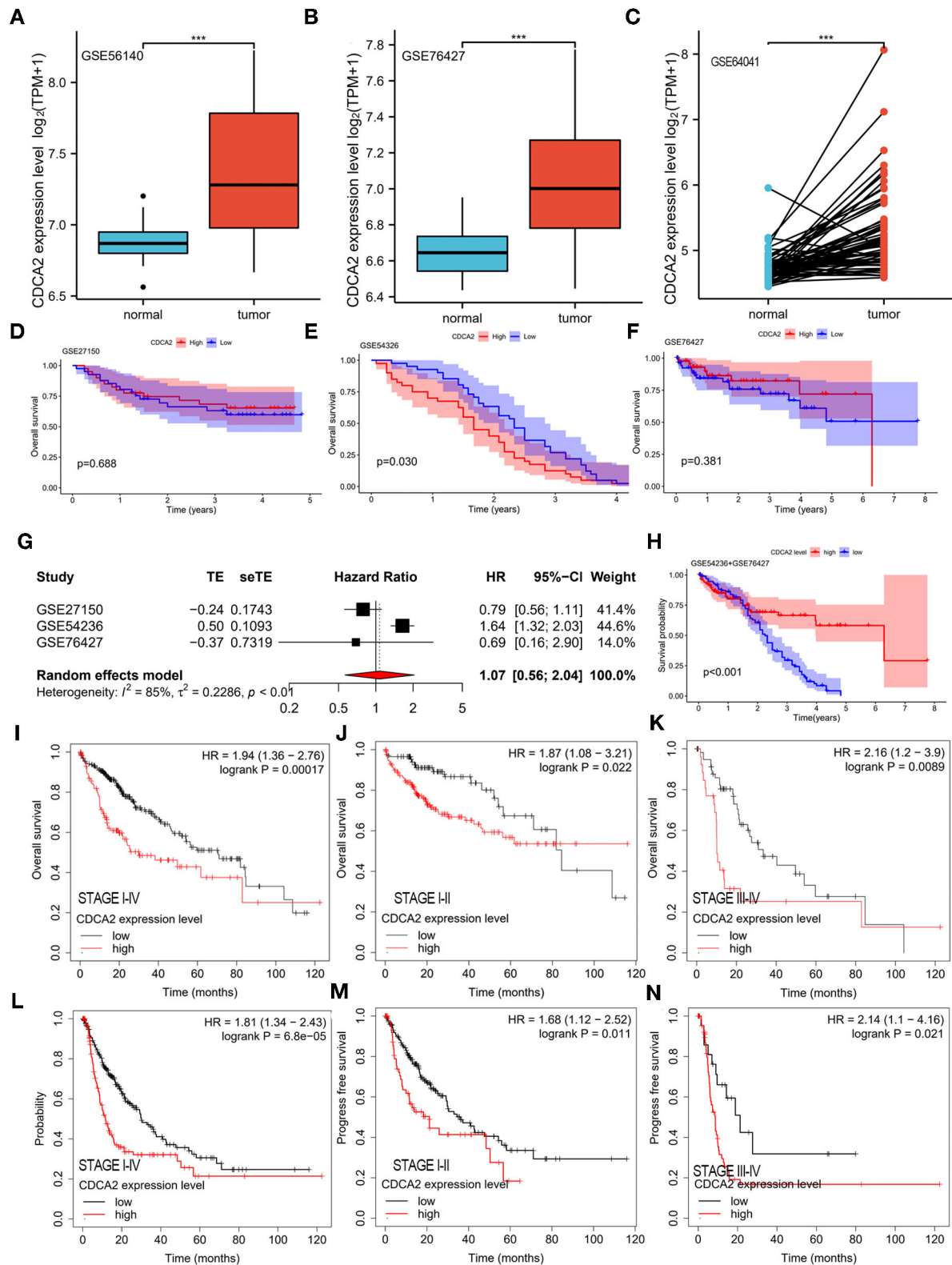


FIGURE 2 | Verification of CDCA2 overexpression in HCC and the correlation between CDCA2 expression with survival by independent datasets. Cell division cycle-associated protein 2 expression in (A) GSE56140, (B) GSE76427, and (C) GSE64041. ***p < 0.001. The overall survivals of patients with high and low CDCA2 expression in (D) GSE27150, (E) GSE54326, and (F) GSE76427. (G) Meta-analysis of hazard ratio (HR) of CDCA2 overexpression to survival. (H) Overall survival of patients with high and low CDCA2 expression in GSE54326+GSE 76427 cohort. (I) Overall survivals of patients with high and low CDCA2 expression

(Continued)

FIGURE 2 | in Kaplan–Meier plotter database. **(J)** Overall survivals of stage I–II subgroup patients with high and low CDCA2 expression in Kaplan–Meier plotter database. **(K)** Overall survivals of stage III–IV subgroup patients with high and low CDCA2 expression in Kaplan–Meier plotter database. **(L)** Progress free survivals of patients with high and low CDCA2 expression in Kaplan–Meier plotter database. **(M)** Progress free survivals of stage I–II subgroup patients with high and low CDCA2 expression in Kaplan–Meier plotter database. **(N)** Progress free survivals of stage III–IV subgroup patients with high and low CDCA2 expression in Kaplan–Meier plotter database.

TABLE 2 | Univariate and multivariate analysis were used to estimate the correlation between clinicopathological characteristics and prognosis of hepatocellular carcinoma (HCC).

Characteristics	Total (N)	Univariate analysis		Multivariate analysis	
		HR (95% CI)	P-value	HR (95% CI)	P-value
Age	373	1.205	0.295		
(>60 vs. ≤60)		(0.850–1.708)			
Gender	373	0.793	0.200		
(Male vs. Female)		(0.557–1.130)			
BMI	336	0.798	0.235		
(>25 vs. ≤25)		(0.550–1.158)			
T stage	370	2.126	<0.001[†]	0.731	0.757
(T2–T4 vs. T1)		(1.481–3.052)		(0.101–5.302)	
Pathologic stage	349	2.090	<0.001[†]	2.658	0.337
(Stage II–IV vs. Stage I)		(1.429–3.055)		(0.362–19.545)	
AFP(ng/ml)	279	1.075	0.772		
(>400 vs. ≤400)		(0.658–1.759)			
Histologic grade	368	1.091	0.636		
(G4–3 vs. G1–2)		(0.761–1.564)			
CDCA2	373	1.694	0.003[†]	1.613	0.013[†]
(High vs. Low)		(1.196–2.401)		(1.108–2.349)	

[†] $p < 0.05$.

In the current study, we analyzed the expression pattern of CDCA2 and its diagnostic and prognostic value in HCC. To explore the potential mechanism by which CDCA2 regulates the tumorigenesis and development of HCC, we analyzed the CDCA2-high phenotype related signal pathways by GSEA. In the TCGA-LIHC cohort, we found that CDCA2 was upregulated in HCC and increased CDCA2 expression was associated with poor prognosis of patients with HCC. To validate the bioinformatic analysis results of the TCGA-LIHC cohort, we searched the GEO database and analyzed CDCA2 expression level in normal tissue and HCC tissue and its association with prognosis. We got consistent results with the results of TCGA-LIHC cohort.

Some reports have indicated that CDCA2 was associated with poor survival in HCC and the correlation between pathologic stage and histologic grade with CDCA2 expression was also reported (15–17). However, the relationship between CDCA2 expression and other clinical features was not analyzed. In the current study, we analyzed the correlation between CDCA2 expression level and clinicopathological features, such as T stage, lymph node invasion, distant metastasis, pathologic stage, histological grade, AFP level, and BMI. Logistic regression showed that CDCA2 expression was significantly associated with

histological grade, AFP level, T stage, and pathologic stage. CDCA2 was increased as histological grade, AFP level, T stage, and pathologic stage increased. These results suggested that CDCA2 participated in the development of HCC. An ROC curve showed that CDCA2 had high diagnostic value for HCC, with an AUC of 0.951.

Univariate analysis showed that CDCA2 expression level, T stage, and pathologic stage may predict poor prognosis of HCC. Multivariate regression analysis further verified that CDCA2 had an independent prognostic value for HCC. The results were consistent with previous reports (15–17). Wang Y et al. demonstrated that low methylation of CDCA2 was related to poor survival. However, they did not study the relationship between CDCA2 expression level and clinical prognosis and the results were not verified by independent dataset (16). Though Wang Z also indicated that increased CDCA2 was related to poor survival in HCC, they did not verify the results by independent dataset (17). Wu B et al. showed that upregulation of CDCA2 was related to poor survival in HCC. They used only one dataset to validate the upregulation of CDCA2 and the correlation between CDCA2 expression and survival (15). However, the sample size was small. The validation dataset contained only 14 pairs of HCC tissues and adjacent tissues and 64 cases of patients with HCC (15). In the current study, we used three independent datasets to verify the upregulation of CDCA2 in HCC. The validation cohort contained larger sample size than previous study. GSE56140 contained 34 normal tissues and 35 tumor tissues. GSE76427 contained 52 normal tissues and 155 tumor tissues. GSE64041 contained 60 pairs of HCC tissues and adjacent tissues. The three independent datasets showed the consistent results. As these three datasets did not contain prognosis information, we used the other three datasets (GSE27150, GSE54236, and GSE76427) to validate the correlation between CDCA2 expression and clinical outcomes. The result from GSE27150 and GSE75427 showed that the prognosis of patients with high and low CDCA2 expression level did not have statistical difference. However, the result from GSE 54236 showed that patients with high CDCA2 expression had worse clinical outcome. The inconsistent results may be attributed to bias introduced by the small sample size in the GSE27150 and GSE76427 datasets. To further analyze the results, meta-analysis was performed. However, meta-analysis indicated that high CDCA2 expression was not related to poor survival. It should be noted that poor heterogeneity existed between the three datasets, with $I^2 = 85\%$. After the data of the three dataset, we found that GSE54236 and GSE76427 were both normalized by the same method (robust spline normalization, RSN) while normalization information of GSE27150 was not provided. After excluding GSE27150 and survival analysis of GSE54236 and GSE76427 by Kaplan–Meier method showed that increased CDCA2 was associated with better survival. The inconsistent

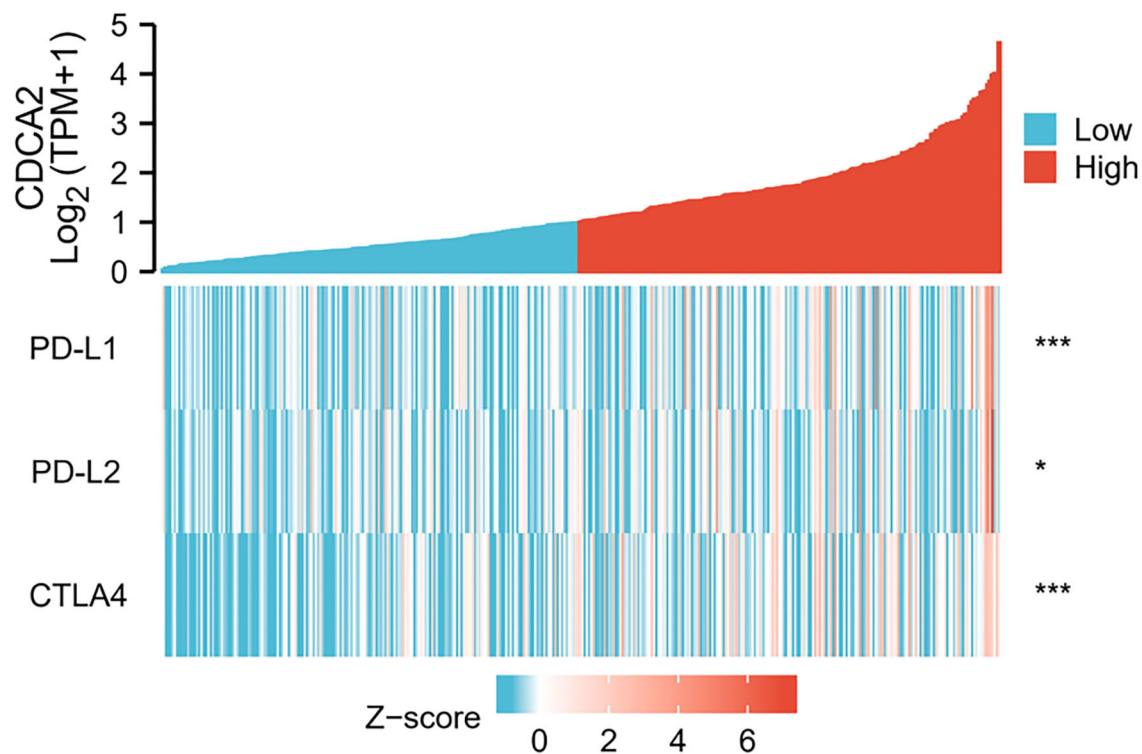


FIGURE 3 | Correlation of CDCA2 with expression level of immune checkpoints. (* $p < 0.05$; *** $p < 0.001$).

TABLE 3 | Cell division cycle-associated protein 2-related pathways screened by gene set enrichment analysis (GSEA).

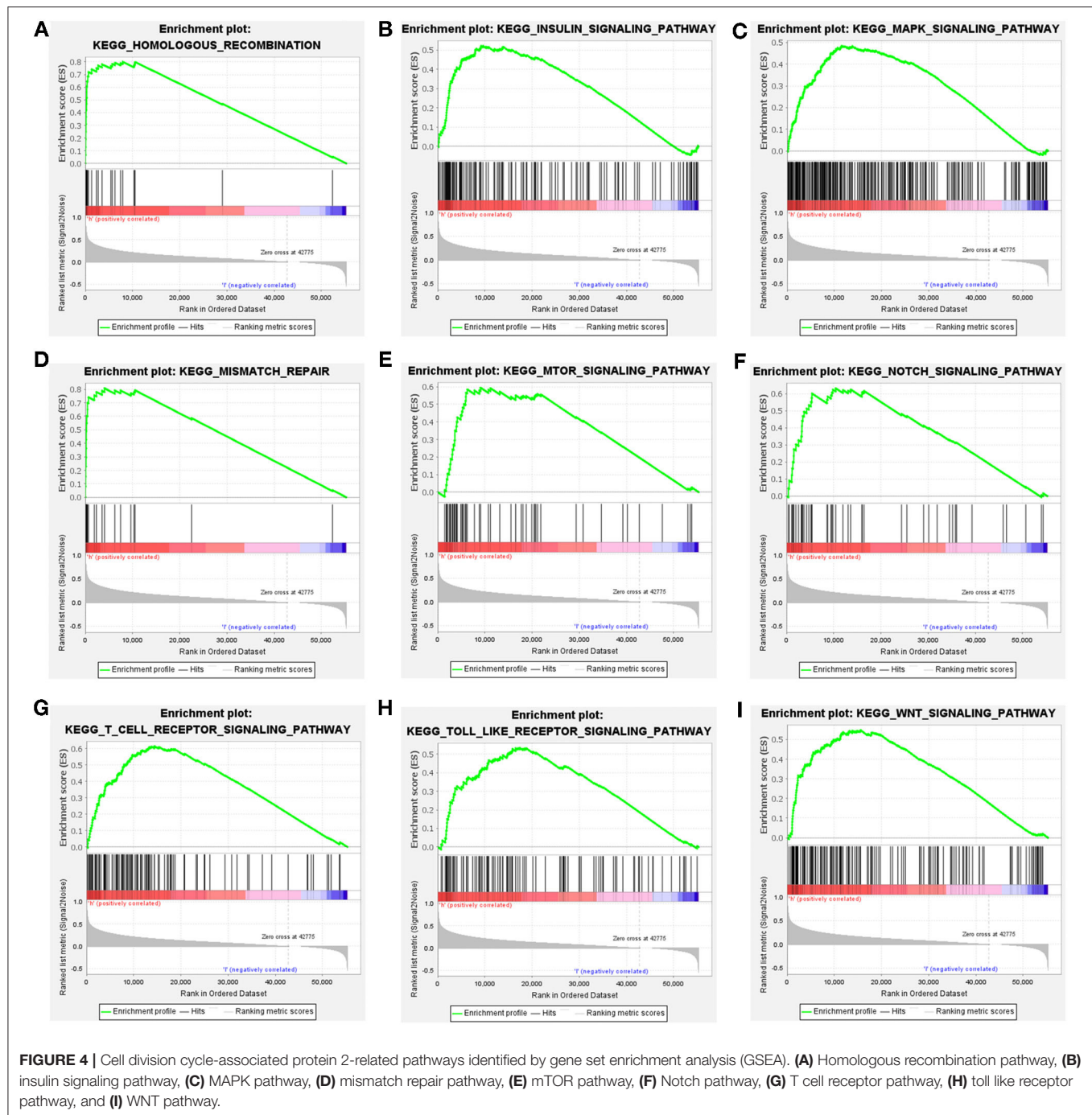
Gene set name	NES	NOM p -value	FDR q -value
KEGG_HOMOLOGOUS_RECOMBINATION	1.99	0.000	0.003
KEGG_INSULIN_SIGNALING_PATHWAY	1.86	0.000	0.008
KEGG_MAPK_SIGNALING_PATHWAY	1.72	0.002	0.023
KEGG_MISMATCH_REPAIR	1.93	0.000	0.005
KEGG_MTOR_SIGNALING_PATHWAY	1.83	0.000	0.011
KEGG_NOTCH_SIGNALING_PATHWAY	1.85	0.004	0.010
KEGG_P53_SIGNALING_PATHWAY	2.10	0.000	0.000
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	1.81	0.002	0.013
KEGG_TOLL LIKE RECEPTOR SIGNALING_PATHWAY	1.74	0.002	0.021
KEGG_WNT_SIGNALING_PATHWAY	1.83	0.000	0.011

result may be due to the obviously higher loss of follow-up rate in the high CDCA2 group. The meta-analysis result of the two datasets confirmed the high CDCA2 expression was related to poor survival.

Some research have studied the mechanisms of CDCA2 in HCC. It was reported that CDCA2 protected against oxidative stress by activating BRCA1-NRF2 pathways in HCC (18).

Li et al. reported that CDCA2 promoted cell proliferation of HCC by activating AKT/CCND1 pathway (19). However, the relationship between immune checkpoint and CDCA2 expression has not been reported. As immune checkpoint inhibitors have become one of the standard treatments for HCC, we estimated whether the CDCA2 expression was related to immune checkpoint expression. Spearman's correlation analysis showed that increased expression of CDCA2 was associated with increased expression of immune checkpoints. It has been indicated that increased immune checkpoint was associated with inhibition of immune cells activity (20). The above results revealed that upregulation of CDCA2 may affect the prognosis by inhibiting immune cell activity.

Gene set enrichment analysis was performed to explore the potential mechanisms of CDCA2 in HCC. We found that homologous recombination pathway, insulin signaling pathway, MAPK pathway, mismatch repair pathway, mTOR pathway, Notch pathway, T cell receptor pathway, toll like receptor pathway, and WNT pathway were enriched in CDCA2 high expression phenotype. Homologous recombination pathway is a signal pathway associated with DNA double-strand breaks repair (21). Drugs targeting homologous recombination deficiency (HRD), such as poly(ADP-ribose) polymerase (PARP) inhibitors, have been proved to have an antitumor activity in some types of tumors, such as breast cancer and ovarian cancer (21, 22). Mismatch repair pathway is another DNA damage repair pathway, which promotes DNA damage response mediated by



ataxia telangiectasia mutated (ATM) and ataxia-telangiectasia mutated (ATR) (23). The previous studies reported that BRCA1 was recruited by CDCA2 (18) and BRCA1 functioned in DNA repair process (24). These results supported that homologous recombination pathway and mismatch repair pathway were enriched in CDCA2 high phenotype. Insulin signaling pathway, which can be activated by IGF-1, promotes cell growth, proliferation, and inhibits apoptosis. It has been indicated that insulin signaling pathway activation was associated with increased risk of breast cancer (25) and colorectal cancer (26).

MAPK pathway is the ubiquitous signal transduction pathway which involves in many processes of life and often alters in many disease (27). MAPK pathway regulates cellular activities during development of cancers, such as cell proliferation, cell apoptosis, and immune escape. Inhibiting the upstream kinase of MAPK pathway has become a therapeutic strategy of some cancer (28). The mTOR pathway involves in the regulation of protein synthesis, glucose metabolism, lipid metabolism, glutamine metabolism, and nucleotide synthesis in cancer cells. The mTOR pathway has become a therapeutic target for cancer

therapy. mTOR inhibitors, such as rapamycin and everolimus, have been approved for the treatment of some types of cancers (29). It was reported that CDCA2 activated AKT related pathways and promoted HCC proliferation (19). mTOR was one of the downstream effectors of PI3K/AKT pathways (30). The current study showed that mTOR pathway was enriched in CDCA2 high phenotype. The result was consistent with previous reports. Notch pathway plays a vital role in promoting tumor development by changing tumor microenvironment and recruiting immunosuppressive cells (31). Moreover, Notch pathway can interact with WNT pathway and promote HCC development (32). Toll like receptors are important factors affecting the immune system and initiation of inflammatory response. It has been revealed that inhibiting toll like receptors suppresses the proliferation of HCC cells (33). The above finding results from GSEA provided information to explore the mechanism by which CDCA2 promoting the development of HCC.

However, some limitations existed in the current study. First, the number of tumor tissues in TCGA and GEO database was much higher than number of normal tissues, which were used as a control. Second, we only analyzed the CDCA2 mRNA expression of the tissue. The protein expression level of CDCA2 was not assessed. And finally, we only explored the potential involved pathways related to CDCA2 by bioinformatic analysis and the molecular mechanism was not explored in depth by molecular biology experiment. In addition, it should be noticed that the meta-analysis of the three GEO datasets indicated that CDCA2 was not associated with survival, though multiple Cox regression analyses pointed out that CDCA2 was associated with survival independently. The relationship between CDCA2 expression and survival should be validated clinically.

In conclusion, we analyzed the CDCA2 expression data of TCGA database and validated the results using independent cohorts from GEO database. The results showed that CDCA2 was increased in HCC and had a high diagnostic power for HCC. Kaplan–Meier analysis and univariate Cox regression analysis indicated that CDCA2 was associated with poor survival for HCC. Increased CDCA2 expression was associated with the upregulation of PD-L1, PD-L2, and CTLA4. In addition, we

also screened the potential signal pathways related to CDCA2 in HCC. However, the prognostic value of CDCA2 in HCC needs further clinical exploration and validation.

PUBLIC DATABASE AND TOOLS USED IN THE CURRENT STUDY

1. The Cancer Genome Atlas (TCGA) database. Link: <https://portal.gdc.cancer.gov/>
2. Kaplan-Meier Plotter database. Link: <https://kmplot.com/analysis/>
3. Gene Expression Omnibus (GEO) database. Link: www.ncbi.nlm.nih.gov/geo/

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material** and access to datasets can be found here: The Cancer Genome Atlas (TCGA) database. Link: <https://portal.gdc.cancer.gov/>, Kaplan-Meier Plotter database. Link: <https://kmplot.com/analysis/>, Gene Expression Omnibus (GEO) database. Link: www.ncbi.nlm.nih.gov/geo/. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

MT, ML, XA, and GH had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis, designed the study, analysis and interpretation of data, and drafting and revising of the manuscript. MT and ML acquisition of data and statistical analysis. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fmed.2021.773724/full#supplementary-material>

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin.* (2019) 69:7–34. doi: 10.3322/caac.21551
2. Yan H, Li Z, Shen Q, Wang Q, Tian J, Jiang Q, et al. Aberrant expression of cell cycle and material metabolism related genes contributes to hepatocellular carcinoma occurrence. *Pathol Res Pract.* (2017) 213:316–21. doi: 10.1016/j.prp.2017.01.019
3. Tabrizian P, Jibara G, Shrager B, Schwartz M, Roayaie S. Recurrence of hepatocellular cancer after resection: patterns, treatments, and prognosis. *Ann Surg.* (2015) 261:947–55. doi: 10.1097/SLA.0000000000000710
4. Yang JD, Hainaut P, Gores GJ, Amadou A, Plymoth A, Roberts LR, et al. Global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat Rev Gastroenterol Hepatol.* (2019) 16:589–604. doi: 10.1038/s41575-019-0186-y
5. Li K, Wang HT, He YK, Guo T. New idea for treatment strategies for Barcelona clinic liver cancer stages based on a network meta-analysis. *Medicine.* (2017) 96:e6950. doi: 10.1097/MD.00000000000006950
6. Li F, Zhang H, Li Q, Wu F, Wang Y, Wang Z, et al. CDCA2 acts as an oncogene and induces proliferation of clear cell renal cell carcinoma cells. *Oncol Lett.* (2020) 19:2466–74. doi: 10.3892/ol.2020.11322
7. Qian J, Lesage B, Beullens M, Van Eynde A, Bollen M. PP1/Repo-man dephosphorylates mitotic histone H3 at T3 and regulates chromosomal aurora B targeting. *Curr Biol.* (2011) 21:766–73. doi: 10.1016/j.cub.2011.03.047
8. Jin WH, Zhou AT, Chen JJ, Cen Y. CDCA2 promotes proliferation and migration of melanoma by upregulating CCAD1. *Eur Rev Med Pharmacol Sci.* (2020) 24:6858–63. doi: 10.26355/eurrev_202006_21675
9. Feng Y, Qian W, Zhang Y, Peng W, Li J, Gu Q, et al. CDCA2 promotes the proliferation of colorectal cancer cells by activating the AKT/CCND1 pathway *in vitro* and *in vivo*. *BMC Cancer.* (2019) 19:576. doi: 10.1186/s12885-019-5793-z
10. Zhang Y, Cheng Y, Zhang Z, Bai Z, Jin H, Guo X, et al. CDCA2 inhibits apoptosis and promotes cell proliferation in prostate cancer and

- is directly regulated by HIF-1 α pathway. *Front Oncol.* (2020) 10:725. doi: 10.3389/fonc.2020.00725
11. Peng A, Lewellyn AL, Schiemann WP, Maller JL. Repo-man controls a protein phosphatase 1-dependent threshold for DNA damage checkpoint activation. *Curr Biol.* (2010) 20:387–96. doi: 10.1016/j.cub.2010.01.020
 12. Ward JF. Complexity of damage produced by ionizing radiation. *Cold Spring Harb Symp Quant Biol.* (2000) 65:377–82. doi: 10.1101/sqb.2000.65.377
 13. Shi R, Zhang C, Wu Y, Wang X, Sun Q, Sun J, et al. CDCA2 promotes lung adenocarcinoma cell proliferation and predicts poor survival in lung adenocarcinoma patients. *Oncotarget.* (2017) 8:19768–79. doi: 10.18632/oncotarget.15519
 14. Xiao B, Chen L, Ke Y, Hang J, Cao L, Zhang R, et al. Identification of methylation sites and signature genes with prognostic value for luminal breast cancer. *BMC Cancer.* (2018) 18:405. doi: 10.1186/s12885-018-4314-9
 15. Wu B, Huang Y, Luo Y, Ma A, Wu Z, Gan Y, et al. The diagnostic and prognostic value of cell division cycle associated gene family in hepatocellular carcinoma. *J Cancer.* (2020) 11:5727–37. doi: 10.7150/jca.46554
 16. Wang Y, Yang Y, Gao H, Ouyang T, Zhang L, Hu J, et al. Comprehensive analysis of CDCAs methylation and immune infiltrates in hepatocellular carcinoma. *Front Oncol.* (2020) 10:566183. doi: 10.3389/fonc.2020.566183
 17. Wang Z, Xu J, Zhang S, Chang L. Expression of cell divisioncycle-associated genes and their prognostic significance in hepatocellular carcinoma. *Int J Clin Exp Pathol.* (2021) 14:151–69.
 18. Wang S, Cao K, Liao Y, Zhang W, Zheng J, Li X, et al. CDCA2 protects against oxidative stress by promoting BRCA1-NRF2 signaling in hepatocellular carcinoma. *Oncogene.* (2021) 40:4368–83. doi: 10.1038/s41388-021-01855-w
 19. Li J, Chen Y, Wang X, Wang C. CDCA2 triggers *in vivo* and *in vitro* proliferation of hepatocellular carcinoma by activating the AKT/CCND1 signaling. *J BUON.* (2021) 26:882–8.
 20. Li B, Chan HL, Chen P. Immune checkpoint inhibitors: basics and challenges. *Curr Med Chem.* (2019) 26:3009–25. doi: 10.2174/0929867324666170804143706
 21. Prakash R, Zhang Y, Feng W, Jasin M. Homologous recombination and human health: the roles of BRCA1, BRCA2, and associated proteins. *Cold Spring Harb Perspect Biol.* (2015) 7:a016600. doi: 10.1101/cshperspect.a016600
 22. Hoppe MM, Sundar R, Tan DSP, Jeyasekharan AD. Biomarkers for homologous recombination deficiency in cancer. *J Natl Cancer Inst.* (2018) 110:704–13. doi: 10.1093/jnci/djy085
 23. Li Z, Pearlman AH, Hsieh P. DNA mismatch repair and the DNA damage response. *DNA Repair.* (2016) 38:94–101. doi: 10.1016/j.dnarep.2015.11.019
 24. Varol U, Kucukzeybek Y, Alacacioglu A, Somali I, Altun Z, Aktas S, et al. BRCA genes: BRCA 1 and BRCA 2. *J BUON.* (2018) 23:862–6.
 25. Goodwin PJ. Insulin in the adjuvant breast cancer setting: a novel therapeutic target for lifestyle and pharmacologic interventions? *J Clin Oncol.* (2008) 26:833–4. doi: 10.1200/JCO.2007.14.7132
 26. Sridhar SS, Goodwin PJ. Insulin-like growth factor axis and colon cancer. *J Clin Oncol.* (2009) 27:165–7. doi: 10.1200/JCO.2008.19.8937
 27. Lee S, Rauch J, Kolch W. Targeting MAPK signaling in cancer: mechanisms of drug resistance and sensitivity. *Int J Mol Sci.* (2020) 21. doi: 10.3390/ijms21031102
 28. Peluso I, Yarla NS, Ambra R, Pastore G, Perry G. MAPK signalling pathway in cancers: olive products as cancer preventive and therapeutic agents. *Semin Cancer Biol.* (2019) 56:185–95. doi: 10.1016/j.semcancer.2017.09.002
 29. Magaway C, Kim E, Jacinto E. Targeting mTOR and metabolism in cancer: lessons and innovations. *Cells.* (2019) 8. doi: 10.3390/cells8121584
 30. Song M, Bode AM, Dong Z, Lee MH. AKT as a therapeutic target for cancer. *Cancer Res.* (2019) 79:1019–31. doi: 10.1158/0008-5472.CAN-18-2738
 31. Jackstadt R, van Hooft SR, Leach JD, Cortes-Lavaud X, Lohuis JO, Ridgway RA, et al. Epithelial NOTCH signaling rewires the tumor microenvironment of colorectal cancer to drive poor-prognosis subtypes and metastasis. *Cancer Cell.* (2019) 36:319–36.e7. doi: 10.1016/j.ccell.2019.08.003
 32. Kim W, Khan SK, Gvozdenovic-Jeremic J, Kim Y, Dahlman J, Kim H, et al. Hippo signaling interactions with Wnt/beta-catenin and notch signaling repress liver tumorigenesis. *J Clin Invest.* (2017) 127:137–52. doi: 10.1172/JCI88486
 33. Huang Y, Cai B, Xu M, Qiu Z, Tao Y, Zhang Y, et al. Gene silencing of Toll-like receptor 2 inhibits proliferation of human liver cancer cells and secretion of inflammatory cytokines. *PLoS ONE.* (2012) 7:e38890. doi: 10.1371/journal.pone.0038890

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

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

Copyright © 2022 Tang, Liao, Ai and He. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership