frontiers | Research Topics

Computational genomics and structural bioinformatics in personalized medicines, volume II

Edited by
George Priya Doss C., Thirumal Kumar D.
and Balu Kamaraj

Published in
Frontiers in Medicine
Frontiers in Physiology
Frontiers in Molecular Biosciences

ISSN 1664-8714
ISBN 978-2-8325-3833-3
DOI 10.3389/978-2-8325-3833-3

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, volume II

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. (2023). *Computational genomics and structural bioinformatics in personalized medicines, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-3833-3

Table of contents

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 18 November 2022
DOI 10.3389/fmed.2022.1036874

OPEN ACCESS

EDITED BY
Balu Kamaraj,
Imam Abdulrahman Bin Faisal University, Saudi Arabia

REVIEWED BY
Qingbing Zheng,
Xiamen University, China
E. Srinivasan,
Indian Institute of Science (IISc), India
Karthick Vasudevan,
Reva University, India

*CORRESPONDENCE
Luke Elizabeth Hanna
hannatrc@yahoo.com
Umashankar Vetrivel
umashankar.v@icmr.gov.in

†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 05 September 2022
ACCEPTED 26 October 2022
PUBLISHED 18 November 2022

CITATION
Vivekanandan S, Vetrivel U and Hanna LE (2022) Design of human immunodeficiency virus-1 neutralizing peptides targeting CD4-binding site: An integrative computational biologics approach. *Front. Med.* 9:1036874.
doi: 10.3389/fmed.2022.1036874

Design of human immunodeficiency virus-1 neutralizing peptides targeting CD4-binding site: An integrative computational biologics approach

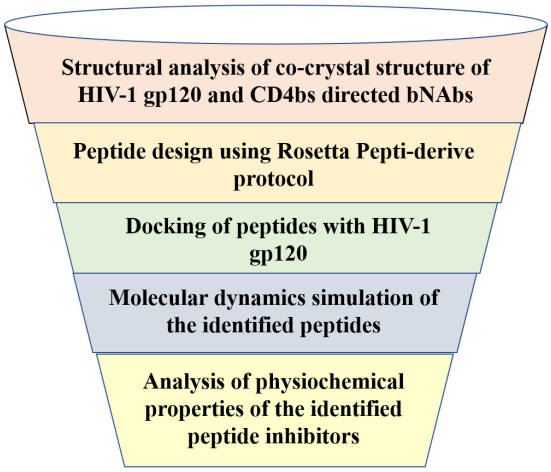
Sandhya Vivekanandan[1,2], Umashankar Vetrivel[1*†] and Luke Elizabeth Hanna[1*†]

[1]Department of Virology and Biotechnology, ICMR-National Institute for Research in Tuberculosis, Chennai, India, [2]University of Madras, Chennai, India

Peptide therapeutics have recently gained momentum in antiviral therapy due to their increased potency and cost-effectiveness. Interaction of the HIV-1 envelope gp120 with the host CD4 receptor is a critical step for viral entry, and therefore the CD4-binding site (CD4bs) of gp120 is a potential hotspot for blocking HIV-1 infection. The present study aimed to design short peptides from well-characterized CD4bs targeting broadly neutralizing antibodies (bNAbs), which could be utilized as bNAb mimetics for viral neutralization. Co-crystallized structures of HIV-1 gp120 in complex with CD4bs-directed bNAbs were used to derive hexameric peptides using the Rosetta Peptiderive protocol. Based on empirical insights into co-crystallized structures, peptides derived from the heavy chain alone were considered. The peptides were docked with both HIV-1 subtype B and C gp120, and the stability of the peptide–antigen complexes was validated using extensive Molecular Dynamics (MD) simulations. Two peptides identified in the study demonstrated stable intermolecular interactions with SER365, GLY366, and GLY367 of the PHE43 cavity in the CD4 binding pocket, and with ASP368 of HIV-1 gp120, thereby mimicking the natural interaction between $ASP368_{gp120}$ and $ARG59_{CD4-RECEPTOR}$. Furthermore, the peptides featured favorable physico-chemical properties for virus neutralization suggesting that these peptides may be highly promising bNAb mimetic candidates that may be taken up for experimental validation.

KEYWORDS

HIV-1, peptide therapeutics, CD4-binding site, neutralizing peptides, molecular dynamics simulation



GRAPHICAL ABSTRACT
Graphical flow chart of methodology.

Introduction

Human Immuno-deficiency Virus (HIV), the causative agent of Acquired Immuno-Deficiency Syndrome (AIDS), continues to be a tenacious global public health challenge. According to the UNAIDS 2021 report, there were 37.7 million people living with HIV (PLHIV), of which 27.5 million people were on Anti-Retroviral Treatment (ART) and 1.5 million people were newly infected with HIV in 2020[1]. Though 40 years have passed since the discovery of HIV, a preventive vaccine against HIV continues to be a dream of the future (1, 2). However, the introduction of combinatorial Anti-Retroviral Therapy (cART)/Highly Active ART (HAART) has revolutionized the treatment of HIV infection and contributed significantly to viral suppression in infected individuals and control of transmission (3, 4). However, the emergence of drug resistance and the establishment of long-lived latent reservoirs remain major obstacles to the cure of HIV infection and elimination of the disease (5, 6).

In recent years, broadly neutralizing antibodies (bNAbs) that can neutralize diverse HIV-1 strains by targeting vulnerable epitopes on the HIV-1 envelope and thereby block HIV-1 infection have gained attention as potential adjuncts to antiretroviral therapy (7, 8). Recent studies have demonstrated that the administration of bNAbs is effective in suppressing viremia (9) and protecting against lentiviral infection in animal models (10, 11), thus providing valuable insights for the design of effective HIV-1 vaccines (12, 13). Very recently, researchers have directed their attention towards the development of therapeutic proteins and peptides targeting HIV, due to their advantages such as specificity and selective nature of action as compared to drugs and antibodies (14–16). Enfuvirtide (also known as Fuzeon or T20), an FDA-approved peptide-based drug, prevents the completion of HIV fusion events and has been used in combination with other anti-retroviral drugs for treating HIV infection (17). However, the drug has limited clinical application due to the emergence of resistant HIV-1 strains (12, 18).

Selective interaction of the HIV-1 envelope glycoprotein (gp120) with the CD4 molecule which serves as the primary cellular receptor, and one of the chemokine receptors CCR5/CXCR4 or both, constitutes a crucial step in HIV-1 infection (19–21). Regardless of the genomic and antigenic variation between HIV-1 strains, the CD4 binding site (CD4bs) is known to be well-conserved among the different HIV-1 subtypes and is reported to be one of the potential targets of neutralizing antibodies (22–24). The CD4bs is centered in a cavity formed at the interface of the gp120 outer and inner domains, where the hydrophobic residues present in the deep pocket constitute the point of contact with Phe-43 of the CD4 receptor (also called the Phe43 cavity) (25, 26). In addition, Arg59 of the CD4 receptor forms a salt bridge with D368 of gp120 to stabilize the CD4 binding site interaction (27, 28).

As early as 1999, Vita et al. reported that oligo-peptides targeting the CD4bs could inhibit the binding of gp120 with the CD4 receptor and thereby prevent HIV infection (29). The present study is based on the hypothesis that short peptides derived from the paratope of broadly neutralizing antibodies might function as potent mimics of these antibodies. This is based on earlier reports that ultra-short peptides of size up to seven amino acids have several useful features including biocompatibility, tunability, non-immunogenicity, biodegradability, and most importantly, efficient survival against proteolytic degradation in the gastrointestinal tract, as compared to longer peptides (30). We chose ultra-short peptides of 6-amino acids length (hexamers) for our study. Taking advantage of the available HIV-1 gp120-neutralizing antibody crystal structure complexes, we made an attempt to identify hexameric peptides from the paratope of neutralizing antibodies and characterized them using *in silico* methods like Molecular modeling, interacting interface analysis, and Molecular Dynamic (MD) simulation to understand their usefulness as therapeutic tools for HIV.

Materials and methods

Selection of co-crystal structures of broadly neutralizing antibody with HIV-1 envelope gp120

A number of CD4bs-directed neutralizing antibodies have been identified and reported. Based on their mode of

1 https://www.unaids.org/en

recognition and B-cell ontogeny, CD4bs antibodies fall into two categories: VH-gene restricted antibodies derived from the heavy chain germline genes VH1-2 or VH1-46, and CDRH3 dominated antibodies in which the antibody binding interfaces are dominated by the complementary-determining region three (CDR3) (13, 31, 32). The CD4bs directed bNAbs used for this study included VRC01 and 8ANC131, considered to be the first identified members of the VH-gene restricted "VRC01-class and 8ANC131-class" bNAbs (33) since the co-crystal structures of these antibodies with HIV-1B and C envelopes were available. VRC01 (VH1-2) and 8ANC131 (VH1-46) are both potent bNAbs found to be capable of neutralizing about 91 and 78% of the HIV-1 strains, respectively (34). The co-crystal structures of 8ANC131 with the HIV-1 subtype B envelope YU-2 gp120 (PDB ID: 4RWY 2.13 Å resolution) and VRC01 with the HIV-1 subtype C envelope ZM176.66 gp120 (PDB ID: 4LST 2.55 Å resolution) were downloaded from PDB (Protein Data Bank) (**Figure 1**). Both co-crystal structures included the heavy and light chains of the respective antibodies complexed with HIV-1 envelope gp120. The Fab (Fragment antigen-binding) regions of the antibodies were bound to the CD4bs in HIV-1 gp120.

Design of short linear peptides targeting the CD4-binding site

The Rosetta Peptiderive is a computational tool designed to predict possible inhibitory peptides from the crystal structures of protein complexes based on their interacting interface, was used to identify short linear peptides that would target the CD4bs and bring about virus neutralization. This tool is hosted online in ROSIE (Rosetta Online Server that Includes Everyone) web interface and can be accessed at https://rosie.rosettacommons.org/peptiderive. The antigen (HIV-1 gp120)–antibody (bNAb) complex was uploaded on the Rosetta peptiderive tool in PDB format, with optimal parameters defining the Receptor and Partner. The tool automatically refines the antigen–antibody complex by removing local clashes and extracts potential peptide fragments of specified window size. The binding energies of the identified peptide–antigen complexes were calculated using the Rosetta energy function (35). Peptides with the most significant binding scores were shortlisted, and their position, sequence, interface score and relative interface score were obtained (36). Intermolecular interactions of the identified peptide-antigen complexes were visualized in the PDBsum webserver (37) and CHIMERA (38).

Docking of peptides with human immunodeficiency virus-1 gp120

To validate the binding of the identified peptides with HIV-1 gp120, peptide–antigen docking was performed using HADDOCK (High Ambiguity Driven protein-protein DOCKing) webserver (Version 2.2) in the EASY interface available at https://wenmr.science.uu.nl/haddock2.4/. The antigen and peptides were docked by generating Ambiguous Interaction Restraints (AIR) with the interface residues identified from the PDBsum analysis of the Rosetta peptiderive complexes (39, 40). The docked structures were summarized in clusters, and each cluster was assigned a HADDOCK score, cluster size, RMSD from the overall lowest energy conformations, Z-score and buried surface area along with bonding energies (Vander Waal's, electrostatic, desolvation, and restraints violation energies). The best-docked complex (topmost cluster suggested by HADDOCK) replicating the desired residual interactions was identified and selected for further analysis. The binding affinity (ΔG) and dissociation constant (K_d) of the docked complexes were calculated using the PRODIGY webserver, available at https://wenmr.science.uu.nl/prodigy/. This webserver predicts binding affinities based on inter-molecular contacts within a distance cut-off of 5.5 Å (41, 42).

Molecular dynamics simulations of the peptides with human immunodeficiency virus-1 envelopes

The peptide-HIV-1 envelope complexes identified using Rosetta peptiderive were subjected to Molecular dynamics (MD) simulations to deduce their dynamic behavior under physiologically simulated conditions (43). MD simulations were performed using the DESMOND software package (44) with OPLS_2005 as a force field and implemented as in Muthukumaran et al. (45). To begin with, the system was built in an auto-calculated cubic box and solvated with explicit Single Point Charge (SPC) water molecules. The solvated system was energy minimized and the MD run was carried out for 200 ns by implementing an NPT ensemble with a sampling interval of 10 ps. During the MD run, the whole system was maintained at an equilibrium of 300 K temperature and 1 atm pressure. Analytical tools available in DESMOND were used to infer the Root Mean Square Deviation (RMSD) of the protein backbone, the Root Mean Square Fluctuation (RMSF) of the residues, the radius of gyration, and other structural transitions throughout the simulations.

Molecular mechanics-poisson boltzmann surface area calculation for the top-scoring stable neutralizing peptide-antigen complexes

The binding free energy (ΔG) of the final frames of stable neutralizing peptide–antigen complexes obtained from the MD simulation was calculated by implementing MM-PBSA (Molecular Mechanics-Poisson Boltzmann Surface Area) protocol in farPPI (fast amber rescoring for Protein–Protein interaction Inhibitors) webserver, available at http://cadd.zju.edu.cn/farppi/. Precise binding energies of the docked poses were evaluated by the MM-PBSA method which combines energy calculations based on implicit solvent

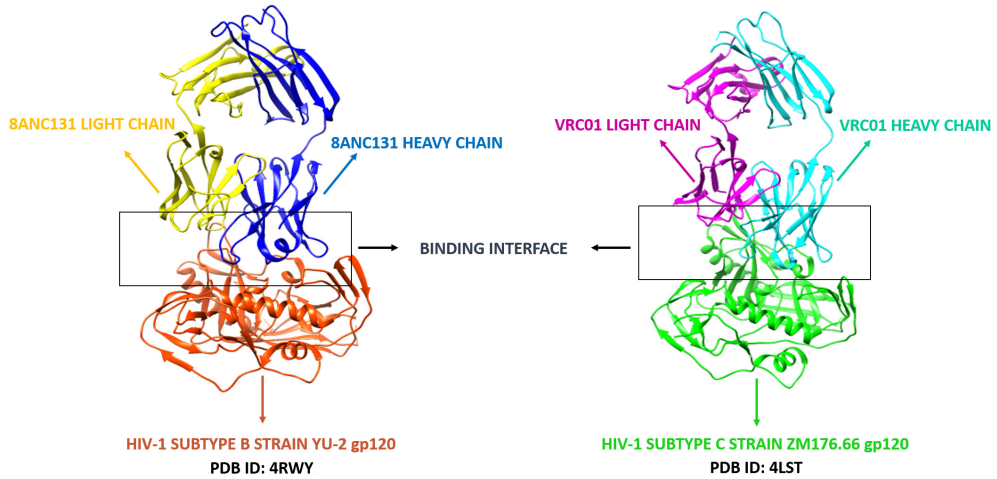


FIGURE 1
Co-crystal structures of 8ANC131-YU-2 gp120 and VRC01-ZM176.66 gp120.The secondary structure elements (alpha helix and beta sheets) color coded. PDB ID: 4RWY—Orange red: HIV-1 subtype B envelope, Blue: 8ANC131 Heavy chain, Yellow: 8ANC131 Light chain. PDB ID: 4LST—Green: HIV-1 subtype C envelope, Cyan: VRC01 Heavy chain, Magenta: VRC01 Light chain. The interface between the antibodies and HIV-1 gp120 are highlighted and shown as the binding interface/CD4-binding site. (These two antibodies were selected based on their neutralization profile and availability of crystal structure with HIV-1 subtype B and C envelope gp120 in Protein Data Bank).

and molecular mechanics model (46). Among the MM-PBSA procedures, PB3 based approach was found to be highly accurate as compared to the other approaches in farPPI, as two force fields, GAFF2 and ff14SB, were applied to the peptide and antigen, respectively (47, 48). Hence, this method was adopted to score the binding free energy of the peptide–antigen complexes.

KDeep absolute binding affinity calculation for the most stable neutralizing peptide-antigen complexes

In addition to MM-PBSA, absolute binding affinity (ΔG) of the topmost neutralizing peptide–antigen complexes was calculated using KDeep, a protein–ligand affinity predictor tool available at https://playmolecule.com/Kdeep/. This predictor works based on a machine learning approach using a state-of-the-art 3D convolutional neural network (49). The input was voxelized into pharmacophore features like aromaticity, hydrophobicity, total excluded volume, etc., and passed onto the DCNN (Deep Convolutional Neural Network) model, which is pre-trained by the PDBbind benchmark (v.2006). Based on the implemented algorithm, the binding affinity of the identified neutralizing peptide–antigen complexes was calculated as discussed by Karlov et al. (50) and Varela-rial et al. (51).

Additional computational predictions

The identified peptides were subjected to alanine scanning using Bude Alanine Scan[2] (52, 53) and Robetta Alanine scan[3] (54) webservers to infer the energetically significant amino acids at the peptide–antigen interface. This prediction helps to prioritize key residues in the identified peptides. Toxicity and physico-chemical properties of the peptides were predicted using ToxinPred[4] (55) and the peptide analyzing tool provided by Thermo-fisher Scientific[5].

Results

Neutralizing peptides derived from the CD4-binding site-directed neutralizing antibodies

Four hexameric peptides were derived through structure-based sequence inference from the 8ANC131 and VRC01 neutralizing antibody-HIV-1 gp120 complexes using the Rosetta peptiderive protocol as shown in **Figures 2**, **3**. From the hot segments in the bNAbs (that contribute to the most significant binding interaction with the HIV-1 envelope gp120 protein), two peptides were identified from each of the two antigen–antibody complexes. These included the peptide Arg-Asp-Arg-Ser-Thr-Gly (RDRSTG) from the H chain of 8ANC131, which had an interface score of −9.447 and contributed to 29% of binding energy, and the peptide

2 https://pragmaticproteindesign.bio.ed.ac.uk/balas/
3 https://robetta.bakerlab.org/queue.jsp
4 http://crdd.osdd.net/raghava/toxinpred/
5 https://www.thermofisher.com/in/en/home/life-science/protein-biology/peptides-proteins/custom-peptide-synthesis-services/peptide-analyzing-tool.html

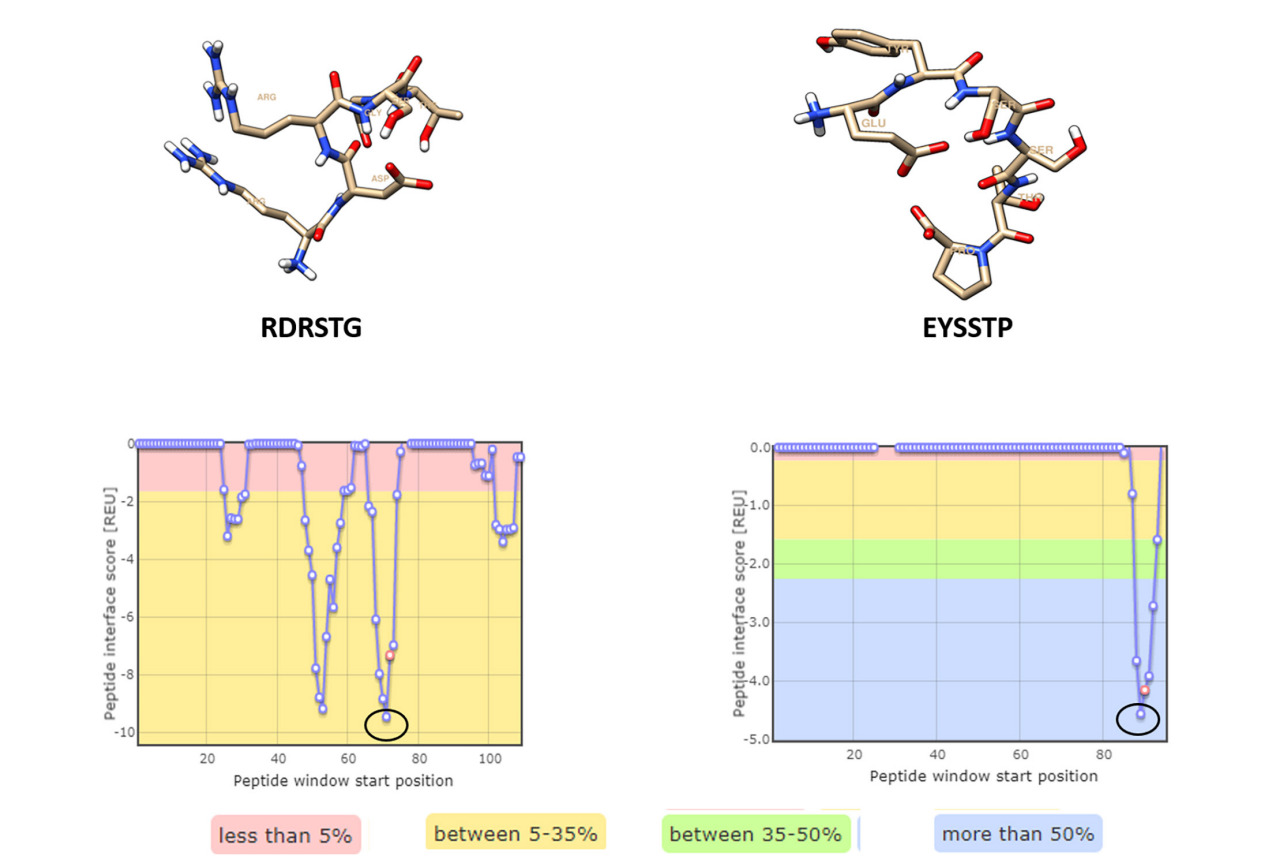


FIGURE 2
Peptides derived from 8ANC131 with the predicted hotspot regions on the antigen-antibody interface. The regions from where the peptides are derived are highlighted. Peptides were visualized in Chimera, version 1.16.

Glu-Tyr-Ser-Ser-Thr-Pro (EYSSTP) from the L chain, which had an interface score of −4.554 and contributed to 101% of binding energy. Two other hexamers were derived from the PDB crystal structure of VRC01-HIV-1C envelope, namely, Val-Asn-Tyr-Ala-Arg-Pro (VNYARP) from the H chain, which had an interface score of −9.982 and contributed to 30% of binding energy, and QQYEFF (Gln-Gln-Tyr-Glu-Phe-Phe) from the L chain, which had an interface score of −6.839 and contributed to 68% of binding energy (**Table 1**). In general, the peptides derived from the heavy chain of the antibodies gave comparatively lower interface scores than peptides derived from the light chain, signifying better binding affinity of the former. Among the four peptides, RDRSTG peptide having an interface score of −9.447 showed the most significant binding to the HIV-1 envelope.

Molecular docking of peptides with antigens

Structural analysis of the 8ANC131-subtype B gp120 (PDB ID: 4RWY) and VRC01-subtype C gp120 (PDB ID: 4LST) complexes revealed close interaction between the antibody Heavy chains and the HIV-1 gp120 CD4-binding site, while the light chains protruded beyond the CD4bs, particularly the D Loop and V5 regions (**Figure 4**). Therefore, we excluded the peptides derived from the light chains as they did not engage our target, i.e., the CD4bs. Residues 365–371 of HIV-1 gp120 were found to be the key residues involved in making critical contacts with Phe43 and Arg59 residues of the CD4 receptor (56). The VRC01 antibody showed a non-bonded interaction with $Ser365_{gp120}$, $Gly366_{gp120}$, and $Gly367_{gp120}$ of the Phe43 cavity, while in 8ANC131, $Gly366_{gp120}$ and $Gly367_{gp120}$ were found to be involved in the interaction (57). Furthermore, $ASP368_{gp120}$ was observed to mediate the interaction with $ARG71_{8ANC131/VRC01}$ by forming hydrogen bonds and salt bridges, which mimicked the natural interaction between $ARG59_{CD4RECEPTOR}$ and $ASP368_{gp120}$ (25, 31, 34) (**Figure 5**).

We also performed intermolecular interaction analysis of the Rosetta-derived peptide-antigen complexes and observed similar interactions as seen in the PDB crystal structures (**Supplementary Figure 1**). Among the peptides derived from the antibody heavy chains, RDRSTG was found to form two hydrogen bonds with ASP368 (2.75 Å and 2.82 Å) and MET426 (2.72 Å and 3.20 Å), and one hydrogen bond with GLY431

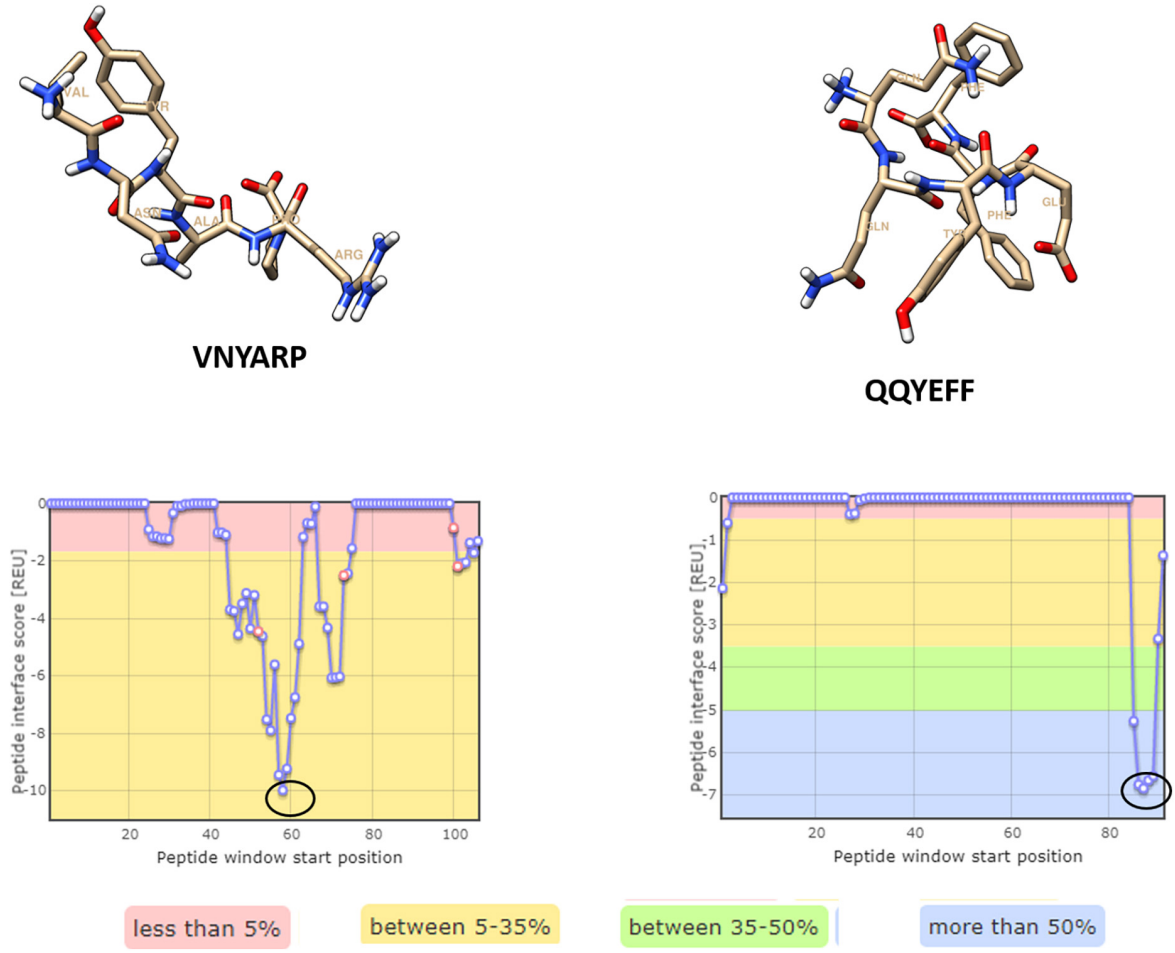


FIGURE 3
Peptides derived from VRC01 with the predicted hotspot regions on the antigen–antibody interface. The regions from where the peptides are derived are highlighted in a circle [Peptides were visualized in Chimera, version 1.16].

TABLE 1 Peptides derived from neutralizing antibodies and their interface scores.

PDB ID (Co-crystal structure)	Peptide sequence	Receptor (Envelope gp120)	Antibody chain (H-Heavy/ L-Light)	Position in neutralizing antibody (Crystal structure)	Interface score	Total interface score (REU)	Relative interface score (%)
4RWY (8ANC131-subtype B gp120)	RDRSTG	A	H	71–76	−9.447	−33.11	28.54
	EYSSTP	A	L	90–95	−4.554	−4.49	101.32
4LST (VRC01-subtype C gp120)	VNYARP	G	H	57–62	−9.982	−33.57	29.73
	QQYEFF	G	L	89–91, 96–98	−6.839	−10.03	68.23

*Highlighted peptides contribute significantly to binding with the respective antigen.

(2.76 Å); the other crystal structure residues were found to have non-bonded contacts in the vicinity of <5 Å (LEU122, VAL430, TRP427 and LYS432). The VNYARP peptide formed two hydrogen bonds with GLY458 (3.08 Å and 3.12 Å) and one hydrogen bond with ARG456 (2.73 Å) and THR467 (3.00 Å). In addition, salt bridges were also observed at ASP461

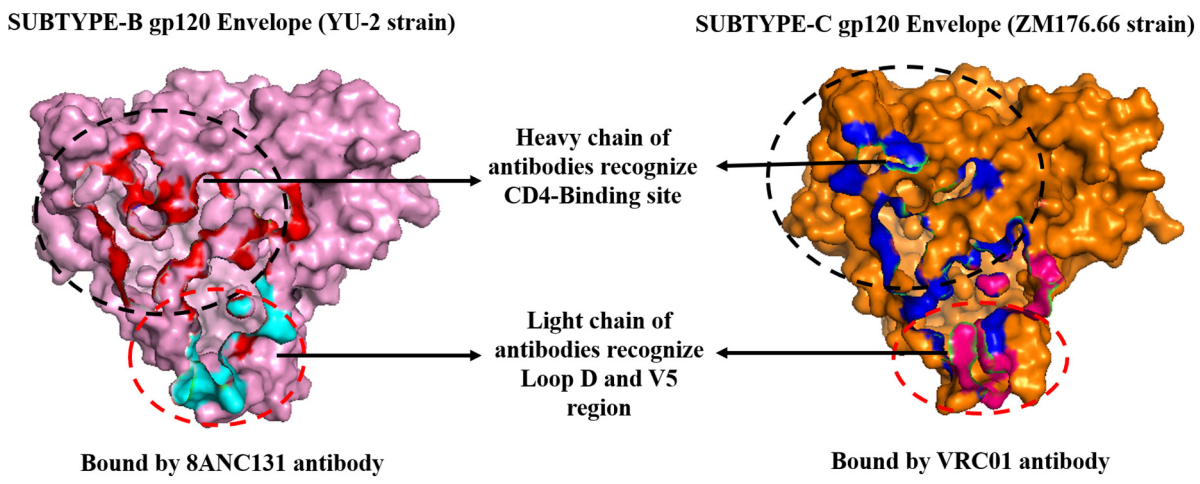


FIGURE 4
Binding pockets of 8ANC131 and VRC01 in HIV-1 subtype B and subtype C gp120.

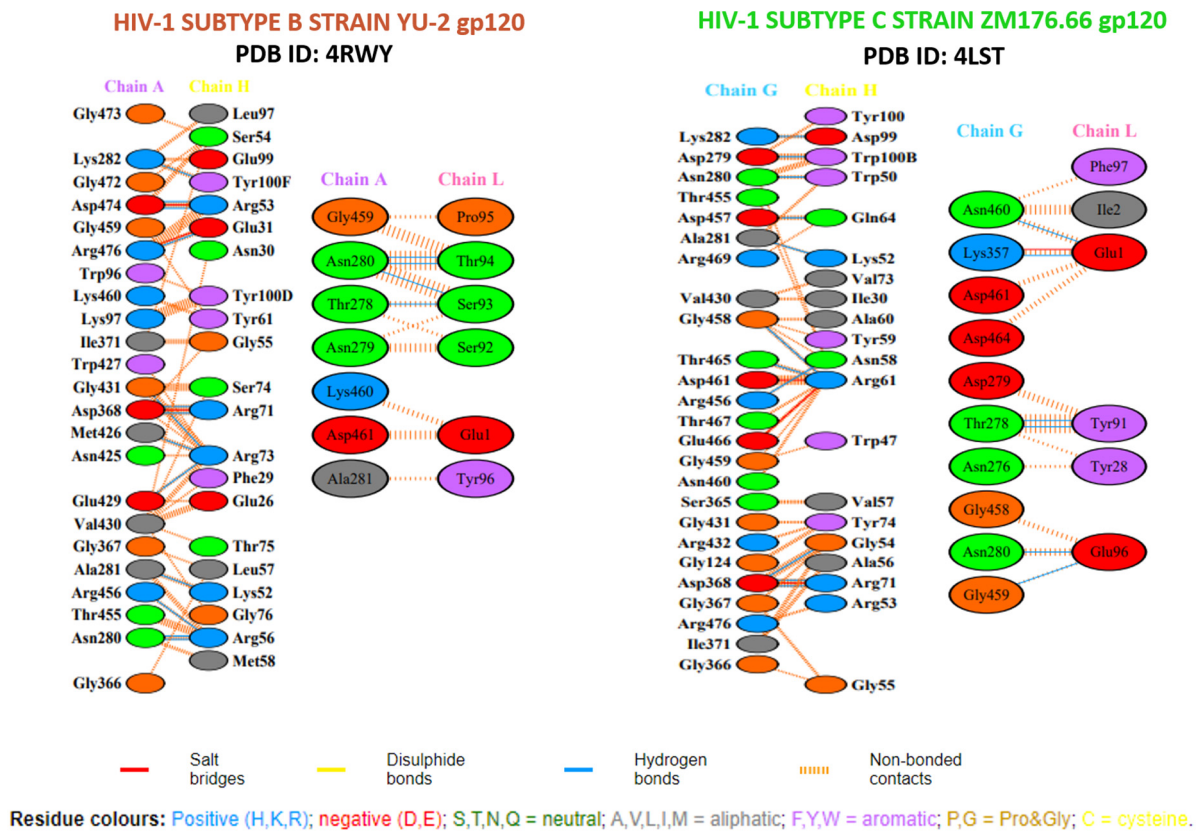


FIGURE 5
Hydrogen bonding (HB) interactions plot of the crystal structure of antigen–neutralizing antibody complexes.

and GLU466. Other non-bonded contacting residues were ASN280, THR465, GLY366, SER365, ASN460 and ASP457. Based on these observations, we docked the neutralizing antibody 8ANC131-derived peptide RDRSTG with subtype C (ZM176.66) gp120, and the VRC01-derived peptides VNYARP with subtype B (YU-2) gp120, to examine the closeness of the interaction patterns (especially ASP368 and SER365) with that seen in the native crystal structures. To revalidate the

observed interactions in the Rosetta derived complexes, we performed re-docking of the peptide RDRSTG with subtype B (YU-2) gp120 and VNYARP with subtype C (ZM176.66) gp120. (Positions of residues are different in PDB crystal structures and 2D LIGPLOT—Supplementary Table 1; Residues stated here are in accordance with the crystal structure but different from that in LIGPLOT).

Docking with subtype B gp120

The RDRSTG peptide derived from 8ANC131 was found to form hydrogen bonds with ASP368 (2.96 Å, 2.62 Å), TRP427 (2.81 Å, 2.97 Å), GLY198, GLU370, ASN425, MET426, GLU429 and LYS432, with a binding affinity (ΔG) of −9.1 kCal/mol and K_d of 2.2E-07. The peptide VNYARP derived from VRC01 showed hydrogen bond interactions with ASP368 (2.68 Å) and GLY431 (2.86 Å) with a binding affinity (ΔG) of −8.6 kCal/mol and K_d of 5.0E-07 (Supplementary Figure 2).

Docking with subtype C gp120

The RDRSTG peptide featured interactions at positions SER365 (2.75 Å and 2.69 Å), GLY366, ASP457 (2.78 Å and 2.67 Å), GLY458 and ASN460, with a binding affinity (ΔG) of −8.8 kCal/mol and K_d of 3.4E-07. In the case of VNYARP-subtype C gp120 re-docking, hydrogen bond interactions were observed at ASN280 (2.87 Å and 3.10 Å), LYS360, HIS364, ASP457, ASP461 (2.65 Å, 3.23 Å), THR465, GLU466, THR467 and ARG469, thus concurring with the Rosetta peptiderive prediction. However, the main residue SER365 was noticed to form a non-bonded contact with a binding affinity (ΔG) of −9.9 kCal/mol and K_d of 5.9E-08 (Supplementary Figure 3). The redocking study demonstrated the predictive accuracy of the methods implemented.

Molecular dynamics simulation analysis of the peptide-antigen complexes

To start with, the HIV-1 subtype B and subtype C gp120 antigens (without peptides) were subjected to a production run of 200 ns, and trajectory analysis was performed. The system of subtype B gp120 antigen comprised of 49,311 atoms with 14,688 water molecules in the neutralized state, while the subtype C gp120 antigen system comprised of 48,478 atoms with 14,405 water molecules in the neutralized state. The RMSD plot of both antigens revealed that the Cα deviations were stable and within the range of 3 Å, and were found to converge toward the final stages of simulation (Supplementary Figure 4). The RMSF plot identified the peaks which represent the regions/residues that fluctuated the most during the simulation: 175–200 (4.7 Å) and 225–250 (5.0 Å) regions in subtype B gp120, and 250–275 (4.0 Å) and 300–337 (4.2 Å) regions in subtype C gp120 (Supplementary Figure 5).

We then performed molecular dynamics simulation of the peptide–gp120 complexes. The simulation system of subtype B gp120-RDRSTG solvated complex comprised of 49,305 atoms with 14,654 water molecules, and was neutralized by adding one cl^- ion (1.241 mM). On trajectory analysis, the protein–ligand RMSD plot revealed that the complex converged at 10 ns with a 0.6 Å difference between the peptide and antigen-bound state (Figure 6). The ligand RMSD value was in the range of 3.0 Å with reference to the backbone of the antigen and was found to be well-bound to the binding regions. The RMSF plot revealed that RDRSTG (Supplementary Figure 6) interacted well at regions 50–100, 220–250, 250–300 and 300–337, despite fluctuations. Fluctuations posed by the peptide throughout the simulation were inferred from the ligand RMSF plot (Supplementary Figure 7), where it was found to be stable in the range of 4 Å. The structural compactness of the peptide was measured based on the radius of gyration (rGyr). This analysis revealed that the peptide RDRSTG maintained its compactness up to 150 ns in the range of 1 Å (Supplementary Figure 8). The bonded interactions between the antigenic residues and the RDRSTG peptide were analyzed from the ligand–protein contacts plot (Figure 7), wherein it was found that for about 79 and 52% of the duration of the run, Asp229 (ASP368) and Glu274 (GLU429) interacted by means of hydrogen bonds, ionic bonds and water bridges, respectively (Supplementary Figure 10).

Subtype C gp120-VNYARP peptide was made up of 48,414 atoms with 14,350 water molecules in the neutralized state. The antigen–peptide RMSD plot inferred that the complex converged at 75 ns with a 0.6 Å difference between the peptide and antigen-bound states (Figure 6). However, the peptide VNYARP evolved to make stable interactions between 85 and 160 ns in the vicinity of <3 Å. The ligand RMSD value was in the range of 2.0 Å with a major fluctuation at 75 ns. Residues in the region 150–180, 200–250, and 300–339 were found to sustain bonded interactions with the peptide as per the RMSF plot (Supplementary Figure 6). The ligand RMSF plot inferred that the peptide is stable as the fluctuations were within the range of 4 Å (Supplementary Figure 7). The rGyr analysis revealed a minimum deviation of 6.0–6.5 Å, indicating that the peptide sustained high compactness during the entire simulation process (Supplementary Figure 9). With regard to peptide–antigen contacts (Figure 8), Gly305 (GLY458) was found to interact 96% of the time during the entire run by means of hydrogen bonds and water bridges. Asp304 (ASP457—70%), Asp227 (ASP368—67%), Gly226 (GLY367—62%) and Ser224 (SER365—62%) formed hydrogen bond interactions and water bridges, with the exception of Asp227 (ASP368), where an additional ionic interaction featured. The least interacting residue was Arg303 (ARG456), which revealed sustained binding (hydrogen bonds and water bridges) around 53% of the 200 ns production run (Supplementary Figure 11).

The MD trajectories revealed RDRDTG and VNYARP peptides to be highly stable in terms of bonded interactions during the 200 ns of simulation. The dynamic evolution

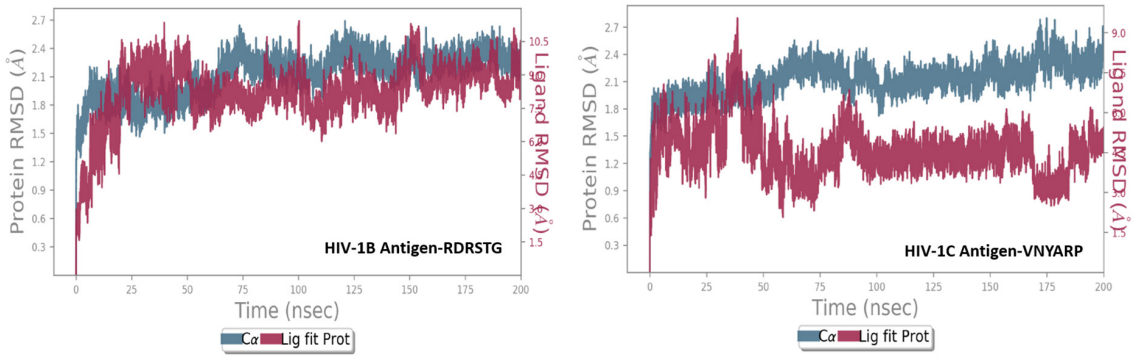


FIGURE 6
Root mean square deviation (RMSD) plot of the peptide–antigen complexes.

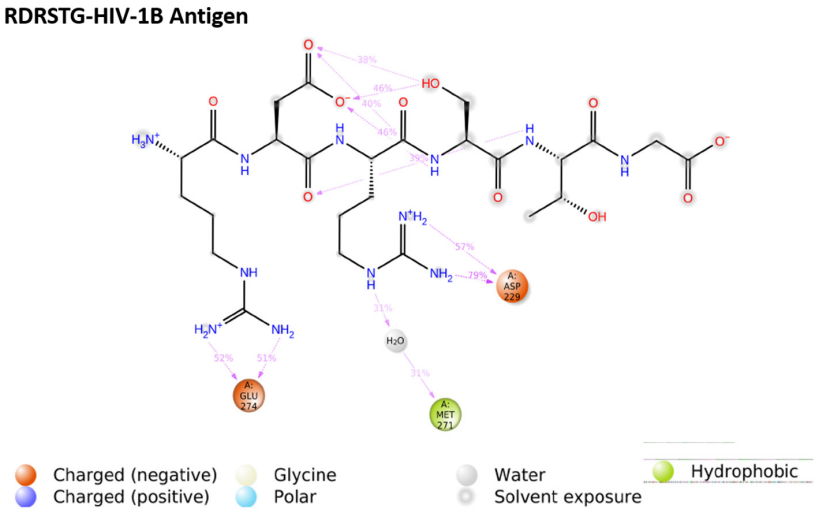


FIGURE 7
Ligand–protein contacts of peptide Arg-Asp-Arg-Ser-Thr-Gly (RDRSTG) within the subtype B gp120 complex.

of the peptides RDRSTG and VNYARP are illustrated in Figures 9, 10. The MD trajectory analyses revealed that the peptides RDRSTG and VNYARP were stable binders, as they feature stable contacts with the key residues namely, SER365, GLY366, GLY367, and ASP368 across the production run (Supplementary Figure 12). The binding free energies (ΔG) of the complexes (RDRSTG-subtype B gp120 and VNYARP-subtype C gp120) were calculated over the MD simulation trajectory for the frames sampled at an interval of 20 ns and subjected to MM-PBSA (PB3) using the far-ppi server and binding affinity calculation using Kdeep, respectively. MM-PBSA calculations of RDRSTG-subtype B gp120 and VNYARP-subtype C gp120 complexes gave an average of −13.58 ± 2.85 (Mean ± SD) kCal/mol and −16.04 ± 8.77 (Mean ± SD) kCal/mol, respectively. Similarly, KDeep calculations gave an average of −9.32 ± 0.80 (Mean ± SD) kCal/mol for RDRSTG-subtype B gp120 and −10.18 ± 0.63 (Mean ± SD) kCal/mol for VNYARP-subtype C gp120, respectively (Supplementary Figure 17).

The Alanine scan analysis for RDRSTG-Subtype B gp120 and VNYARP-Subtype C gp120 complexes using Robetta and Bude scan identified the cumulative energetically important amino acids in the peptides across the binding interface as R, D, R, S and T in the RDRSTG peptide and V, N, Y and R in the VNYARP peptide. The binding affinities of the alanine mutated peptides are provided in Supplementary Table 2. The results of the physico-chemical analysis are provided in Table 2. Further, the peptides were found to be non-toxic.

Discussion

The CD4-binding site of the HIV-1 envelope has been a key target of therapeutics for many years. However, not

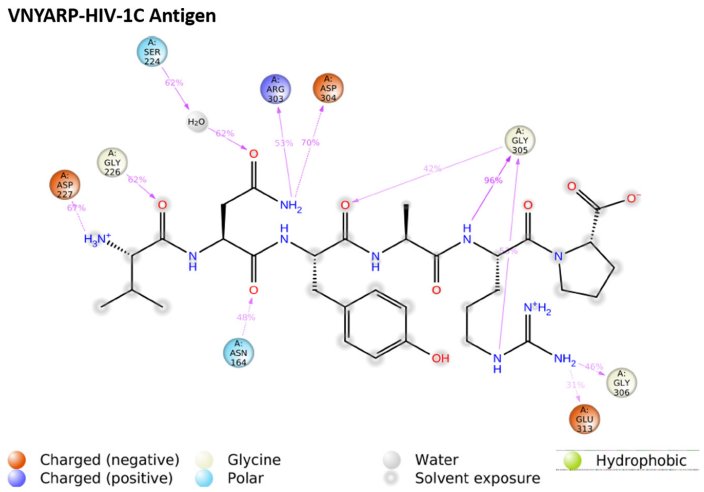


FIGURE 8
Ligand–protein contacts of peptide Val-Asn-Tyr-Ala-Arg-Pro (VNYARP) within the subtype B gp120 complex.

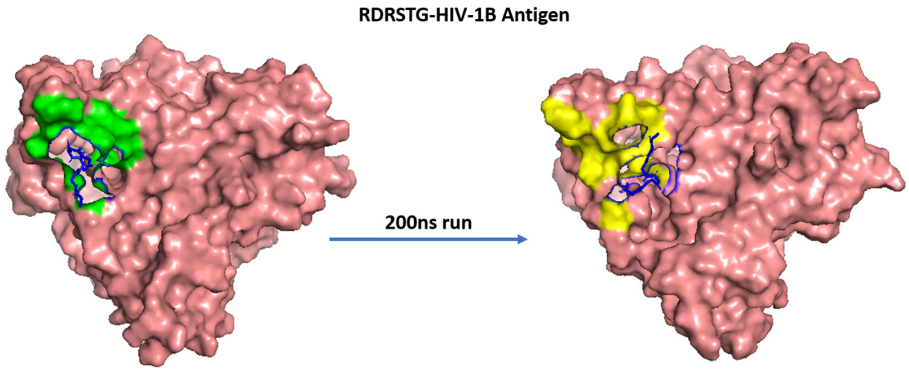


FIGURE 9
Dynamic evolution of the RDRSTG-HIV-1B gp120 complex. Salmon—HIV-1B YU-2 gp120 envelope; Blue—RDRSTG peptide; shaded regions indicate the interactions before and after simulation.

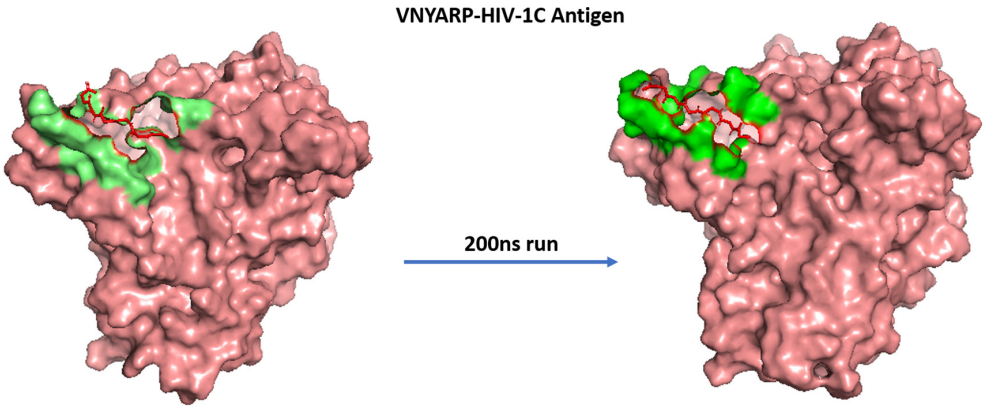


FIGURE 10
Dynamic evolution of the VNYARP-HIV-1C gp120 complex. Salmon—HIV-1C ZM176.66 gp120 envelope; Red—peptide VNYARP; shaded regions indicate their interactions before and after simulation.

TABLE 2 Predicted physico-chemical properties of the neutralizing antibody heavy chain derived peptides.

Peptide	Alanine scan		ToxinPred	Hydrophobicity	Charge	GRAVY	MW Avg. g/mol	MW Mono-isotopic	Theoretical PI
	Bude	Robetta							
RDRSTG	RDRSTG	RDRSTG	Non-toxic	2.66	+1	−2.40	690.7193	690.3409	10.9
VNYARP	VNYARP	VNYARP	Non-toxic	6.63	+1	−0.82	718.8183	718.3763	9.9

*Highlighted letters indicate hot spot residues in the alanine scan. MW, Molecular weight.

a single drug targeting the CD4bs has been approved by the US FDA to date (58). With the discovery of broadly neutralizing antibodies, various new approaches have been explored to improve treatment strategies for HIV infection. Despite advancements witnessed in the treatment of HIV, the development of immune therapeutics still remains a cumbersome, time-consuming, and highly expensive process. In recent decades, peptide therapeutics have gained significance in the field of medicine, for being highly specific and efficacious, with good tolerability and safety profiles (59). The interest in peptide therapeutics has been mitigated by certain limitations; these include the relatively short half-life, physiological instability, and difficulty in oral administration (60). However, there have been ongoing efforts to eliminate the obstacles in utilizing peptides, through half-life extension and stability enhancement under physiological conditions (61). Numerous studies have demonstrated the usefulness of short inhibitory peptides in the treatment of several diseases, particularly cancer (62–65). More recently, peptide therapeutics have also shown promise for the treatment of HIV infection (12, 16).

A number of studies in the past have attempted to identify potent peptide inhibitors targeting the CD4bs (29, 66–70), but without much success. This is because a successful inhibitor should not only block the binding of the HIV envelope to the CD4 receptor but should also efficiently block co-receptor interaction which is important for HIV-1 entry into the target cell (71). This kind of inhibition is actually accomplished very well by neutralizing antibodies, which target specific epitopes on the virus and lead to virus neutralization, thereby preventing HIV infection. Modern methods in computer-aided drug design have catalyzed the ability to reduce cost and time which limits the development of novel therapeutics (72).

Andrianov et al. (73) utilized a computer-aided strategy to screen a public web-oriented virtual screening platform (pepMMsMIMIC) to identify a few promising peptidomimetic candidates from the broadly neutralizing antibody VRC01 (73). In a similar line, we undertook an in-depth analysis of the co-crystal structures of the bNAb 8ANC131-subtype B YU-2 gp120 and VRC01-subtype C ZM176.66 gp120 complexes and inferred that the contacts made by each CD4bs-directed broadly neutralizing antibody with the HIV-1 gp120 were highly variable. However, it was observed that the heavy chain of the CD4bs-directed neutralizing antibodies engaged well with the CD4bs, i.e., the Phe43 cavity, which is highly conserved among the different bNAbs. Based on earlier studies as well as our analysis of the co-crystal structures of the antibody-antigen complexes, we decided to narrow down on hexameric peptides that would be short and at the same time target the critical residues in the CD4bs. Subsequently, potential hexamers were derived from the crystal structures of 8ANC131-subtype B gp120 and VRC01-subtype C gp120. Two peptides were predicted from each crystal structure, one from the heavy chain and another from the light chain. Only the peptides derived from heavy chains were taken up for further computational evaluations as they bound best to the CD4bs. The heavy chain derived peptides were docked with subtype B and subtype C envelopes, to identify interactions with the key residues in the CD4bs. Based on the 2D-interaction plot of the crystallized complexes, peptides RDRSTG and VNYARP, derived from the heavy chain of 8ANC131 and VRC01, respectively, were shortlisted as they interacted with the key residues of the CD4bs mentioned earlier. Molecular dynamics simulation of the RDRSTG-subtype B gp120 and VNYARP-subtype C gp120 complexes across the 200 ns trajectory (frames sampled at an interval of 20 ns) revealed that the peptides RDRSTG and VNYARP precisely target the binding site of the CD4 receptor (Phe43 and Arg59 contacts) and interact with the critical residues through hydrogen bonds and Vander Waal's interactions with an average binding free energy (ΔG) (MM-PBSA) of −13.58 ± 2.85 (Mean ± SD) kCal/mol and −16.04 ± 8.77 (Mean ± SD) kCal/mol, respectively. The sampled frames were also subjected to KDeep calculation, wherein, the peptides RDRSTG and VNYARP scored a significant average binding affinity (ΔG) of −9.32 ± 0.80 (Mean ± SD) kCal/mol and −10.18 ± 0.63 (Mean ± SD) kCal/mol, respectively. In the case of VNYARP, one of the frames at the 60th ns gave a higher MMPBSA value (ΔG = +3.18 kCal/mol) due to a major conformation change; however, the lower binding free energy state was quickly regained around the 80th ns.

The energetically significant amino acids in the topmost stable peptide-antigen complexes of RDRSTG-subtype B gp120 and VNYARP-subtype C gp120 were found to be R, D, R, S, T, V, N, Y and R, as inferred from the cumulative results of the alanine scan (52, 53) and Robetta analyses

(54, 74). The two peptides were also predicted to possess favorable physicochemical properties including non-toxicity, hydrophobicity of 2.66 and 6.63, and GRAVY (Grand Average of Hydropathy) of −2.40 and −0.82 (75, 76), respectively (**Table 2**) making these highly promising therapeutic candidates. A striking finding to be noted is that the RDRSTG peptide is derived from the site that is involved in the critical interaction between $ARG59_{CD4-RECEPTOR}$ and $ASP368_{gp120}$. This could be the likely reason for this peptide standing out as the best CD4bs-targeting neutralizing peptide, as compared to all other peptides.

We further analyzed the co-crystal structures of other VH-gene-restricted (VRC01-class and 8ANC131-class) and CDR-H3-dominated antibodies with gp120 envelope for their residual interactions. The VRC01-class antibodies 3BNC117 (PDB ID: 4JPV), N6 (PDB ID: 5TE7) and NIH45-46 Fab (PDB ID: 4JDV) revealed interactions between the conserved $ARG71_{HC}/_{HeavyChain}$ residue and $ASP368_{gp120}$. In addition, these antibodies also interacted with $SER365_{gp120}$, $GLY366_{gp120}$ and $ASP368_{gp120}$ through $Leu44_{CD4}$ and $Lys46_{CD4}$ (10, 57, 77). In case of 8ANC131-class antibodies (1B2530; PDB ID: 4YFL) and CDR-H3 dominated antibody (CH103; PDB ID: 4JAN), the key contacts were $ASP368_{gp120}$ through $ARG72_{HC}$ and $ARG97_{HC}$, respectively. These antibodies also showed interaction with residues of the PHE43 cavity in gp120 (34, 78). Given these observations, we speculate that peptides derived from these neutralizing antibodies could also be explored for the identification of novel neutralizing peptide mimetics against HIV.

The binding of HIV-1 gp120 with the CD4 receptor on the target cell triggers a conformational change that uncovers epitopes called CD4-induced (CD4i) epitopes that bind to the chemokine co-receptors on the host cell, either CCR5 or CXCR4. Since the binding of the candidate bNAb mimetics to the CD4bs prevents conformational changes in the HIV-1 gp120 and obsoletes binding to the co-receptor, the process of viral entry into the target cells is also inhibited. Thus, the peptide mimetics identified in this study hold promise as highly potent candidates for HIV therapeutics.

Conclusion

Using modern computational tools the present study identified two short, hexameric peptides from the heavy chain of two well-characterized CD4bs-targeting bNAbs, 8ANC131 and VRC01, that hold promise as potential therapeutic candidates that can be exploited for the treatment of HIV-infected persons. This study is the first of its kind to identify short peptides that can bind to and possibly neutralize HIV-1. Given the potential of the identified candidate peptides to function as mimetics of HIV-1 broadly neutralizing antibodies, *in vitro* studies are in progress to validate their efficacy in HIV-1 neutralization in our laboratory (20).

Data availability statement

The original analyses presented in this study are included in the article/**Supplementary material**. Further inquiries can be directed to the corresponding authors.

Author contributions

UV and LH: conceptualization, resources, writing—review and editing, and supervision. SV: methodology, formal analysis, investigation, data curration, and writing—original draft preparation. SV, UV, and LH: validation. LH: project administration. All the authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors thank the funding agency, Regional Centre for Biotechnology (RCB)/Department of Biotechnology (DBT) for awarding her fellowship (DBT/2020/NIRT/1321).

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.1036874/full#supplementary-material

References

1. Euler Z, Schuitemaker H. Cross-reactive broadly neutralizing antibodies: timing is everything. *Front Immun.* (2012) 3:215. doi: 10.3389/fimmu.2012.00215

2. Bi W, Xu W, Cheng L, Xue J, Wang Q, Yu F, et al. IgG Fc-binding motif-conjugated HIV-1 fusion inhibitor exhibits improved potency and in vivo half-life: potential application in combination with broad neutralizing antibodies. *PLoS Pathog.* (2019) 15:e1008082. doi: 10.1371/journal.ppat.1008082

3. Atta MG, De Seigneux S, Lucas GM. Clinical pharmacology in HIV therapy. *CJASN.* (2019) 14:435–44. doi: 10.2215/CJN.02240218

4. Ghosn J, Taiwo B, Seedat S, Autran B, Katlama C. HIV. *Lancet.* (2018) 392:685–97. doi: 10.1016/S0140-6736(18)31311-4

5. Simon V, Ho DD, Abdool Karim Q. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet.* (2006) 368:489–504. doi: 10.1016/S0140-6736(06)69157-5

6. Abner E, Jordan A. HIV "shock and kill" therapy: in need of revision. *Antiviral Res.* (2019) 166:19–34. doi: 10.1016/j.antiviral.2019.03.008

7. Haynes BF, Burton DR, Mascola JR. Multiple roles for HIV broadly neutralizing antibodies. *Sci Transl Med.* (2019) 11:eaaz2686. doi: 10.1126/scitranslmed.aaz2686

8. Spencer DA, Shapiro MB, Haigwood NL, Hessell AJ. Advancing HIV broadly neutralizing antibodies: from discovery to the clinic. *Front Public Health.* (2021) 9:690017. doi: 10.3389/fpubh.2021.690017

9. Mishra N, Sharma S, Dobhal A, Kumar S, Chawla H, Singh R, et al. Broadly neutralizing plasma antibodies effective against autologous circulating viruses in infants with multivariant HIV-1 infection. *Nat Commun.* (2020) 11:4409. doi: 10.1038/s41467-020-18225-x

10. Huang J, Kang BH, Ishida E, Zhou T, Griesman T, Sheng Z, et al. Identification of a CD4-binding-site antibody to HIV that evolved near-pan neutralization breadth. *Immunity.* (2016) 45:1108–21. doi: 10.1016/j.immuni.2016.10.027

11. Caskey M, Klein F, Nussenzweig MC. Broadly neutralizing anti-HIV-1 monoclonal antibodies in the clinic. *Nat Med.* (2019) 25:547–53. doi: 10.1038/s41591-019-0412-8

12. Chupradit K, Moonmuang S, Nangola S, Kitidee K, Yasamut U, Mougel M, et al. Current peptide and protein candidates challenging HIV therapy beyond the vaccine era. *Viruses.* (2017) 9:281. doi: 10.3390/v9100281

13. Ding C, Patel D, Ma Y, Mann JFS, Wu J, Gao Y. Employing broadly neutralizing antibodies as a human immunodeficiency virus prophylactic & therapeutic application. *Front Immunol.* (2021) 12:697683. doi: 10.3389/fimmu.2021.697683

14. Vetrivel U, Nagarajan H, Thirumudi I. Design of inhibitory peptide targeting *Toxoplasma gondii* RON4-human β−tubulin interactions by implementing structural bioinformatics methods. *J Cell Biochem.* (2018) 119:3236–46. doi: 10.1002/jcb.26480

15. Usmani SS, Kumar R, Bhalla S, Kumar V, Raghava GPS. In silico tools and databases for designing peptide-based vaccine and drugs. In: Donev R editor. *Advances in Protein Chemistry and Structural Biology*. Amsterdam: Elsevier (2018). p. 221–63. doi: 10.1016/bs.apcsb.2018.01.006

16. Pu J, Wang Q, Xu W, Lu L, Jiang S. Development of protein- and peptide-based HIV entry inhibitors targeting gp120 or gp41. *Viruses.* (2019) 11:705. doi: 10.3390/v11080705

17. Jamjian MC, McNicholl IR. Enfuvirtide: first fusion inhibitor for treatment of HIV infection. *Am J Health Syst Pharm.* (2004) 61:1242–7. doi: 10.1093/ajhp/61.12.1242

18. Lalezari JP, Trottier B, Chung J, Salgo M. Enfuvirtide, an HIV-1 fusion inhibitor, for drug-resistant HIV Infection in North and South America. *N Engl J Med.* (2003) 348:2175–85. doi: 10.1056/NEJMoa035026

19. Georgiev IS, Gordon Joyce M, Zhou T, Kwong PD. Elicitation of HIV-1-neutralizing antibodies against the CD4-binding site. *Curr Opin HIV AIDS.* (2013) 8:382–92. doi: 10.1097/COH.0b013e328363a90e

20. Bashir T, Patgaonkar M, Kumar CS, Pasi A, Reddy KVR. HbAHP-25, an in-silico designed peptide, inhibits HIV-1 entry by blocking gp120 binding to CD4 receptor. *PLoS One.* (2015) 10:e0124839. doi: 10.1371/journal.pone.0124839

21. Landais E, Moore PL. Development of broadly neutralizing antibodies in HIV-1 infected elite neutralizers. *Retrovirology.* (2018) 15:61. doi: 10.1186/s12977-018-0443-0

22. Li H, Guan Y, Szczepanska A, Moreno-Vargas AJ, Carmona AT, Robina I, et al. Synthesis and anti-HIV activity of trivalent CD4-mimetic miniproteins. *Bioorgan Med Chem.* (2007) 15:4220–8. doi: 10.1016/j.bmc.2007.03.064

23. Lynch RM, Tran L, Louder MK, Schmidt SD, Cohen M, Chavi 001 Clinical Team Members, et al. The development of CD4 binding site antibodies during HIV-1 infection. *J Virol.* (2012) 86:7588–95. doi: 10.1128/JVI.00734-12

24. Pancera M, Zhou T, Druz A, Georgiev IS, Soto C, Gorman J, et al. Structure and immune recognition of trimeric pre-fusion HIV-1 Env. *Nature.* (2014) 514:455–61. doi: 10.1038/nature13808

25. Curreli F, Belov DS, Ramesh RR, Patel N, Altieri A, Kurkin AV, et al. Design, synthesis and evaluation of small molecule CD4-mimics as entry inhibitors possessing broad spectrum anti-HIV-1 activity. *Bioorgan Med Chem.* (2016) 24:5988–6003. doi: 10.1016/j.bmc.2016.09.057

26. Parker Miller E, Finkelstein MT, Erdman MC, Seth PC, Fera DA. Structural update of neutralizing epitopes on the HIV envelope, a moving target. *Viruses.* (2021) 13:1774. doi: 10.3390/v13091774

27. Kassler K, Meier J, Eichler J, Sticht H. Structural basis for species selectivity in the HIV-1 gp120-CD4 interaction: restoring affinity to gp120 in murine CD4 mimetic peptides. *Adv Bioinformat.* (2011) 2011:1–12. doi: 10.1155/2011/736593

28. Meier J, Kassler K, Sticht H, Eichler J. Peptides presenting the binding site of human CD4 for the HIV-1 envelope glycoprotein gp120. *Beilstein J Org Chem.* (2012) 8:1858–66. doi: 10.3762/bjoc.8.214

29. Vita C, Drakopoulou E, Vizzavona J, Rochette S, Martin L, Ménez A, et al. Rational engineering of a miniprotein that reproduces the core of the CD4 site interacting with HIV-1 envelope glycoprotein. *Proc Natl Acad Sci USA.* (1999) 96:13091–6. doi: 10.1073/pnas.96.23.13091

30. Apostolopoulos V, Bojarska J, Chai T-T, Elnagdy S, Kaczmarek K, Matsoukas J, et al. A global review on short peptides: frontiers and perspectives. *Molecules.* (2021) 26:430. doi: 10.3390/molecules26020430

31. Scheid JF, Mouquet H, Ueberheide B, Diskin R, Klein F, Oliveira TYK, et al. Sequence and structural convergence of broad and potent HIV antibodies that mimic CD4 binding. *Science.* (2011) 333:1633–7. doi: 10.1126/science.1207227

32. Bonsignori M, Zhou T, Sheng Z, Chen L, Gao F, Joyce MG, et al. Maturation pathway from germline to broad HIV-1 neutralizer of a CD4-mimic antibody. *Cell.* (2016) 165:449–63. doi: 10.1016/j.cell.2016.02.022

33. Wibmer CK, Moore PL, Morris L. HIV broadly neutralizing antibody targets. *Curr Opin HIV AIDS.* (2015) 10:135–43. doi: 10.1097/COH.0000000000000153

34. Zhou T, Lynch RM, Chen L, Acharya P, Wu X, Doria-Rose NA, et al. Structural repertoire of HIV-1-neutralizing antibodies targeting the CD4 supersite in 14 donors. *Cell.* (2015) 161:1280–92. doi: 10.1016/j.cell.2015.05.007

35. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, Kuroda D, et al. Serverification of molecular modeling applications: the rosetta online server that includes everyone (ROSIE). *PLoS One.* (2013) 8:e63906. doi: 10.1371/journal.pone.0063906

36. Sedan Y, Marcu O, Lyskov S, Schueler-Furman O. Peptiderive server: derive peptide inhibitors from protein–protein interactions. *Nucleic Acids Res.* (2016) 44:W536–41. doi: 10.1093/nar/gkw385

37. Laskowski RA, Jabłońska J, Pravda L, Vaøeková RS, Thornton JM. PDBsum: structural summaries of PDB entries. *Protein Sci.* (2018) 27:129–34. doi: 10.1002/pro.3289

38. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. Chimera?A visualization system for exploratory research and analysis. *J Comput Chem.* (2004) 25:1605–12. doi: 10.1002/jcc.20084

39. van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J Mol Biol.* (2016) 428:720–5. doi: 10.1016/j.jmb.2015.09.014

40. Honorato RV, Koukos PI, Jiménez-García B, Tsaregorodtsev A, Verlato M, Giachetti A, et al. Structural biology in the clouds: the WeNMR-EOSC ecosystem. *Front Mol Biosci.* (2021) 8:729513. doi: 10.3389/fmolb.2021.729513

41. Vangone A, Bonvin AM. Contacts-based prediction of binding affinity in protein–protein complexes. *Elife.* (2015) 4:e07454. doi: 10.7554/eLife.07454

42. Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics.* (2016) 32:3676–8. doi: 10.1093/bioinformatics/btw514

43. Chow E, Klepeis JL, Rendleman CA, Dror RO, Shaw DE. 9.6 new technologies for molecular dynamics simulations. In: Egelman E editor. *Comprehensive Biophysics*. Amsterdam: Elsevier (2012). p. 86–104. doi: 10.1016/B978-0-12-374920-8.00908-5

44. Bowers KJ, Chow DE, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters.

Proceedings of the ACM/IEEE SC 2006 Conference (SC'06). Tampa, FL: IEEE (2006). p. 43–43. doi: 10.1109/SC.2006.54

45. Muthukumaran S, Sulochana KN, Umashankar V. Structure based design of inhibitory peptides targeting ornithine decarboxylase dimeric interface and *in vitro* validation in human retinoblastoma Y79 cells. *J Biomol Struct Dyn.* (2021) 39:5261–75. doi: 10.1080/07391102.2020.1785331

46. 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

47. Singh N, Chaput L, Villoutreix BO. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace. *Brief Bioinformat.* (2021) 22:1790–818. doi: 10.1093/bib/bbaa034

48. Umashankar V, Deshpande SH, Hegde HV, Singh I, Chattopadhyay D. Phytochemical moieties from indian traditional medicine for targeting dual hotspots on SARS-CoV-2 spike protein: an integrative in-silico approach. *Front Med.* (2021) 8:672629. doi: 10.3389/fmed.2021.672629

49. Jiménez J, Škaliè M, Martínez-Rosell G, De Fabritiis G. KDEEP?: protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *J Chem Inf Model.* (2018) 58:287–96. doi: 10.1021/acs.jcim.7b00650

50. Karlov DS, Sosnin S, Fedorov MV, Popov P. graphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes. *ACS Omega.* (2020) 5:5150–9. doi: 10.1021/acsomega.9b04162

51. Varela-Rial A, Maryanow I, Majewski M, Doerr S, Schapin N, Jiménez-Luna J, et al. PlayMolecule glimpse: understanding protein–ligand property predictions with interpretable neural networks. *J Chem Inf Model.* (2022) 62:225–31. doi: 10.1021/acs.jcim.1c00691

52. Ibarra AA, Bartlett GJ, Hegedüs Z, Dutt S, Hobor F, Horner KA, et al. Predicting and experimentally validating hot-spot residues at protein–protein interfaces. *ACS Chem Biol.* (2019) 14:2252–63. doi: 10.1021/acschembio.9b00560

53. Wood CW, Ibarra AA, Bartlett GJ, Wilson AJ, Woolfson DN, Sessions RB. BAlaS: fast, interactive and accessible computational alanine-scanning using BudeAlaScan. *Bioinformatics.* (2020) 36:2917–9. doi: 10.1093/bioinformatics/btaa026

54. Kortemme T, Kim DE, Baker D. Computational alanine scanning of protein-protein interfaces. *Sci STKE.* (2004) 2004:l2. doi: 10.1126/stke.2192004pl2

55. Gupta S, Kapoor P, Chaudhary K, Gautam A, Kumar R. Open source drug discovery consortium, raghava GPS. In silico approach for predicting toxicity of peptides and proteins. *PLoS One.* (2013) 8:e73957. doi: 10.1371/journal.pone.0073957

56. Prévost J, Tolbert WD, Medjahed H, Sherburn RT, Madani N, Zoubchenok D, et al. The HIV-1 Env gp120 inner domain shapes the Phe43 cavity and the CD4 binding site. *mBio.* (2020) 11:e00280-20. doi: 10.1128/mBio.00280-20.

57. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, et al. Multidonor Analysis reveals structural elements, genetic determinants, and maturation pathway for HIV-1 neutralization by VRC01-class antibodies. *Immunity.* (2013) 39:245–58. doi: 10.1016/j.immuni.2013.04.012

58. Curreli F, Belov DS, Kwon YD, Ramesh R, Furimsky AM, O'Loughlin K, et al. Structure-based lead optimization to improve antiviral potency and ADMET properties of phenyl-1H-pyrrole-carboxamide entry inhibitors targeted to HIV-1 gp120. *Eur J Med Chem.* (2018) 154:367–91. doi: 10.1016/j.ejmech.2018.04.062

59. Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. *Drug Discovery Today.* (2015) 20:122–8. doi: 10.1016/j.drudis.2014.10.003

60. Lau JL, Dunn MK. Therapeutic peptides: historical perspectives, current development trends, and future directions. *Bioorgan Med Chem.* (2018) 26:2700–7. doi: 10.1016/j.bmc.2017.06.052

61. Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. *Nat Rev Drug Discov.* (2021) 20:309–25. doi: 10.1038/s41573-020-00135-8

62. Amit C, Muralikumar S, Janaki S, Lakshmipathy M, Therese KL, Umashankar V, et al. Designing and enhancing the antifungal activity of corneal specific cell penetrating peptide using gelatin hydrogel delivery system. *IJN.* (2019) 14:605–22. doi: 10.2147/IJN.S184911

63. Cabri W, Cantelmi P, Corbisiero D, Fantoni T, Ferrazzano L, Martelli G, et al. Therapeutic peptides targeting PPI in clinical development: overview, mechanism of action and perspectives. *Front Mol Biosci.* (2021) 8:697586. doi: 10.3389/fmolb.2021.697586

64. Baig MH, Ahmad K, Saeed M, Alharbi AM, Barreto GE, Ashraf GM, et al. Peptide based therapeutics and their use for the treatment of neurodegenerative and other diseases. *Biomed Pharmacother.* (2018) 103:574–81. doi: 10.1016/j.biopha.2018.04.025

65. Chen RP. From nose to brain: the promise of peptide therapy for Alzheimer's Disease and other neurodegenerative diseases. *J Alzheimers Dis Parkinson.* (2017) 7:314. doi: 10.4172/2161-0460.1000314

66. Ferrer M, Harrison SC. Peptide ligands to human immunodeficiency virus type 1 gp120 identified from phage display libraries. *J Virol.* (1999) 73:5795–802. doi: 10.1128/JVI.73.7.5795-5802.1999

67. Martin L, Stricher F, Missé D, Sironi F, Pugnière M, Barthe P, et al. Rational design of a CD4 mimic that inhibits HIV-1 entry and exposes cryptic neutralization epitopes. *Nat Biotechnol.* (2003) 21:71–6. doi: 10.1038/nbt768

68. Stricher F, Huang C, Descours A, Duquesnoy S, Combes O, Decker JM, et al. Combinatorial optimization of a CD4-mimetic miniprotein and cocrystal structures with HIV-1 gp120 envelope glycoprotein. *J Mol Biol.* (2008) 382:510–24. doi: 10.1016/j.jmb.2008.06.069

69. Van Herrewege Y, Morellato L, Descours A, Aerts L, Michiels J, Heyndrickx L, et al. CD4 mimetic miniproteins: potent anti-HIV compounds with promising activity as microbicides. *J Antimicrobial Chemother.* (2008) 61:818–26. doi: 10.1093/jac/dkn042

70. Choi YH, Rho WS, Kim ND, Park SJ, Shin DH, Kim JW, et al. Short peptides with induced β-turn inhibit the interaction between HIV-1 gp120 and CD4. *J Med Chem.* (2001) 44:1356–63. doi: 10.1021/jm000403

71. Biorn AC, Cocklin S, Madani N, Si Z, Ivanovic T, Samanen J, et al. Mode of action for linear peptide inhibitors of HIV-1 gp120 interactions. *Biochemistry.* (2004) 43:1928–38. doi: 10.1021/bi035088i

72. Andrianov AM, Kashyn IA, Tuzikov AV. Computational identification of novel entry inhibitor scaffolds mimicking primary receptor CD4 of HIV-1 gp120. *J Mol Model.* (2017) 23:18. doi: 10.1007/s00894-016-3189-4

73. Andrianov AM, Kashyn IA, Tuzikov AV. Computational discovery of novel HIV-1 entry inhibitors based on potent and broad neutralizing antibody VRC01. *J Mol Graph Modell.* (2015) 61:262–71. doi: 10.1016/j.jmgm.2015.08.003

74. Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc Natl Acad Sci USA.* (2002) 99:14116–21. doi: 10.1073/pnas.202485799

75. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* (1982) 157:105–32. doi: 10.1016/0022-2836(82)90515-0

76. Chang KY, Yang J-R. Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS One.* (2013) 8:e70166. doi: 10.1371/journal.pone.0070166

77. Scharf L, West AP, Gao H, Lee T, Scheid JF, Nussenzweig MC, et al. Structural basis for HIV-1 gp120 recognition by a germ-line version of a broadly neutralizing antibody. *Proc Natl Acad Sci USA.* (2013) 110:6049–54. doi: 10.1073/pnas.1303682110

78. Liao H-X, Lynch R, Zhou T, Gao F, Alam SM, Boyd SD, et al. Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature.* (2013) 496:469–76. doi: 10.1038/nature12053

frontiers | Frontiers in Physiology

TYPE Original Research
PUBLISHED 23 November 2022
DOI 10.3389/fphys.2022.1031996

OPEN ACCESS

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

REVIEWED BY
Yue Victor Zhang,
Shenzhen Futian Hospital for Rheumatic Diseases, China
Cheng-Rong Yu,
National Eye Institute (NIH), United States

*CORRESPONDENCE
Kanglai Tang,
tangkanglai@tmmu.edu.cn
Taotao Liang,
liangtaotao2018@foxmail.com

SPECIALTY SECTION
This article was submitted to Computational Physiology and Medicine,
a section of the journal Frontiers in Physiology

RECEIVED 30 August 2022
ACCEPTED 07 November 2022
PUBLISHED 23 November 2022

CITATION
Guo J, Tang C, Shu Z, Guo J, Tang H, Huang P, Ye X, Liang T and Tang K (2022), Single-cell analysis reveals that Jinwu Gutong capsule attenuates the inflammatory activity of synovial cells in osteoarthritis by inhibiting AKR1C3. *Front. Physiol.* 13:1031996.
doi: 10.3389/fphys.2022.1031996

Single-cell analysis reveals that Jinwu Gutong capsule attenuates the inflammatory activity of synovial cells in osteoarthritis by inhibiting AKR1C3

Junfeng Guo[1], Chuyue Tang[1], Zhao Shu[2], Junfeng Guo[3], Hong Tang[1], Pan Huang[1], Xiao Ye[1], Taotao Liang[1]* and Kanglai Tang[1]*

[1]State Key Laboratory of Trauma, Burn and Combined Injury, Department of Orthopedics/Sports Medicine Center, Southwest Hospital, Third Military Medical University, Chongqing, China, [2]The First Affiliated Hospital of Chongqing Medical University, Chongqing, China, [3]Department of Stomatology, The 970th Hospital of the Joint Logistics Support Force, Yantai, China

Jinwu Gutong capsule (JGC) is a traditional Chinese medicine formula for the treatment of osteoarthritis (OA). Synovitis is a typical pathological change in OA and promotes disease progression. Elucidating the therapeutic mechanism of JGC is crucial for the precise treatment of OA synovitis. In this study, we demonstrate that JGC effectively inhibits hyperproliferation, attenuates inflammation, and promotes apoptosis of synovial cells. Through scRNA-seq data analysis of OA synovitis, we dissected two distinct cell fates that influence disease progression (one fate led to recovery while the other fate resulted in deterioration), which illustrates the principles of fate determination. By intersecting JGC targets with synovitis hub genes and then mimicking picomolar affinity interactions between bioactive compounds and binding pockets, we found that the quercetin-AKR1C3 pair exhibited the best affinity, indicating that this pair constitutes the most promising molecular mechanism. *In vitro* experiments confirmed that the expression of AKR1C3 in synovial cells was reduced after JGC addition. Further overexpression of AKR1C3 significantly attenuated the therapeutic efficacy of JGC. Thus, we revealed that JGC effectively treats OA synovitis by inhibiting AKR1C3 expression.

Abbreviations: OA, osteoarthritis; JGC, Jinwu Gutong capsule; GDP, gross domestic product; NSAIDs, nonsteroidal anti-inflammatory drugs; GSs, glucocorticoids; DMSO, dimethyl sulfoxide; CCK-8, Cell Counting Kit-8; BEAM, branched expression analysis modeling; CB, Cibotium barometz; ED, Epimedium; CR, Clematidis radix; ZD, Zaocys dhumnades; ABB, *Achyranthes* bidentata Blume; CS, Chaenomeles sinensis; PL, Pueraria lobata; CL, curcuma longa; PCL, psoralea corylifolia Linn.; CJB, Campanumoea javanica bl; FITC, fluorescein isothiocyanate isomer; ROS, reactive oxygen species; SFs, synovial fibroblasts; PCR, polymerase chain reaction; UMAP, uniform manifold approximation and projection.

KEYWORDS
Jinwu Gutong capsule, synovitis, osteoarthritis, scRNA-seq, AKR1C3

Introduction

Osteoarthritis (OA) is the most common age-related chronic degenerative whole-joint disease and affects more than 300 million people worldwide (Choi et al., 2019; Boer et al., 2021). OA imposes a severe social and economic burden, and its total costs are estimated to equal 1%–2.5% of a country's gross domestic product (GDP) (Hiligsmann et al., 2013; Brown et al., 2021). The main pathological features of OA are cartilage degeneration and synovial inflammation (Sellam and Berenbaum, 2010). Increasing evidence indicates that synovial inflammation not only is directly linked to clinical symptoms such as joint swelling and inflammatory pain but also increases cartilage injury (Atukorala et al., 2016; Labinsky et al., 2020). Thus, inhibiting synovitis is a crucial aspect of preventing OA development.

The current treatments for synovitis mainly include nonsteroidal anti-inflammatory drugs (NSAIDs) and glucocorticoids (GSs), but their effects are often short-lived and may even lead to a greater degree of cartilage loss (Conaghan et al., 2019; Pontes-Quero et al., 2021). Jinwu Gutong capsule (JGC) is a traditional botanical formula widely used in China for OA treatment and is widely believed to have considerable potential with respect to clinical efficacy (Zhao et al., 2022). Indeed, the combined application of JGC with NSAIDs or GS can significantly improve the efficacy of OA treatment. However, the pharmacological mechanism of JGC remains unclear and warrants further research.

Single cell sequencing provides insights into the underlying mechanisms of OA development. Early research mainly focused on cartilage degeneration: Tang et al. identified seven molecularly defined populations of chondrocytes in the human OA cartilage (Ji et al., 2019); Jeon et al. (2017) found that $p16^{INK4a}$ positive senescent chondrocytes contribute to the development of spontaneous and injury-induced OA. In recent years, people have increasingly recognized the important role of synovitis in the development of OA. Nanus et al. (2021) illustrated that there are distinct synovial fibroblast subsets in early OA and end-stage OA. Knights et al. (2022) displayed $Prg4^{hi}$ lining fibroblasts secrete Rspo2, which drives pathological joint crosstalk after injury.

In this study, we demonstrate the therapeutic effect of JGC on synovial inflammation and hyperplasia. A single-cell synovial atlas was produced, which allowed an in-depth exploration of the synovial microenvironment. Further transcriptional dynamics analysis revealed a cell fate decision mechanism that affects disease progression and recovery. We also identified the target of JGC in treating OA synovitis and verified this target through computer simulations and biological experiments.

TABLE 1 Molecular docking results.

Bioactive compounds	Targets	affinity (kcal/mol)
quercetin	AKR1C3	−10.1
syringetin	CYP1B1	−9.3
apigenin	CYP1B1	−8.3
quercetin	MMP2	−8.2
quercetin	CYP1B1	−7.9
chlorogenic acid	MMP2	−7.7
apigenin	PTGS2	−7.7
icariside F2	VEGFA	−2.3

Materials and methods

Preprocessing of Jinwu Gutong capsule

Commercial JGC (specification: 0.5 g per pill) was purchased from Guizhou SSLF Pharmaceutical Co., Ltd. (Guizhou, China, approval number: Z20043621). According to the literature (Sridhar et al., 2021), JGC was powdered and extracted using a Soxhlet extractor with 6 times the amount of 90% ethanol. The solvent was then concentrated using an electrically heated blast drying oven at 45°C. Subsequently, the concentrate was lyophilized with a freeze dryer and weighed. The JGC extract was dissolved in DMSO (20 mg/ml) and stored at −80°C for later use.

Cell culture

The human synovial cell line SW982 was kindly provided by Procell Life Science and Technology Co., Ltd. (Wuhan, China). SW982 cells have been shown to possess characteristic features similar to synovial fibroblasts which makes them an ideal tool to study synovitis in OA (Karuppagounder et al., 2022). The cells were cultured in DMEM/Ham's F12 medium (DMEM/F12; HyClone, Logan, UT, United States) with 10% fetal bovine serum (PAN Biotech, Aidenbach, Germany) and 1% penicillin/streptomycin (Gibco, Grand Island, NY, United States).

Detection of cell proliferation

The cell proliferative capacity was determined by Cell Counting Kit-8 assays (CCK-8, Biosharp, Guangzhou, China). Cells (10,000/well) were plated in 96-well plates, and DMSO, CTGF or JGC was added according to the experimental design. CTGF is a pro-inflammatory cytokine, that is, upregulated in OA

and contributes to synovial hyperplasia (MacDonald et al., 2021). The working concentration of CTGF was 25 ng/ml, and that of JGC was 20 μg/ml. After 24 h, the supernatant was replaced with CCK-8 working solution, and the absorbance at 450 nm was measured.

Apoptosis detection

An Annexin V-FITC Assay Kit (Merck, NJ, United States) was used to detect apoptosis in synovial cells. The cells were plated in 6-well plates (50,000/well) and processed as described above. After 24 h, the cells were dissociated and stained according to the instructions provided with the kit. In brief, cells were digested with trypsin, washed gently with PBS, resuspended in buffer solution to 1×10^6 cells/ml. Then 5 μl Annexin V-FITC was added, and the mixture was incubated in the dark for 5 min 5 μl propidium iodide (PI) was added to the cells before analyzed. We measured the proportion of FITC(+) cells by flow cytometry.

Data sources and processing

Single-cell sequencing data for synovial cells were downloaded from the GEO database (no. GSE176308), and 10X genomics data were loaded into the R package Seurat (v4.0.2). Synovial cells were obtained from 4 patients with early-stage OA (both painful and non-painful sites) and 4 patients with end-stage OA (painful sites) (Nanus et al., 2021). Cell quality control was applied to remove low-quality cells with less than 300 detected genes or with more than 10% mitochondrial genes. After normalizing the data, the cells were dimensionally reduced and clustered according to the top 2,000 highly variable genes. The FindIntegrationAnchors algorithm found a set of anchors between Seurat objects from different patients. These anchors could be used to integrate the objects using the IntegrateData function. Harmony package (v1.0) was used to remove the batch effect, the diversity clustering penalty parameter was set to 2 and the ridge regression penalty parameter was set to 1.

Pseudotime analysis

The dynamic states of synovial cells were assessed using the Monocle algorithm (v2.18.0). Monocle uses an unsupervised algorithm to order whole-transcriptome profiles of single cells and produce a 'trajectory' of an individual cell's progress through differentiation. We applied the "reduceDimension" function to compute the CellDataSet object as a lower dimensional trajectory. The Discriminative Dimensionality Reduction with Trees (DDRTree) method was chosen for its ability to reduce dimensionality while discriminating between different data points. Following dimension reduction, the two features with the most significant amount of information were extracted and used as the coordinate axes to visualize the trajectory. Branched expression analysis modeling (BEAM) was performed to identify genes with branch-dependent expression and thus elucidate fate decision mechanisms.

Cell cycle analysis

Independent cell cycle analysis was performed for each synovial cell. The "CellCycleScoring" function in the Seurat package was used to assign cell cycle scores according to S- and G2/M-phase genes, which were identified following procedures described in a previous study (Kan et al., 2022). The number of control features selected from the same bin per analyzed feature was set to 100 and the random seed was set to 1. The cells were classified into G1, S, and G2/M phases based on the maximal score of each cell cycle phase program.

Jinwu Gutong capsule target prediction

We obtained information regarding the main raw materials from the JGC drug manual. Information about the main active ingredients of these raw materials was obtained from the relevant literature (Supplementary Table S2). The SDF format files of molecular structures were downloaded from the Pubchem database (https://pubchem.ncbi.nlm.nih.gov/). Targets of these molecular structures were predicted using the SwissTargetPrediction database (http://www.swisstargetprediction.ch/) (Daina et al., 2019). The species was confined to "*Homo sapiens*", and the predicted targets with a probability more than 0.3 were included in this study.

Molecular docking

Macromolecular structures were downloaded from the RCSB PDB database (https://www.rcsb.org), and biological ligands were accessed from PubChem database. PDB files were converted to the PDBQT format. We used AutoDockTools software to search for possible active pockets, removed all water molecules and assigned hydrogen polarities. AutoDock Vina was employed to conduct molecular docking between the active ingredients and targets, then took the conformation with the highest docking score (Affinity). Finally, we used the PyMOL software to visualize the results of molecular docking.

Statistical analysis

Bilateral tests were performed for all statistical tests. A p-value lower than 0.05 was considered to indicate statistical

significance. R software version 4.0.2 (https://www.r-project.org/) was used for the analysis. The following R language packages were used in this study: "dplyr", "Seurat", "monocle", "monocle", and "iTALK". The "drug-material-target" network was visualized using Cytoscape_3.7.2 (https://cytoscape.org).

Results

Jinwu Gutong capsule exerts ideal therapeutic effects on reducing inflammation and hyperplasia of synovial cells

JGC is widely used for OA treatment with ideal clinical efficacy. According to the instructions, the main raw materials of JGC include *Cibotium barometz* (CB [Cyatheaceae; *Cibotium barometz* (L.) J. Sm]), Epimedium (ED [Berberidaceae; *Epimedium sagittatum* (Siebold & Zucc.) Maxim]), Clematidis radix (CR [Ranunculaceae; Clematis chinensis Osbeck]), Zaocys dhumnades (ZD [Colubridae]), *Achyranthes bidentata* Blume (ABB [Amaranthaceae; *Achyranthes bidentata* Blume]), Chaenomeles sinensis (CS [Rosaceae; *Pseudocydonia sinensis* (Dum.Cours.) C.K. Schneid]), *Pueraria lobata* (PL [Fabaceae; *Pueraria montana* var. lobata (Willd.) Maesen & S.M. Almeida ex Sanjappa & Predeep]), *Curcuma longa* (CL [Zingiberaceae; *Curcuma longa* L., Sp. Pl.: 2 (1753)]), *Psoralea corylifolia* Linn. (PCL [Fabaceae; *Cullen corylifolium* (L.) Medik]), and *Campanumoea javanica* bl (CJB [Campanulaceae; *Codonopsis javanica* (Blume) Hook. f. & Thomson, Ill. Himal. Pl. t.16 B (1855)]). Certain materials (ED, ABB, CS, PL, CL, and CR) reportedly have significant anti-inflammatory and antioxidant activities, and the aqueous extract of CR exerts a good anti-osteoarthritis effect (Cheng et al., 2013; Lin et al., 2019; Cheng et al., 2020; Jeon et al., 2020; Lin et al., 2021; Razavi et al., 2021). The reasonable compatibility of these materials guarantees curative efficacy.

Synovial tissue shows discordant hyperplasia and inflammation during OA progression. The human synovial cell line SW982 was treated with JGC to assess the effect of this drug on synovial hyperplasia. In normal synovial cells, the inhibition of proliferation by JGC was not significant, indicating tolerable drug toxicity. We then induced hyperproliferation using the growth factor CTGF, and JGC exerted a more pronounced inhibitory effect on the proliferation of active synovial cells (Figure 1A). Flow cytometry showed that the proportion of FITC(+) synovial cells was significantly increased, showing the apoptosis-promoting effect of JGC on SW982 cells (Figure 1B). The inflammatory cytokine IL-1β was applied to induce intense inflammation in synovial cells. Although the expression levels of numerous inflammatory genes (IL-1β, IL-6, IL-8, NOS2, and TNF-α) were clearly increased, JGC treatment significantly reversed the increase in expression caused by inflammatory stimulation (Figure 1C). We also found similar trends for the intracellular reactive oxygen species (ROS) levels: inflammation led to increased ROS levels in SW982 cells, and this increase was relieved after JGC addition (Figure 1D). These results confirm the therapeutic effect of JGC on synovitis *in vitro*.

Cellular composition and communication of synovial microenvironment in osteoarthritis

To deeply dissect the molecular mechanism of JGC in the treatment of OA synovitis, scRNA-seq data from 4145 synovial fibroblasts (SFs) were examined in this study. SFs were clustered into nine color-labeled subsets based on their unbiased transcriptome signatures (Figure 2A). The cell cluster properties were preliminarily assessed based on cluster-specific markers (Figures 2B,C; Supplementary Figure S1; Supplementary Table S1): the cells in SF-0 expressed high levels of IGFBP6, MFAP5, and SEMA3C, indicating their high proliferative capacity; the cells in SF-1 overexpressed CXCL12 and ID1, suggesting a stronger inflammatory stimulus; the cells in SF-2 expressed MMP2 and WISP2, which play decisive roles in fibrosis; the cells in SF-5 showed relatively high expression of Piezo2, a mechanosensitive channel; the cells in SF-6 expressed RNASE1, indicating decreased adhesion to cartilage; the cells in SF-7 expressed genes critical for synovial angiogenesis (expressing SCUBE3); and the cells in SF-8 expressed relatively high levels of a cell cycle-related gene (CENPM).

We further calculated module scores to assess their inflammatory and proliferative activities, which are the two most prominent pathological features of synovitis. Consistent with the abovementioned results, the SF-1 synovial cells showed the highest level of inflammation, whereas the SF-0 cells exhibited an excessive proliferative capacity (Figures 3A,B). Overall, the proportions of cells from patients with or without pain, according to clinical information, did not significantly differ among the clusters; however, higher proportions of cells in SF-0, SF-1, and SF-2 were obtained from end-stage OA patients (Figures 3C,D). A cell–cell communication analysis revealed complex ligand–receptor interactions in the synovial microenvironment, and intercellular crosstalk was mainly divided into cytokines, growth factors and others (Figure 3E). Based on the cytokine categories, the synovial cells in SF-1 expressed higher levels of CXCL12, which interacts with the ITGB1 receptor of surrounding cells to regulate proinflammatory cytokine production (Kong et al., 2020). The growth factor category revealed that CTGF secreted by SF-7 cells interacts with LRP1, which is highly expressed on the surface of cells in other clusters, to induce pathological progression (Schnieder et al., 2020).

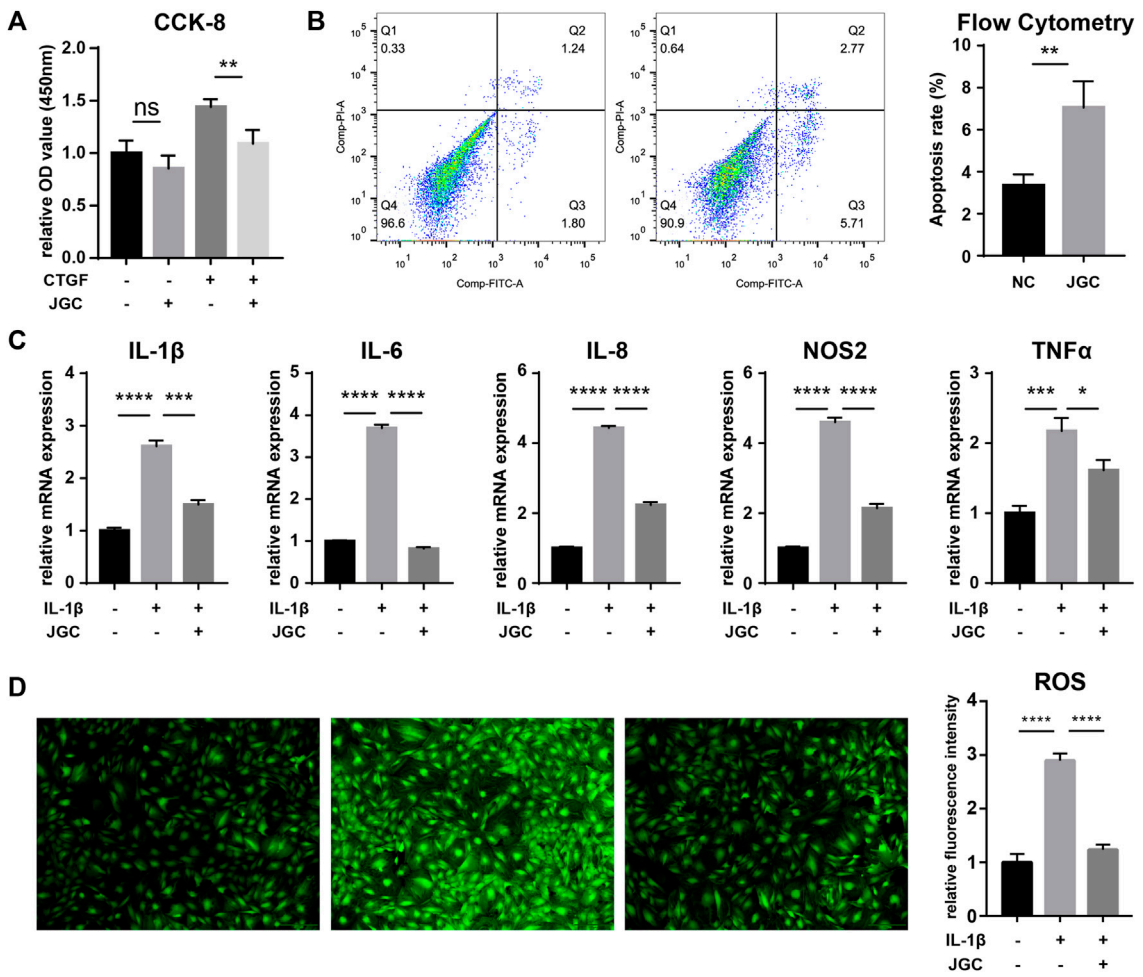


FIGURE 1
Therapeutic effect of JGC on synovitis. **(A)** CCK-8 assay showing the effect of JGC on cell proliferation. **(B)** Flow cytometry showing the effect of JGC on apoptosis. **(C)** PCR showing that JGC effectively inhibits synovial inflammation. **(D)** JGC clearly reduces the intracellular ROS levels.

Transcriptional dynamics analysis reveals the regulation of synovial cell fate decisions

The Monocle pseudotime algorithm was used to profile the fate trajectory of synovial cells. The cells were dimensionally descended and arranged in a typical dendritic shape (Figure 4A), and the fate trajectory was divided into three cell states based on bifurcation points (Figure 4B, state 1 to state 3). By comparing the gene patterns in distinct cell states, we found certain classical progenitor/stem cell markers to be significantly overexpressed in cell state 1 (OCT-4, TRA-1-81, SSEA4, NANOG, etc.). Thus, cell state 1 was defined as the origin of the trajectory (Figure 4C), and the synovial cells gradually differentiated into two distinctive fates as the trajectory progressed (Figure 4D).

We screened for "branch-dependent" genes that changed as the cell fate developed and divided these genes into two gene modules. A Gene Ontology (GO) enrichment analysis of "branch-dependent" genes helped annotate the cellular properties across different cell fates (Figure 4E). Certain functions that are beneficial to synovitis recovery were significantly activated in cell fate 1 (e.g., negative regulation of the inflammatory response and cell growth). However, some terms that suggest pathogenesis were enhanced in cell fate 2 (such as positive regulation of angiogenesis). The expression patterns of some canonical synovitis regulators were further assessed, and certain restorative genes (such as NMB, APOE and SMAD7) were highly expressed in cell fate 1 but decreased in cell fate 2. In addition, some pathogenic genes, such as ASPN and ACTA2, showed completely contrary trends (Figure 4F). A cell cycle analysis showed that the proportion of actively proliferating cells (G2/M) was significantly higher in cell fate 2, indicating likely tissue hyperplasia (Figure 4G). What's more, the two pathways associated with pain (prostanoid and eicosanoid signaling) showed increased activation in cell fate 2, suggesting that these cells were more likely to induce

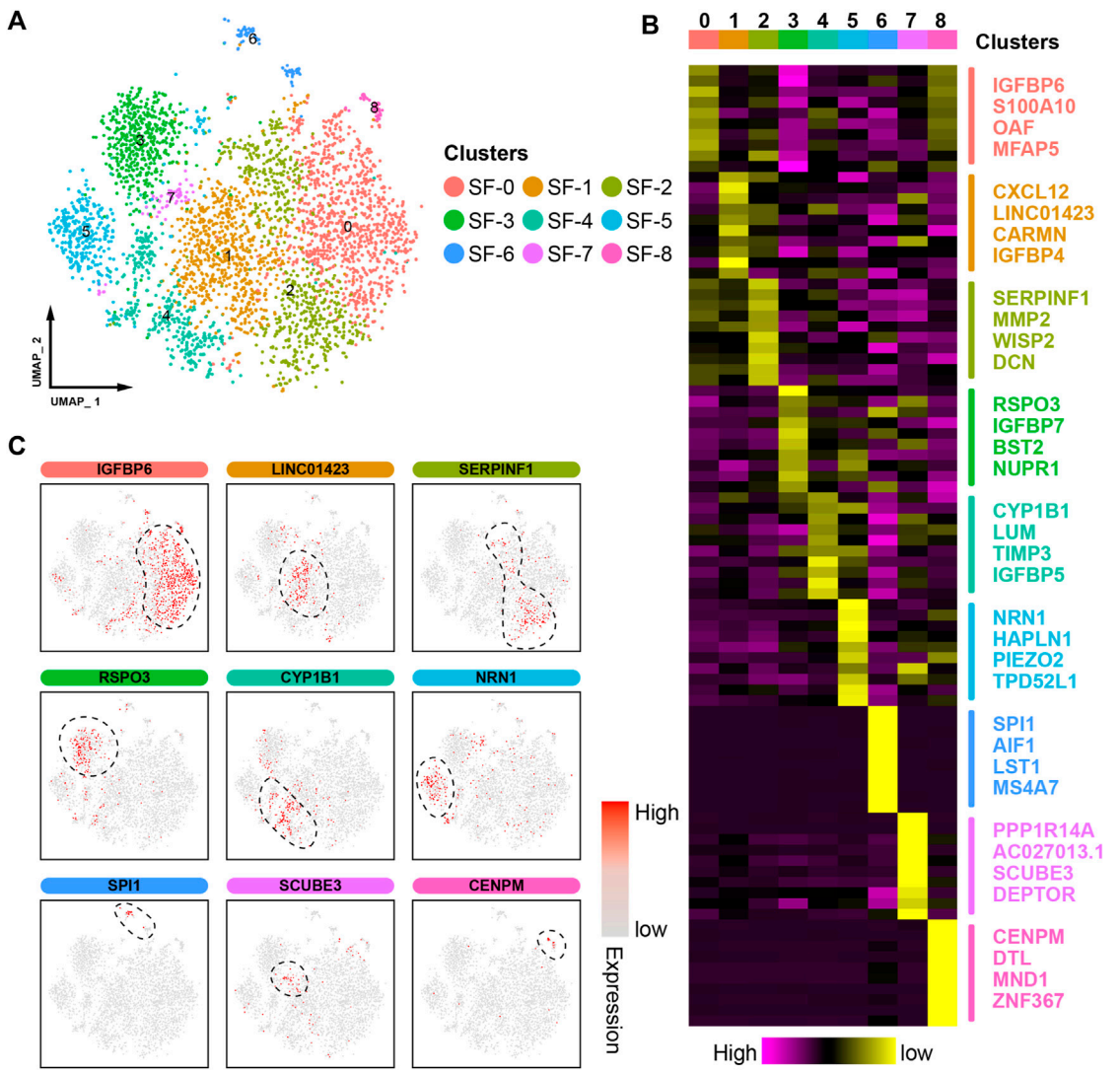


FIGURE 2
ScRNA-seq profiling of synovitis microenvironments. **(A)** A uniform manifold approximation and projection (UMAP) plot showing the color-coded cell clusters in the synovitis microenvironment. **(B)** Heatmap showing the marker gene expression in the different cell clusters. **(C)** UMAP plot showing the marker gene expression in the different cell clusters.

clinical symptoms (Figure 4H). In summary, these results suggest that cells in cell fate 1 contribute to recovery and that cells in cell fate 2 lead to synovitis progression.

Jinwu Gutong capsule treats synovitis by inhibiting AKR1C3

A differential expression analysis between the two cell fates identified a total of 403 key synovitis genes, including 195 and 208 upregulated genes in cell fate 1 and cell fate 2, respectively (Figure 5A). Furthermore, by summarizing previous research results, we collected 122 bioactive molecules from the raw materials of JGC (Supplementary Table S2). Subsequently, 151 potential targeting relationship pairs were predicted from the SwissTargetPrediction database (Supplementary Table S3), and a "drug-material-target" network was generated to visualize the potential therapeutic mechanism (Figure 5B). By taking the intersection of JGC targets with key genes of synovitis, five promising functional targets (AKR1C3, VEGFA, CYP1B1, MMP2, and PTGS2) were obtained (Figure 5C). Molecular docking was performed to simulate the interaction between bioactive compounds and binding pockets, which revealed a molecular basis for this picomolar affinity (Supplementary Figure S2). The quercetin-AKR1C3 pair exhibited the best affinity, indicating that this pair

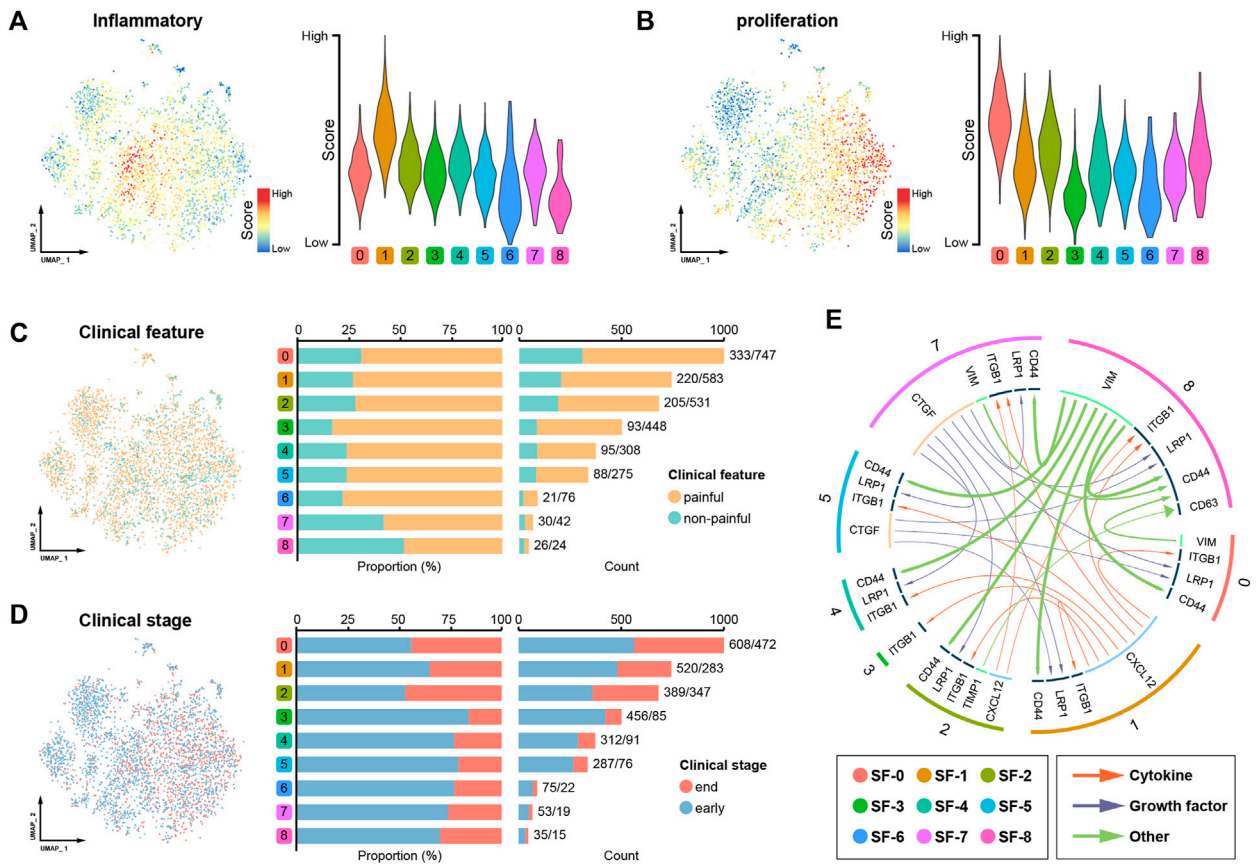


FIGURE 3
Assessment of the synovial microenvironment and intercellular communication. **(A)** UMAP plot showing the level of inflammation in the different cell clusters. **(B)** UMAP plot showing the proliferation ability of the different cell clusters. **(C)** Distribution of cells from patients with or without pain. **(D)** Distribution of cells from early- and end-stage OA patients. **(E)** Cell–cell communication in the synovial microenvironment.

constitutes the most promising molecular mechanism (Figures 5D,E; Table 1).

Further PCR results confirmed the hypothesis that AKR1C3 expression was elevated in inflamed synovial cells and effectively inhibited by the addition of JGC (Figure 6A). Rescue experiments were performed to characterize the regulatory relationship. AKR1C3 overexpression significantly attenuated the JGC-induced inhibitory effect on synovial cell proliferation (Figure 6B). Similarly, the anti-inflammatory effect of JGC on synovial cells was clearly counteracted by AKR1C3 overexpression (Figure 6C). Taken together, our findings suggest that JGC treats synovitis in osteoarthritis by inhibiting AKR1C3.

Discussion

OA is a chronic degenerative disease that involves pain and disability, resulting in poor quality of life (Xie et al., 2021). Severe synovitis is one of the typical pathological features of OA and leads to disease progression (Jin et al., 2011; Zhang et al., 2022). Certain botanical drugs, such as saponins and kaempferol, have been shown to act as effective therapeutics in inflammatory diseases (Devi et al., 2015; Dong et al., 2019). As a traditional botanical formula, JGC has been widely used in clinical practice and exerts good curative effects on OA synovitis. Thus, elucidating the molecular mechanism of JGC has important academic value and broad application prospects.

The pathological changes occurring in the OA synovium mainly include inflammation, hyperplasia and fibrosis, all of which usually coexist (Kuang et al., 2020). Our study shows that JGC effectively inhibits the expression of proinflammatory factors in synovial cells and reduces the intracellular ROC levels in these cells. Furthermore, JGC restrained the overproliferation of and induced apoptosis in synovial cells. These results confirm the therapeutic effect of JGC on synovitis at the cellular level, which complements the results from previous studies.

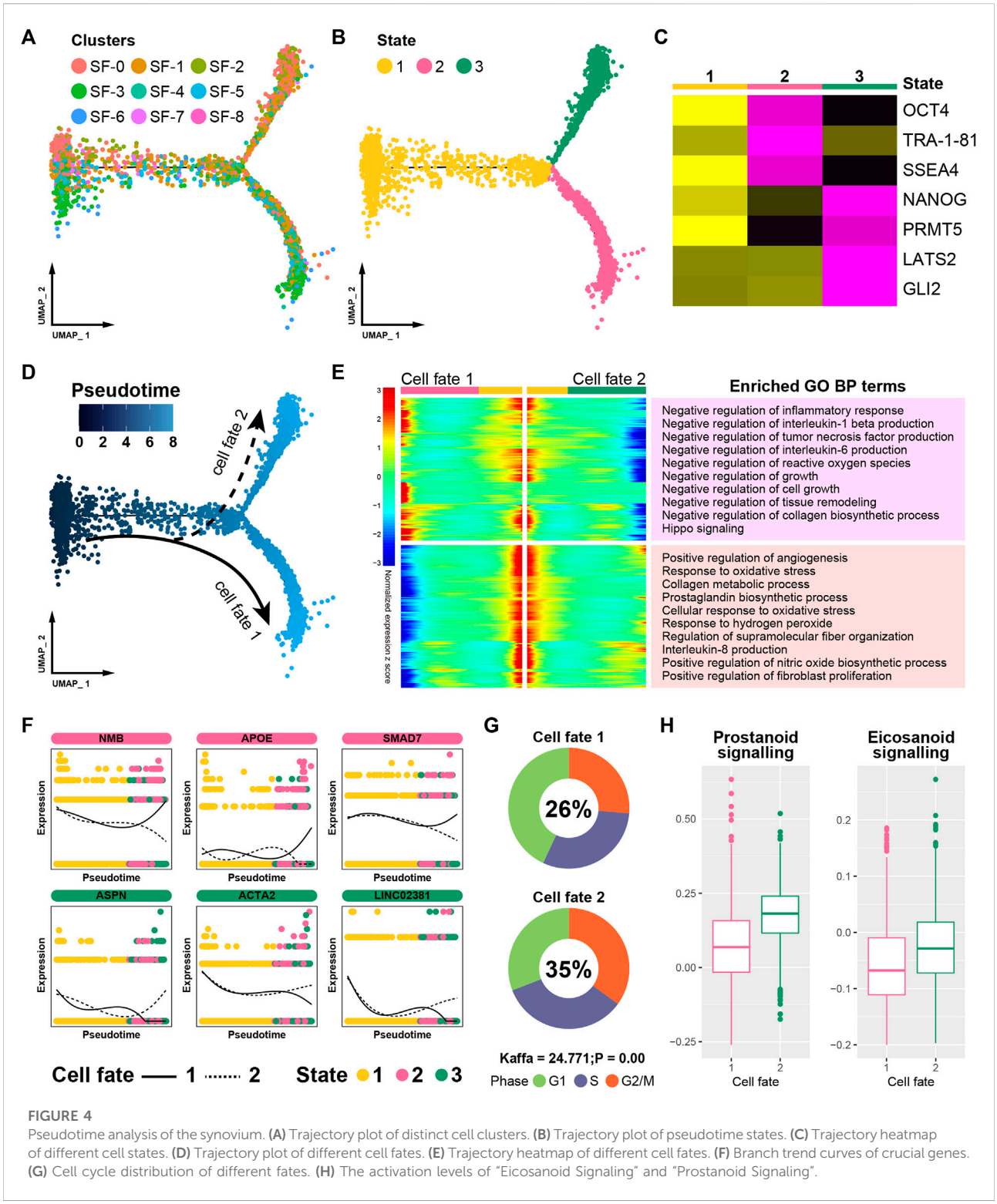


FIGURE 4
Pseudotime analysis of the synovium. **(A)** Trajectory plot of distinct cell clusters. **(B)** Trajectory plot of pseudotime states. **(C)** Trajectory heatmap of different cell states. **(D)** Trajectory plot of different cell fates. **(E)** Trajectory heatmap of different cell fates. **(F)** Branch trend curves of crucial genes. **(G)** Cell cycle distribution of different fates. **(H)** The activation levels of "Eicosanoid Signaling" and "Prostanoid Signaling".

A pseudotime analysis revealed the transcriptional dynamics and cell trajectory fates of synovial cells. In addition to the inflammation-, proliferation-, and fibrosis-related terms mentioned above, we found that the Hippo pathway was significantly activated in cell fate 1. The cells in cell fate 1 were identified as synovitis repair cells, and certain previous studies support our conclusion that activation of the Hippo pathway by verteporfin significantly reduces the severity of

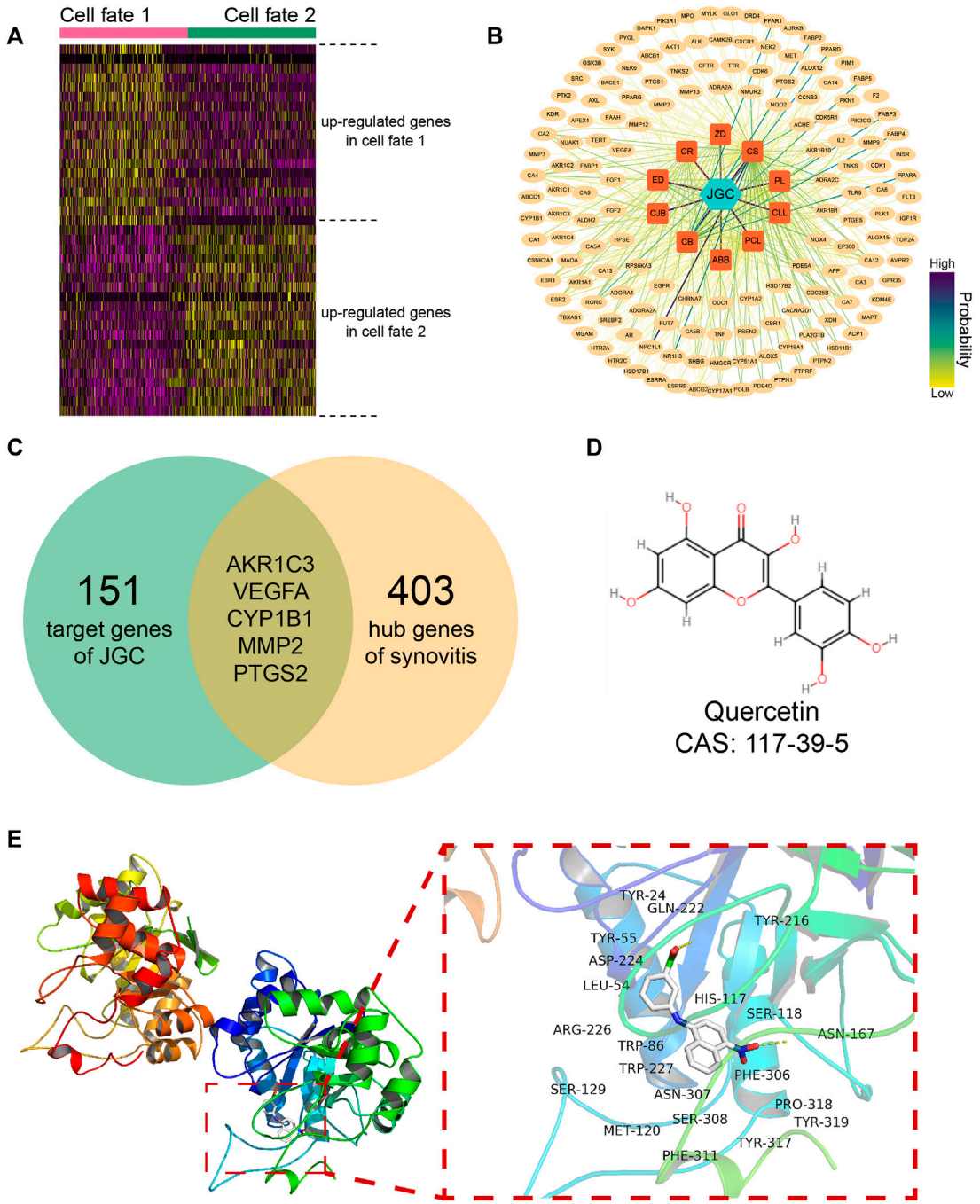


FIGURE 5
JGC treats synovitis by inhibiting AKR1C3. **(A)** Heatmap showing differentially expressed genes among distinct cell fates. **(B)** Network showing predicted targets of JGC. **(C)** Venn diagram showing the intersection of JGC targets with hub genes of synovitis. **(D)** Molecular structure of quercetin. **(E)** Molecular docking pattern of the quercetin-AKR1C3 pair.

synovitis (Caire et al., 2021; Symons et al., 2022). Certain key genes (APOE and SMAD7) were found to silence cell fate 2. Apolipoprotein E, a major apoprotein of the chylomicron, inhibits synovial activation and ectopic bone formation (de Munter et al., 2016); in contrast, Smad7 loss promotes synovial inflammation and fibrosis (Blaney Davidson et al., 2006; Zhou et al., 2018). Moreover, the expression of several disease progression genes (ASPN, ACTA2 and LINC02381) was

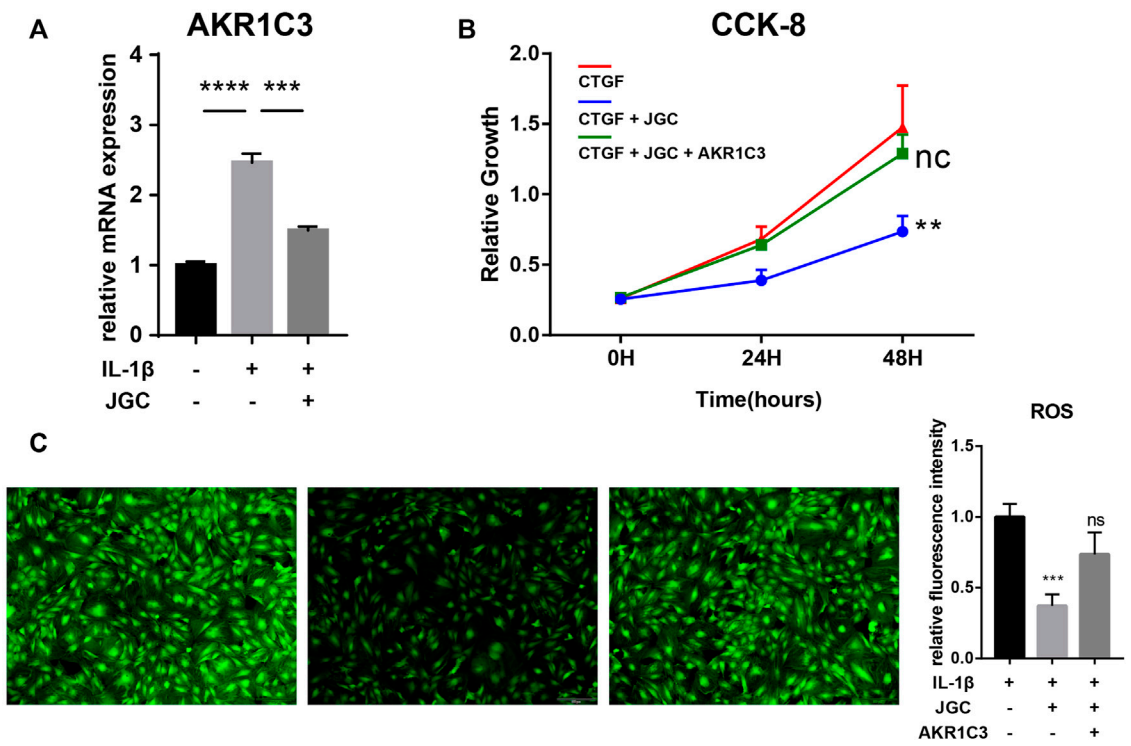


FIGURE 6
Rescue experiments of AKR1C3. **(A)** PCR showing that AKR1C3 is inhibited by JGC. **(B)** CCK-8 assay showing that AKR1C3 overexpression attenuates the JGC-mediated inhibition of cell proliferation. **(C)** ROS staining showing that AKR1C3 overexpression counteracts the anti-inflammatory effect of JGC.

increased in cell fate 2 (Yang et al., 2018; Wang and Zhao, 2020; Wei et al., 2021). Joint pain is the predominant symptom of OA. "Eicosanoid Signaling" and "Prostanoid Signaling" are thought to be the main contributors to OA pain (Sanchez-Lopez et al., 2022). Several enzymes of the eicosanoid receptors are well-recognized targets of anti-inflammatory drugs that can reduce synovial inflammation (Korotkova and Jakobsson, 2014). Interestingly, our study found that cells in fate 2 were more active in both pathways. This finding indicated that as synovial cells progress toward fate 2, the patient's pain symptoms will likely become more severe. Overall, the consistency of our results with those from previous studies bolsters the reliability of our findings on cell fate determination.

We found that quercetin, an active component of JGC, well matched the active pocket of AKR1C3, and a PCR analysis confirmed a regulatory relationship. The steroidogenic enzyme AKR1C3 plays an important role in many diseases, such as prostaglandin disorder, metastatic breast tumors and atopic dermatitis (Mantel et al., 2012; Evans et al., 2019; Li et al., 2020). AKR1C3 mediates hyperproliferation, oxidative stress and drug resistance in various tissues (González-Muniesa et al., 2013; Yepuru et al., 2013; Thoma, 2015). Although AKR1C3 is a promising therapeutic target, no AKR1C3-targeting drugs have been approved for clinical use to date (Pippione et al., 2017). As a natural product, quercetin has been extensively evaluated for its efficacy and pharmacological safety (Hu et al., 2017; Ulusoy and Sanlier, 2020; Lai and Wong, 2021; Yan et al., 2022). Our study verifies the therapeutic effect of quercetin on OA synovitis by targeting AKR1C3, which further broadens the potential application of quercetin.

This study has some limitations. There were relatively few synovitis scRNA-seq samples and a lack of corresponding chondrocytes and subchondral bone samples. Analysis of additional samples would be conducive to eliminating the heterogeneity caused by individual differences. Simultaneous analysis of data from multiple tissues (synovium, cartilage, subchondral bone) is beneficial to deepen our understanding of OA disease process.

In summary, our study confirms the beneficial influence of JGC in OA synovitis and thus shows that JGC effectively suppresses inflammation and hyperproliferation in synovial cells. An in-depth profiling of the synovitis microenvironment and transcriptional dynamics revealed two distinct cell fates that resolve or advance the disease. We also identified the pharmacological mechanism of the quercetin-AKR1C3 pair of

JGC in the treatment of OA synovitis. These efforts will help researchers better elucidate OA synovitis and improve treatment outcomes.

Data availability statement

The raw ordinary scRNA-seq data for synovitis can be accessed from GEO (GSE176308). The software programs and packages used to analyze the dataset are freely available. Further inquiries can be directed to the corresponding authors.

Author contributions

Conception and design: KT and TL. Development of methodology: JG1 and ZS. Analysis and interpretation of the data: JG1, HT and JG3. Statistical analysis: JG3, CT, XY and ZS. Drafting of the manuscript: JG1, JG3, PH and CT. Critical revision of the manuscript: KT and JG1. Obtained funding: KT and TL. All the authors read and approved the final manuscript.

Funding

This research was supported by grants from the National Natural Science Foundation of China (nos. 82072516, 82130071, and 82102635), the Sports Injury Repair Research and Innovation Group (csts2020jcyj-cxttX0004) and the Personalization Training Program for the Training Object of the Outstanding Talents of Army Medical University (4139Z2C2).

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphys.2022.1031996/full#supplementary-material

References

Atukorala, I., Kwoh, C. K., Guermazi, A., Roemer, F. W., Boudreau, R. M., Hannon, M. J., et al. (2016). Synovitis in knee osteoarthritis: a precursor of disease? *Ann. Rheum. Dis.* 75, 390–395. doi:10.1136/annrheumdis-2014-205894

Blaney Davidson, E. N., Vitters, E. L., van Den Berg, W. B., and van der Kraan, P. M. (2006). TGF beta-induced cartilage repair is maintained but fibrosis is blocked in the presence of Smad7. *Arthritis Res. Ther.* 8, R65. doi:10.1186/ar1931

Boer, C. G., Hatzikotoulas, K., Southam, L., Stefánsdóttir, L., Zhang, Y., Coutinho de Almeida, R., et al. (2021). Deciphering osteoarthritis genetics across 826, 690 individuals from 9 populations. *Cell* 184, 6003–6005. doi:10.1016/j.cell.2021.11.003

Brown, W. E., Huang, B. J., Hu, J. C., and Athanasiou, K. A. (2021). Engineering large, anatomically shaped osteochondral constructs with robust interfacial shear properties. *NPJ Regen. Med.* 6, 42. doi:10.1038/s41536-021-00152-0

Caire, R., Audoux, E., Courbon, G., Michaud, E., Petit, C., Dalix, E., et al. (2021). YAP/TAZ: Key players for rheumatoid arthritis severity by driving fibroblast like synoviocytes phenotype and fibro-inflammatory response. *Front. Immunol.* 12, 791907. doi:10.3389/fimmu.2021.791907

Cheng, H., Feng, S., Jia, X., Li, Q., Zhou, Y., and Ding, C. (2013). Structural characterization and antioxidant activities of polysaccharides extracted from Epimedium acuminatum. *Carbohydr. Polym.* 92, 63–68. doi:10.1016/j.carbpol.2012.09.051

Cheng, X. C., Guo, X. R., Qin, Z., Wang, X. D., Liu, H. M., and Liu, Y. L. (2020). Structural features and antioxidant activities of Chinese quince (Chaenomeles sinensis) fruits lignin during auto-catalyzed ethanol organosolv pretreatment. *Int. J. Biol. Macromol.* 164, 4348–4358. doi:10.1016/j.ijbiomac.2020.08.249

Choi, W. S., Lee, G., Song, W. H., Koh, J. T., Yang, J., Kwak, J. S., et al. (2019). The CH25H-CYP7B1-RORα axis of cholesterol metabolism regulates osteoarthritis. *Nature* 566, 254–258. doi:10.1038/s41586-019-0920-1

Conaghan, P. G., Cook, A. D., Hamilton, J. A., and Tak, P. P. (2019). Therapeutic options for targeting inflammatory osteoarthritis pain. *Nat. Rev. Rheumatol.* 15, 355–363. doi:10.1038/s41584-019-0221-y

Daina, A., Michielin, O., and Zoete, V. (2019). SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. *Nucleic Acids Res.* 47, W357–W364. doi:10.1093/nar/gkz382

de Munter, W., van Den Bosch, M. H., Slöetjes, A. W., Croce, K. J., Vogl, T., Roth, J., et al. (2016). High LDL levels lead to increased synovial inflammation and accelerated ectopic bone formation during experimental osteoarthritis. *Osteoarthr. Cartil.* 24, 844–855. doi:10.1016/j.joca.2015.11.016

Devi, K. P., Malar, D. S., Nabavi, S. F., Sureda, A., Xiao, J., Nabavi, S. M., et al. (2015). Kaempferol and inflammation: From chemistry to medicine. *Pharmacol. Res.* 99, 1–10. doi:10.1016/j.phrs.2015.05.002

Dong, J., Liang, W., Wang, T., Sui, J., Wang, J., Deng, Z., et al. (2019). Saponins regulate intestinal inflammation in colon cancer and IBD. *Pharmacol. Res.* 144, 66–72. doi:10.1016/j.phrs.2019.04.010

Evans, K., Duan, J., Pritchard, T., Jones, C. D., Mcdermott, L., Gu, Z., et al. (2019). OBI-3424, a novel AKR1C3-activated prodrug, exhibits potent efficacy against preclinical models of T-ALL. *Clin. Cancer Res.* 25, 4493–4503. doi:10.1158/1078-0432.CCR-19-0551

González-Muniesa, P., Marrades, M. P., Martínez, J. A., and Moreno-Aliaga, M. J. (2013). Differential proinflammatory and oxidative stress response and vulnerability to metabolic syndrome in habitual high-fat young male consumers

putatively predisposed by their genetic background. *Int. J. Mol. Sci.* 14, 17238–17255. doi:10.3390/ijms140917238

Hiligsmann, M., Cooper, C., Arden, N., Boers, M., Branco, J. C., Luisa Brandi, M., et al. (2013). Health economics in the field of osteoarthritis: an expert's consensus paper from the European society for clinical and economic aspects of osteoporosis and osteoarthritis (ESCEO). *Semin. Arthritis Rheum.* 43, 303–313. doi:10.1016/j.semarthrit.2013.07.003

Hu, K., Miao, L., Goodwin, T. J., Li, J., Liu, Q., and Huang, L. (2017). Quercetin remodels the tumor microenvironment to improve the permeation, retention, and antitumor effects of nanoparticles. *ACS Nano* 11, 4916–4925. doi:10.1021/acsnano.7b01522

Jeon, O. H., Kim, C., Laberge, R. M., Demaria, M., Rathod, S., Vasserot, A. P., et al. (2017). Local clearance of senescent cells attenuates the development of post-traumatic osteoarthritis and creates a pro-regenerative environment. *Nat. Med.* 23, 775–781. doi:10.1038/nm.4324

Jeon, Y. D., Lee, J. H., Lee, Y. M., and Kim, D. K. (2020). Puerarin inhibits inflammation and oxidative stress in dextran sulfate sodium-induced colitis mice model. *Biomed. Pharmacother.* 124, 109847. doi:10.1016/j.biopha.2020.109847

Ji, Q., Zheng, Y., Zhang, G., Hu, Y., Fan, X., Hou, Y., et al. (2019). Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. *Ann. Rheum. Dis.* 78, 100–110. doi:10.1136/annrheumdis-2017-212863

Jin, C., Frayssinet, P., Pelker, R., Cwirka, D., Hu, B., Vignery, A., et al. (2011). NLRP3 inflammasome plays a critical role in the pathogenesis of hydroxyapatite-associated arthropathy. *Proc. Natl. Acad. Sci. U. S. A.* 108, 14867–14872. doi:10.1073/pnas.1111101108

Kan, T., Zhang, S., Zhou, S., Zhang, Y., Zhao, Y., Gao, Y., et al. (2022). Single-cell RNA-seq recognized the initiator of epithelial ovarian cancer recurrence. *Oncogene* 41, 895–906. doi:10.1038/s41388-021-02139-z

Karuppagounder, V., Pinamont, W., Yoshioka, N., Elbarbary, R., and Kamal, F. (2022). Early gβγ-GRK2 inhibition ameliorates osteoarthritis development by simultaneous anti-inflammatory and chondroprotective effects. *Int. J. Mol. Sci.* 23, 7933. doi:10.3390/ijms23147933

Knights, A. J., Farrell, E. C., Ellis, O. M., Lammlin, L., Junginger, L. M., Rzeczycki, P. M., et al. (2022). Synovial fibroblasts assume distinct functional identities and secrete R-spondin 2 in osteoarthritis. *Ann. Rheum. Dis.*, 2022-222773. doi:10.1136/ard-2022-222773

Kong, J. S., Park, J. H., Yoo, S. A., Kim, K. M., Bae, Y. J., Park, Y. J., et al. (2020). Dynamic transcriptome analysis unveils key proresolving factors of chronic inflammatory arthritis. *J. Clin. Invest.* 130, 3974–3986. doi:10.1172/JCI126866

Korotkova, M., and Jakobsson, P. J. (2014). Persisting eicosanoid pathways in rheumatic diseases. *Nat. Rev. Rheumatol.* 10, 229–241. doi:10.1038/nrrheum.2014.1

Kuang, L., Wu, J., Su, N., Qi, H., Chen, H., Zhou, S., et al. (2020). FGFR3 deficiency enhances CXCL12-dependent chemotaxis of macrophages via upregulating CXCR7 and aggravates joint destruction in mice. *Ann. Rheum. Dis.* 79, 112–122. doi:10.1136/annrheumdis-2019-215696

Labinsky, H., Panipinto, P. M., Ly, K. A., Khuat, D. K., Madarampalli, B., Mahajan, V., et al. (2020). Multiparameter analysis identifies heterogeneity in knee osteoarthritis synovial responses. *Arthritis Rheumatol.* 72, 598–608. doi:10.1002/art.41161

Lai, W. F., and Wong, W. T. (2021). Design and optimization of quercetin-based functional foods. *Crit. Rev. Food Sci. Nutr.* 62, 7319–7335. doi:10.1080/10408398.2021.1913569

Li, Z. Y., Yin, Y. F., Guo, Y., Li, H., Xu, M. Q., Liu, M., et al. (2020). Enhancing anti-tumor activity of sorafenib mesoporous silica nanomatrix in metastatic breast tumor and hepatocellular carcinoma via the Co-administration with flufenamic acid. *Int. J. Nanomed.* 15, 1809–1821. doi:10.2147/IJN.S240436

Lin, L. W., Tsai, F. H., Lan, W. C., Cheng, Y. D., Lee, S. C., and Wu, C. R. (2019). Steroid-enriched fraction of Achyranthes bidentata protects amyloid β peptide 1-40-induced cognitive dysfunction and neuroinflammation in rats. *Mol. Neurobiol.* 56, 5671–5688. doi:10.1007/s12035-018-1436-7

Lin, T. F., Wang, L., Zhang, Y., Zhang, J. H., Zhou, D. Y., Fang, F., et al. (2021). Uses, chemical compositions, pharmacological activities and toxicology of Clematidis Radix et Rhizome- a Review. *J. Ethnopharmacol.* 270, 113831. doi:10.1016/j.jep.2021.113831

Macdonald, I. J., Huang, C. C., Liu, S. C., Lin, Y. Y., and Tang, C. H. (2021). Targeting CCN proteins in rheumatoid arthritis and osteoarthritis. *Int. J. Mol. Sci.* 22, 4340. doi:10.3390/ijms22094340

Mantel, A., Carpenter-Mendini, A. B., Vanbuskirk, J. B., de Benedetto, A., Beck, L. A., and Pentland, A. P. (2012). Aldo-keto reductase 1C3 is expressed in differentiated human epidermis, affects keratinocyte differentiation, and is upregulated in atopic dermatitis. *J. Invest. Dermatol.* 132, 1103–1110. doi:10.1038/jid.2011.412

Nanus, D. E., Badoume, A., Wijesinghe, S. N., Halsey, A. M., Hurley, P., Ahmed, Z., et al. (2021). Synovial tissue from sites of joint pain in knee osteoarthritis patients exhibits a differential phenotype with distinct fibroblast subsets. *EBioMedicine* 72, 103618. doi:10.1016/j.ebiom.2021.103618

Pippione, A. C., Giraudo, A., Bonanni, D., Carnovale, I. M., Marini, E., Cena, C., et al. (2017). Hydroxytriazole derivatives as potent and selective aldo-keto reductase 1C3 (AKR1C3) inhibitors discovered by bioisosteric scaffold hopping approach. *Eur. J. Med. Chem.* 139, 936–946. doi:10.1016/j.ejmech.2017.08.046

Pontes-Quero, G. M., Benito-Garzón, L., PéREZ Cano, J., Aguilar, M. R., and Vázquez-Lasa, B. (2021). Modulation of inflammatory mediators by polymeric nanoparticles loaded with anti-inflammatory drugs. *Pharmaceutics* 13, 290. doi:10.3390/pharmaceutics13020290

Razavi, B. M., Ghasemzadeh Rahbardar, M., and Hosseinzadeh, H. (2021). A review of therapeutic potentials of turmeric (Curcuma longa) and its active constituent, curcumin, on inflammatory disorders, pain, and their related patents. *Phytother. Res.* 35, 6489–6513. doi:10.1002/ptr.7224

Sanchez-Lopez, E., Coras, R., Torres, A., Lane, N. E., and Guma, M. (2022). Synovial inflammation in osteoarthritis progression. *Nat. Rev. Rheumatol.* 18, 258–275. doi:10.1038/s41584-022-00749-9

Schnieder, J., Mamazhakypov, A., Birnhuber, A., Wilhelm, J., Kwapiszewska, G., Ruppert, C., et al. (2020). Loss of LRP1 promotes acquisition of contractile-myofibroblast phenotype and release of active TGF-β1 from ECM stores. *Matrix Biol.* 88, 69–88. doi:10.1016/j.matbio.2019.12.001

Sellam, J., and Berenbaum, F. (2010). The role of synovitis in pathophysiology and clinical symptoms of osteoarthritis. *Nat. Rev. Rheumatol.* 6, 625–635. doi:10.1038/nrrheum.2010.159

Sridhar, A., Ponnuchamy, M., Kumar, P. S., Kapoor, A., Vo, D. N., and Prabhakar, S. (2021). Techniques and modeling of polyphenol extraction from food: a review. *Environ. Chem. Lett.* 19, 3409–3443. doi:10.1007/s10311-021-01217-8

Symons, R. A., Colella, F., Collins, F. L., Rafipay, A. J., Kania, K., Mcclure, J. J., et al. (2022). Targeting the IL-6-Yap-Snail signalling axis in synovial fibroblasts ameliorates inflammatory arthritis. *Ann. Rheum. Dis.* 81, 214–224. doi:10.1136/annrheumdis-2021-220875

Thoma, C. (2015). Prostate cancer: Breaking AKR1C3-mediated enzalutamide resistance by inhibiting androgen synthesis. *Nat. Rev. Urol.* 12, 124. doi:10.1038/nrurol.2015.31

Ulusoy, H. G., and Sanlier, N. (2020). A minireview of quercetin: from its metabolism to possible mechanisms of its biological activities. *Crit. Rev. Food Sci. Nutr.* 60, 3290–3303. doi:10.1080/10408398.2019.1683810

Wang, J., and Zhao, Q. (2020). Linc02381 exacerbates rheumatoid arthritis through adsorbing miR-590-5p and activating the mitogen-activated protein kinase signaling pathway in rheumatoid arthritis-fibroblast-like synoviocytes. *Cell Transpl.* 29, 963689720938023. doi:10.1177/0963689720938023

Wei, Q., Kong, N., Liu, X., Tian, R., Jiao, M., Li, Y., et al. (2021). Pirfenidone attenuates synovial fibrosis and postpones the progression of osteoarthritis by anti-fibrotic and anti-inflammatory properties *in vivo* and *in vitro*. *J. Transl. Med.* 19, 157. doi:10.1186/s12967-021-02823-4

Xie, J., Wang, Y., Lu, L., Liu, L., Yu, X., and Pei, F. (2021). Cellular senescence in knee osteoarthritis: molecular mechanisms and therapeutic implications. *Ageing Res. Rev.* 70, 101413. doi:10.1016/j.arr.2021.101413

Yan, L., Vaghari-Tabari, M., Malakoti, F., Moein, S., Qujeq, D., Yousefi, B., et al. (2022). Quercetin: an effective polyphenol in alleviating diabetes and diabetic complications. *Crit. Rev. Food Sci. Nutr.*, 1–24. doi:10.1080/10408398.2022.2067825

Yang, S., Xing, Z., Liu, T., Zhou, J., Liang, Q., Tang, T., et al. (2018). Synovial tissue quantitative proteomics analysis reveals paeoniflorin decreases LIFR and ASPN proteins in experimental rheumatoid arthritis. *Drug Des. Devel. Ther.* 12, 463–473. doi:10.2147/DDDT.S153927

Yepuru, M., Wu, Z., Kulkarni, A., Yin, F., Barrett, C. M., Kim, J., et al. (2013). Steroidogenic enzyme AKR1C3 is a novel androgen receptor-selective coactivator that promotes prostate cancer growth. *Clin. Cancer Res.* 19, 5613–5625. doi:10.1158/1078-0432.CCR-13-1151

Zhang, M., Hu, W., Cai, C., Wu, Y., Li, J., and Dong, S. (2022). Advanced application of stimuli-responsive drug delivery system for inflammatory arthritis treatment. *Mater. Today. Bio* 14, 100223. doi:10.1016/j.mtbio.2022.100223

Zhao, J., Yang, W., Liang, G., Luo, M., Pan, J., Liu, J., et al. (2022). The efficacy and safety of Jinwu Gutong capsule in the treatment of knee osteoarthritis: A meta-analysis of randomized controlled trials. *J. Ethnopharmacol.* 293, 115247. doi:10.1016/j.jep.2022.115247

Zhou, G., Sun, X., Qin, Q., Lv, J., Cai, Y., Wang, M., et al. (2018). Loss of Smad7 promotes inflammation in rheumatoid arthritis. *Front. Immunol.* 9, 2537. doi:10.3389/fimmu.2018.02537

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 10 January 2023
DOI 10.3389/fmed.2022.1090120

OPEN ACCESS

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

REVIEWED BY
Rabbani Syed,
King Saud University, Saudi Arabia
V. P. Snijesh,
St. John's Research Institute, India
Md Zubbair Malik,
Jawaharlal Nehru University, India

*CORRESPONDENCE
Thoraia Shinawi
✉ tshinawai@kau.edu.sa

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

RECEIVED 04 November 2022
ACCEPTED 21 December 2022
PUBLISHED 10 January 2023

CITATION
Nasser KK and Shinawi T (2023) Genotype-protein phenotype characterization of *NOD2* and *IL23R* missense variants associated with inflammatory bowel disease: A paradigm from molecular modelling, dynamics, and docking simulations.
Front. Med. 9:1090120.
doi: 10.3389/fmed.2022.1090120

Genotype-protein phenotype characterization of *NOD2* and *IL23R* missense variants associated with inflammatory bowel disease: A paradigm from molecular modelling, dynamics, and docking simulations

Khalidah Khalid Nasser[1,2,3] and Thoraia Shinawi[1]*

[1]Department of Medical Laboratory Technology, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, [2]Princess Al-Jawhara Al-Brahim Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia, [3]Centre for Artificial Intelligence in Precision Medicine, King Abdulaziz University, Jeddah, Saudi Arabia

Inflammatory bowel disease (IBD) is a gastrointestinal disease with an underlying contribution of genetic, microbial, environment, immunity factors. The coding region risk markers identified by IBD genome wide association studies have not been well characterized at protein phenotype level. Therefore, this study is conducted to characterize the role of *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln) missense variants on the structural and functional features of corresponding proteins. Thus, we used different variant pathogenicity assays, molecular modelling, secondary structure, stability, molecular dynamics, and molecular docking analysis methods. Our findings suggest that SIFT, Polyphen, GREP++, PhyloP, SiPhy and REVEL methods are very sensitive in determining pathogenicity of *NOD2* and *IL23R* missense variants. We have also noticed that all the tested missense variants could potentially alter secondary (α-helices, β-strands, and coils) and tertiary (residue level deviations) structural features. Moreover, our molecular dynamics (MD) simulation findings have simulated that *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln) variants creates rigid local structures comprising the protein flexibility and conformations. These predictions are corroborated by molecular docking results, where we noticed that *NOD2* and *IL23R* missense variants induce molecular interaction deformities with *RIPK2* and *JAK2* ligand molecules, respectively. These functional alterations could potentially alter the signal transduction pathway cascade involved in inflammation and autoimmunity. Drug library searches and findings from docking studies have identified the inhibitory effects of Tacrolimus and Celecoxib drugs on *NOD2* and *IL23R* variant forms, underlining their potential to contribute to personalized

medicine for IBD. The present study supports the utilization of computational methods as primary filters (pre-*in vitro* and *in vivo*) in studying the disease potential mutations in the context of genptype-protein phenotype characteristics.

KEYWORDS

inflammatory diseases, IBD, genetic variants, molecular docking, protein stability, 3D modelling, MD simulation

1. Introduction

Inflammatory bowel disease (IBD) is chronic autoimmune condition of the digestive tract (GIT) (1). Ulcerative Colitis (UC) and Crohn's disease (CD) constitute the two main clinical forms of IBD. The specific molecular etiology of IBD is yet to be fully understood, but numerous studies show that aberrant interactions between various genetic, immunologic (e.g., mucosal immune cells) and environmental (e.g., gut microbiota) factors play a pivotal role in IBD pathogenesis (1, 2). The genetic basis of IBD is well supported by findings such as increased disease rates in monozygotic twins, and also by disease susceptibility differences among ethnic groups (3). Population genetics investigations have also revealed compelling evidence about the critical role of genetic factors in the etiopathogenesis of IBD. In recent years, the International IBD Genetics Consortium (IIBDGC), has pooled up all the GWAS findings and identified a total of 201 IBD susceptibility loci (4, 5). Among these loci, *NOD2* and *IL23R* still represent the strongest predictors for IBD susceptibility and clinical phenotypes (6–8).

NOD2 (Nucleotide Binding Oligomerization Domain 2) is an intracellular receptor belonging to the family of cytosolic NLRs (NOD, leucine-rich repeat protein) involved in immune response by recognizing the muramyl dipeptide (MDP) component of the bacterial cell wall. *NOD2* variants like Arg70Trp, Gly908Arg, Arg702Trp and Leu1007PfsX2*NOD2* are strongly implicated in Crohn's disease (CD) in Caucasian population (9–12). The *IL23R* gene encodes a transmembrane protein molecule belonging to type I cytokine receptor (13). This molecule initially pairs with *IL12RB1* to bind the *IL23* signaling molecule and activates JAK- STAT and NF-κB signaling pathways. This receptor is highly expressed in dendritic cells and is shown to be involved in controlling infection and chronic autoimmune diseases (14). The polymorphisms in the *IL23R* gene are also known to modulate IL23 responses and have also been reported to influence the risk of IBD development (15, 16).

Although, positive statistical associations of *NOD2* and *IL23R* genes with IBD is well known, the specific mechanisms how these genetic variants contribute to clinical phenotypes is not yet clear. It is reasonable to assume that the disease related amino acid substitution mutations cause changes in the chemical nature or position of the encoded amino acid variant, and potentially influences the bio physical characteristics (like hydrogen bonding, pH dependence and conformational dynamics) of the proteins. Although, both *in vivo* and *in vitro* studies are effective solution in this direction, but they consume lot of time and require a series of laboratory investigations. The alternate strategy for overcoming this difficulty is by predicting the specific biophysical impacts of each mutation through advanced integrated bioinformatics approaches. So many computations programs like SIFT (17), Polyphen (18), M-CAP (19), FATHMM (20), CADD (21) etc., each specializing on different prediction principles, are now available for exploring the relationship between genetic mutations and human diseases. Numerous studies have utilized these programs to screen clinically significant genetic variants in different human diseases (22–26). Therefore, in the present study, we have performed a comprehensive computational analysis of *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln) variants using diverse range of machine learning approaches. The genetic sequence – protein structure relationships were studied different structural parameters like secondary structure, active sites, motifs, domains, and accessible surface areas in both wild type and mutant proteins.

Disease management strategy for IBD patients involves surgery or drug treatment, depending upon the clinical conditions and progression of inflammation (27). IBD treatment regime consists of drugs belonging to five major categories like anti-inflammatory steroids, immunosuppressive, symptomatic relief drugs, antibiotics, and biological agents. The long-term serious side effects and toxicity induction by these steroidal and non-steroidal drugs in IBD patients is seen to be unavoidable. However, this problem can be effectively minimized by screening drugs which have the potential to inhibit mutated target proteins and reduce the drug associated cellular toxicity (28). Our drug library searching revealed us that Tacrolimus and Celecoxib drugs shows specific inhibitory action on mutated forms of *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln), respectively. Hence, our study provides computational evidence to repurpose Tacrolimus and Celecoxib drugs against IBD pathogenesis after conducting comprehensive *in vitro* and *in vivo* experiments.

2. Materials and methods

2.1. Variant data

The details of NOD2 and *IL23R* genes including mRNA accession number, reference number and their concerned protein sequences were retrieved from UniProt, Human Gene Mutation Database (HGMD), ClinVar, 1,000 genomes, Ensemble (and the Single Nucleotide Polymorphism Database (dbSNP). The terms like genetic mutations, genetic variations, and SNPs are used interchangeably throughout this manuscript.

2.2. Prediction and functional annotation of variants

dbNSFP version 2.2 was used for the functional predictions and annotations of NOD2 and *IL23R* missense mutations. The dbNSFP is a comprehensive database for functional predictions and annotations of all the potential human non-synonymous single-nucleotide variants (nsSNVs) (29, 30). The current version (dbNSFP v2.2) of the database is based on the GENCODE 9/Ensemble version 64 and it includes a total of 87,361,054 nsSNVs. The search for the nsSNVs from the database is done using a java program that executes the query in dbNSFP v2.2 on the local machine of the user. For each query it produces two prediction scores and three conservation scores along with other variant and gene annotations. In this study, we produced the prediction data for *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln) genetic variants using six different algorithms e.g., SIFT, PolyPhen-2, GERP++, PhyloP, SiPhy and REVEL.

2.3. Structure analysis of mutations

2.3.1. 3D modeling, secondary structure, and solvent accessibility methods

The structural and functional consequences of any variant can be better understood, by studying them at 3D level. Therefore, we analyzed the 3D model of selected *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln) variants. The Protein Databank (PDB) does not have experimentally solved structures for *NOD2* and *IL23R*, so, we resorted to homology and/or *ab initio* based computer modeling. In this study, we used different homology modeling tools like Modeller,[1] Swiss Model,[2] etc., Another important computational approach used to build a protein model is, *ab initio* modeling. When an identical structure is unavailable or the target sequence has <30% identity, this approach is utilized. The I-Tasser[3] used in the *ab initio* studies relies on the basic principle of multiple-threading alignments by LOMETS and iterative template fragment assembly simulations. The energy minimization of built protein models was done by applying the force-field of steepest descent using SPDV tool.[4] This energy minimization step was carried out to remove the wicked contacts in a simulated protein structure. After the energy minimization step, built protein's structural quality was assessed by Procheck[5] tools.

The secondary structure analysis (such as helices, loops, sheets, etc.) of built models was carried out using the PDBSUM server.[6] The active site analysis were carried out using CastP[7] tool, this tools provide information about the active cavities, conserved amino acids and substrate binding sites present in the protein structure. Electrostatic, superpose, and solvent accessibility analysis were carried out using Pymol, Yasara,[8] and SAS tools.[9] The SAS analysis provides information about exposed and buried residues present in a protein, which is very crucial for comparing wild type and mutated protein models. In order to check the domains in the protein sequence, we submitted our sequence to the SangerPfam web server,[10] which directly searches the protein sequences by expanding typical search methodology with a Pfam wrapper around the HMMER pack. The default E-value threshold used in the HMM search process was 1.0.

2.3.2. Molecular dynamics (MD) simulations

The structural analysis of the *NOD2* and *IL23R* proteins was performed to evaluate the stability of wild type and variant proteins using Gromacs 4.0 and Molecular Operating Environment (MOE) softwares. The energy minimization for initial structures was performed using the steepest descent algorithm in the Gromacs 3.3 software package at a maximum of 2,000 ps time, at 300K temperature. After energy minimizing the wild type and mutated proteins, we applied restraint at 100 ps to allow solvent equilibration (NVT, NPT) around the protein. Finally full MDS was performed on all structures (wild-type and mutant models) at 20,000 ps, separately embedded in a box (box volume > 756.12 nm^3), containing pre-equilibrated water molecules. The van der Waals interaction and particle Mesh Ewald (PME) for long range electrostatic interactions was set to >10 Å. The space between the edge of the box and protein was set at >10 Å. Episodic frontier environments

1 https://salilab.org/modeller/
2 https://swissmodel.expasy.org/
3 https://zhanggroup.org/I-TASSER/
4 https://spdbv.unil.ch/
5 https://saves.mbi.ucla.edu/
6 http://www.ebi.ac.uk/pdbsum
7 http://sts.bioe.uic.edu/castp/index.html
8 www.yasara.org
9 www.abren.net/asaview
10 www.ebi.ac.uk/interpro/

were smeared in all ways. Charged ions were positioned to exchange water molecules in alternate positions, thus building the entire neutral system. The lengths of hydrogen-atom bonds were constrained using the LINCS parameters technique, at a 0.002 ps time step. For every 1 ps, the structures from the dynamic trajectory were saved. The xmgrace analysis package in GROMACS software, was used to perform all the post-dynamic studies of the trajectories (31).

2.4. Genetic interaction networks analysis

The protein association partners of *NOD2* and *IL23R* were studied using GeneMania tool.[11] These databases provide data about protein association based on multiple categories of information, including physical co-occurrences, genomic neighborhood, conserved co-expression, and gene fusion, and these studies are limited to experimentally validated interactions. The input format consists of providing the query gene list. The output is a network of functional relationships for query gene and predicted related genes in the form of nodes and edges. Nodes represent genes and links represent networks. Genes can be linked by more than one type of network.

2.5. Protein-drug interaction analysis by molecular drug docking

At first, the potential therapeutic molecules showing an cut-off interaction score of >0.03 against *NOD2* and *IL23R* genes were identified in Drug–Gene Interaction database (DGIdb) (32). Then molecular docking analysis was performed to elucidate the functional interaction deformities of wild and mutant proteins with the query drugs. AutoDock 4.0, which is based on the Lamarckian Genetic Algorithm, is used to run docking queries for drugs and target proteins. During the docking process, the torsion angles of flexible ligands were identified by 10 independent runs. The protein structures were initially neutralized by removing ions and charges (on histidine), before applying gigaster charges to them. The grid maps were constructed around protein-ligand molecules using Autogrid module of Auto dock software program. The default parameters used in constructing the grid were 60, 60, 60 points in x, y, and z directions, a center spacing of a grid is 0.367A° (approx. 1/4 of the length of c–c covalent bond). Then, the docking parameter file was prepared with AutoDockTools (ADT). When LGA was set to 150 runs, the other default parameters were 150 conformations, population size is 50, and energy evaluations is 25,00,000. For docking parameters, the initial translation was set to 0.2A Å; the torsion to 0.5^o, the quaternion to 5.0^o; and the RMS cluster tolerance to 0.75 Å. The ligands that showed the most promising binding energy were chosen from the protein-ligand docking complex at the end of the docking process. Pymol-0.98 was used to analyze the resulting docking complexes.

11 https://genemania.org/

3. Results

3.1. Pathogenic characterization of IBD variants

The SIFT and PolyPhen-2 predictions, have attributed the deleterious effect to *NOD2* (Arg675Trp and Gly908Arg) and *IL23R* (Gly149Arg and Arg381Gln). The other predictions like GERP++, PhyloP, SiPhy and REVEL scores (GERP++ RS > 0; Phylop > 0; SiPhy > 0, REVEL < 0.5) have also confirmed that these 4 SNPs affect the nucleotide sequences, which are under the high evolutionary significance (**Table 1**).

3.2. Protein structural impact analysis of IBD variants

Structural annotations Workflow of current study represented in **Supplementary Figure 1**.

3.2.1. 3D modeling

Due to unavailability of *NOD2* and *IL23R* crystal statures in Protein Databank, we performed the BLASTp search in protein databank to check the homologous proteins with 45% identity. However, we could not find any homologues protein structures in PDB at the required threshold value. Therefore, to develop *NOD2* and *IL23R* wild type protein models, we resorted to *ab initio* based modeling approach using I-Tasser web server. The resultant output was 5 protein models for *NOD2* and *IL23R*, each. The best model was selected based on its c-scores (ranging from −5 to + 2). The top *NOD2* protein model (**Figure 1A**) had a c-score of −1.23 and *IL23R* had a score of −2.2 (**Figure 1B**). Both *NOD2* and *IL23R* were cured by an energy minimization step to remove all the bad contacts in the protein structure. *NOD2*'s energy was minimized at 2,335 fs, and the released energy was −3,25,428 KJ/Mol. For *IL23R*, energy minimization was done at 3,245 fs, and it resulted in the release of −2,3545 KJ/mol of energy. These models were further evaluated for protein quality using PROCHECK software. The *NOD2* protein model revealed that 97% of residues are in the allowed region and only 3% of residues are present in the disallowed region. For *IL23R*, 96.8% of residues are in the allowed region and 3.2% of residues are in the disallowed region of the protein.

The native *NOD2* and *IL23R* protein structures were further used as templates to create mutant protein versions

TABLE 1 Pathogenicity prediction of coding region mutations using different algorithm.

rsID	ChrPos	Gene	Amino acid substitution	Functional predictions		Evolutionary conservations prediction			Concordance tool
				SIFT[1]	PolyPhen[2]	GREP++[3]	PhyloP[3]	SiPhy[3]	REVEL[4]
rs2066844	16:50712015	NOD2	Arg675Trp	0.01	0.72	5.74	0.742	6.913	0.55051
rs2066845	16:50722629	NOD2	Gly908Arg	0.01	1	5.91	0.871	6.9139	0.705
rs76418789	1:67182913	IL23R	Gly149Arg	0	0.991	5.24	−0.432	6.0229	0.653
rs11209026	1:67240275	IL23R	Arg381Gln	0.02	0.776	5.19	0.024	7.5165	0.449

[1]SIFT < 0.05 the corresponding SNP is "Damaging"; otherwise, it is predicted as "Tolerated."
[2]PolyPhen-2: > 0.5 prediction, "deleterious" and < 0.5, "neutral."
[3]GERP++, PhyloP and Siphy: the larger the score, the more conserved the site.
[4]REVEL < 0.5 is neutral, >0.5 is deleterious.

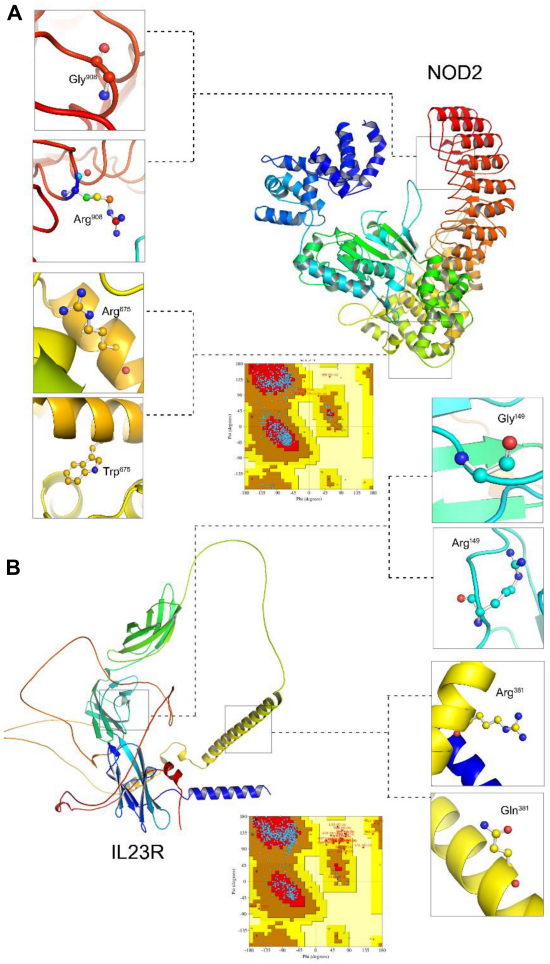


FIGURE 1
Molecular visualization of protein models (NOD2 and IL23R) and Ramachandran plots. **(A)** The NOD2 protein structure, zoom view represent localization of Gly908 (wildtype), Arg908 (mutated). **(B)** The IL23R protein structure zoom view represent localization of Gly149 (wildtype), Arg149 (mutated), Arg381 (wildtype), Gln381 (mutated).

using MODELLER9v3 and Swiss Model server software. All the 100 models (output from MODELLER9v3) generated per each mutant category, were further subjected to energy minimization followed by PROCHECK validation. The mutant model (Gly908Arg and Arg675Trp) of *NOD2* contains 95.2% residues in allowed regions and 4.8% in disallowed regions. The two mutant models (Gly149Arg, Arg381Gln) of *IL23R* consist of 94.2 and 96.8% of residues in the allowed region, whereas 5.8 and 3.2% of disallowed regions, respectively.

3.2.2. Super positioning of native and mutant models

We compared wild and mutant protein models of *NOD2* and *IL23R* to examine their structural drifts induced by amino acid substitutions. The c-alpha backbone of the root mean square

deviation (RMSD) between wild type and mutated models (Arg675Trp and Gly908Arg) of *NOD2* was found to be of 0.04 and 0.06 Å suggesting a limited potential of these mutations in inducing whole structure level alterations, respectively. However, at the amino acid residue level, these deviation was seen to be very high, i.e., 2.45 and 1.78 Å, respectively. The *IL23R* superposed on two mutated models, the C-alpha and backbone RMSDs were 0.048 and 0.052Å, suggesting limited potential of Gly149Arg and Arg381Gln mutations in inducing whole structure level alterations. Similarly, even at amino acid residue level, the deviation was minimal, that is, 1.6 and 1.48 Å (**Table 2**).

3.2.3. Secondary structural annotations of IBD variants

We sought to examine the structural and functional consequences of amino acid substitutions in *NOD2* and *IL23R* proteins through diverse approaches like secondary structure analysis, and clefts analysis.

3.2.4. Secondary structural features, clefts, and active site analysis of NOD2

Secondary structure analysis is crucial to understanding the hierarchical classification of protein structures and their polypeptide folding nature. The secondary structure of *NOD2* consists of different elements like 3 beta sheets, 12 beta-alpha-beta motifs, 2 beta hairpins, 1 beta bulge, 20 strands, 44 helices, 78 helix-helix interfaces, 68 beta turns, and 9 gamma turns. As *NOD2* is a transmembrane protein, it is made up of many helices as well as beta turns to maintain the polypeptide folding, which is important for maintaining its globular shape (**Figure 2A**).

Clefts are defined as gap regions existing in any protein molecule. The size of cleft often determines how protein interacts with their ligand molecules. Most of the active sites in proteins contain both deep and large clefts. The *NOD2* protein contains 4 clefts greater than 1,000 Å, out of which deepest and largest cleft located in between signal recognition and oligomerization regions is 12,085.03 Å in size. This large cleft is made up of 201 residues and consists of 72.13% accessible vertices and 13.77% buried vertices (**Figure 2B**).

NOD2 ligand binding site prediction using PDBSUM showed that ADP interacts with His 603, Ser306, Thr239, Gly302, Thr240, Thr253, Thr307, Gly304 and Lys305 amino acid residue of *NOD2*.

3.2.5. Secondary structural features, clefts, and active site analysis of IL23R

The *IL23R* protein consists of three regions, i.e., the C-terminal signal recognition, transmembrane and cytosolic c-terminal regions. The secondary structural features of this protein are made up of 10 sheets, 7 beta hairpins, 3 beta bulges, 37 strands, 4 helices, 80 beta turns, 40 gamma turns, and one disulfide bridge. The odd secondary structural features of *IL23R* are characterized by a low number of helices and a high ratio of turns, which further helps to maintain the stability of *IL23R* in the membrane (**Figure 2C**).

The *IL23R* contains 4 clefts that are larger than 1,000 Å in size. Out of these, the fourth cleft made up of 91 residues is the deepest and largest, is 6,021Å in size, and it contains 65.91% accessible vertices and 11.59 buried vertices (**Figure 2D**).

IL23R active site prediction using the CASTp server revealed the existence of two different active or ligand-binding sites in between extracellular and intracellular regions. In the extracellular region, the active site acid amino acid residues are as follows, Tyr100, Gln110, Asp118, Leu210, and Arg227. In the intracellular region, Phe530, Asn542, Glu570, Aln587, and Gly599 are predicted as active site residues.

3.2.6. Solvent accessible surface area analysis of IBD variants

The native Arginine at 675th position interacts is in buried condition with more than 30% surface accessible area to solvents but the variant Tryptophan is found in exposed condition and decreases the solvent accessibility. The glycine (native) amino acid at the 908th position of the *NOD2* protein is in buried position and portrays 20% surface accessible area to solvents, whereas the substitution of arginine amino acid, due to its physical conformation, portrays 80% of the surface accessible area to solvents. The *IL23R* Phe149 and Arg381 amino acid (native) residues showed 80% surface accessible regions, with only Arg381 showing a significant shift (80–100%) in its solvent accessibility ability (**Figure 3A**).

3.3. Stability predictions of IBD variants

Any amino acid substitution is likely to affect the stability of protein structures. Therefore, to understand the structural consequences of Gly908Arg of *NOD2* and Gly149Arg and Arg381Gln of *IL23R* on their protein stabilities, we assessed their free energy changes through the DUET web server. **Table 3** reveals that Gly908Arg of *NOD2* and Gly149Arg and Arg381Gln of *IL23R* mutations are destabilizing to protein stability in terms of free energy changes.

3.4. Functional domain analysis of IBD variants

The *NOD2*, Arg675Trp variant is located 68 amino acids downstream from the winged helix domain located from 545th to 597th residues, whereas the Gly908Arg variant is in leucine rich domain 4 spanning between 897th and 1,004th amino acids. The *IL23R,* Gly149 Arg is located in Fibronectin domain 1 (129–217), whereas Arg381Gln variant lies 63 residues downstream to Fibronectin domain 2 (219–318) of the protein (**Figure 3B**).

TABLE 2 RMSD values and H-bond interaction of mutant and wild type models of *NOD2* and *IL23R*.

Protein	Mutated residue	RMSD (Å)		H-Bonds
		Protein level	Residue level	
NOD2	Arg675	–	–	7 H-bods, Arg-675, Val590, Ala589, Arg678, Leu672, Ser591
	Trp675 (M)	0.04	2.453	3-Hobonds Val671, Leu672, Ala679
	Gly908	–	–	2 H-Bonds with Asn880 and Val935
	Arg908 (M)	0.06	1.78	1 H–Bond with Val935
IL23R	Gly149	–	–	–
	Arg149 (M)	0.0479	1.6	1 H-bond with Glu130
	Arg381	–	–	5 H-bonds with Ser379, Thr382, Gly383
	Gln 381 (M)	0.052	1.48	1 H-bond with Ser379

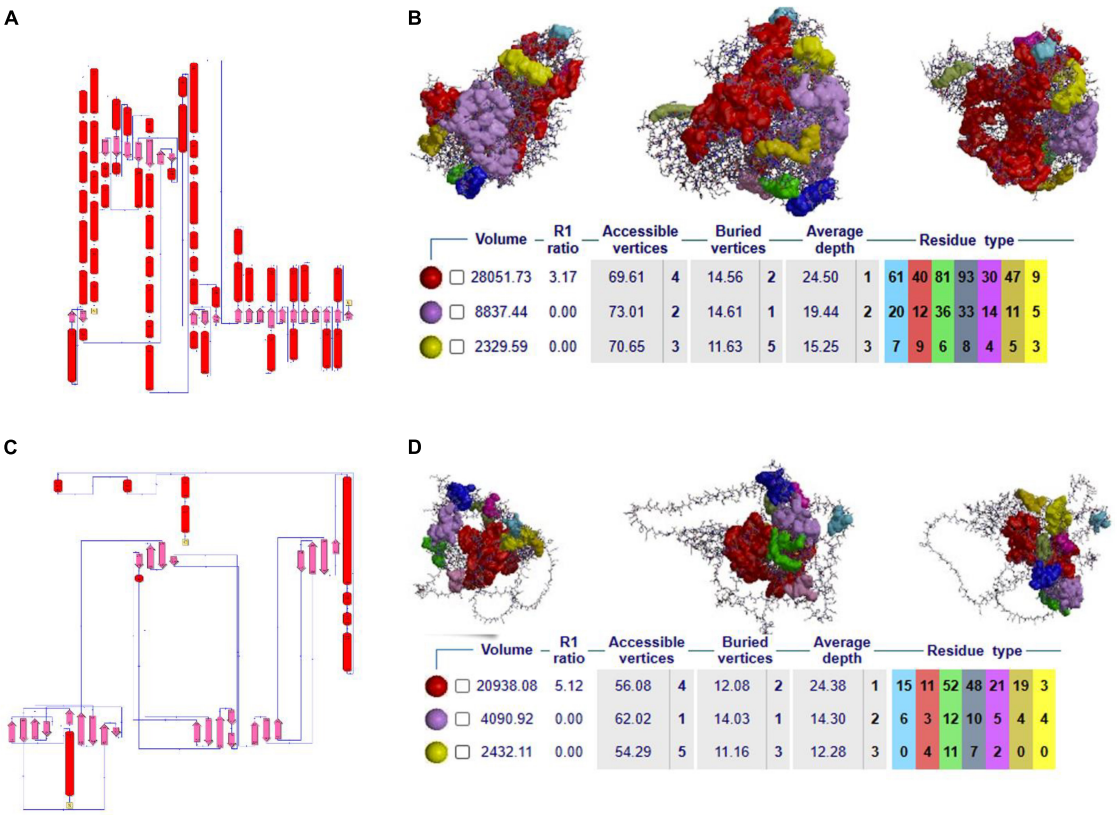


FIGURE 2
(A–C) 2D secondary structural conformation of NOD2 and IL23R. **(B–D)** The NOD2 and IL23R clefts in the structure, shown here as solid surfaces colored according to volume, with the largest shown in red.

3.5. MD simulation findings of IBD variants

The MD analysis was performed to better understand the stability of proteins in both wild and mutant states during the molecular simulation phase. We have also tried to predict physical disturbances in mutant proteins, in terms of their values corresponding to RMSD of C-alpha, radius of gyration (Rg) and solvent accessible surface area (SASA) at a 10ns solvent simulation period. The native energy minimized structures of *NOD2* and *IL23R* were used as references to compute the RMSD values of their mutant forms.

In the case of *NOD2,* the molecular stability in wild type protein was achieved at 3,000 ps (0.58 nm value) over the total 10 ns simulation test period. For Arg675Trp and Gly908Arg variants, the RMSD values increased sharply after 4,000 ps and

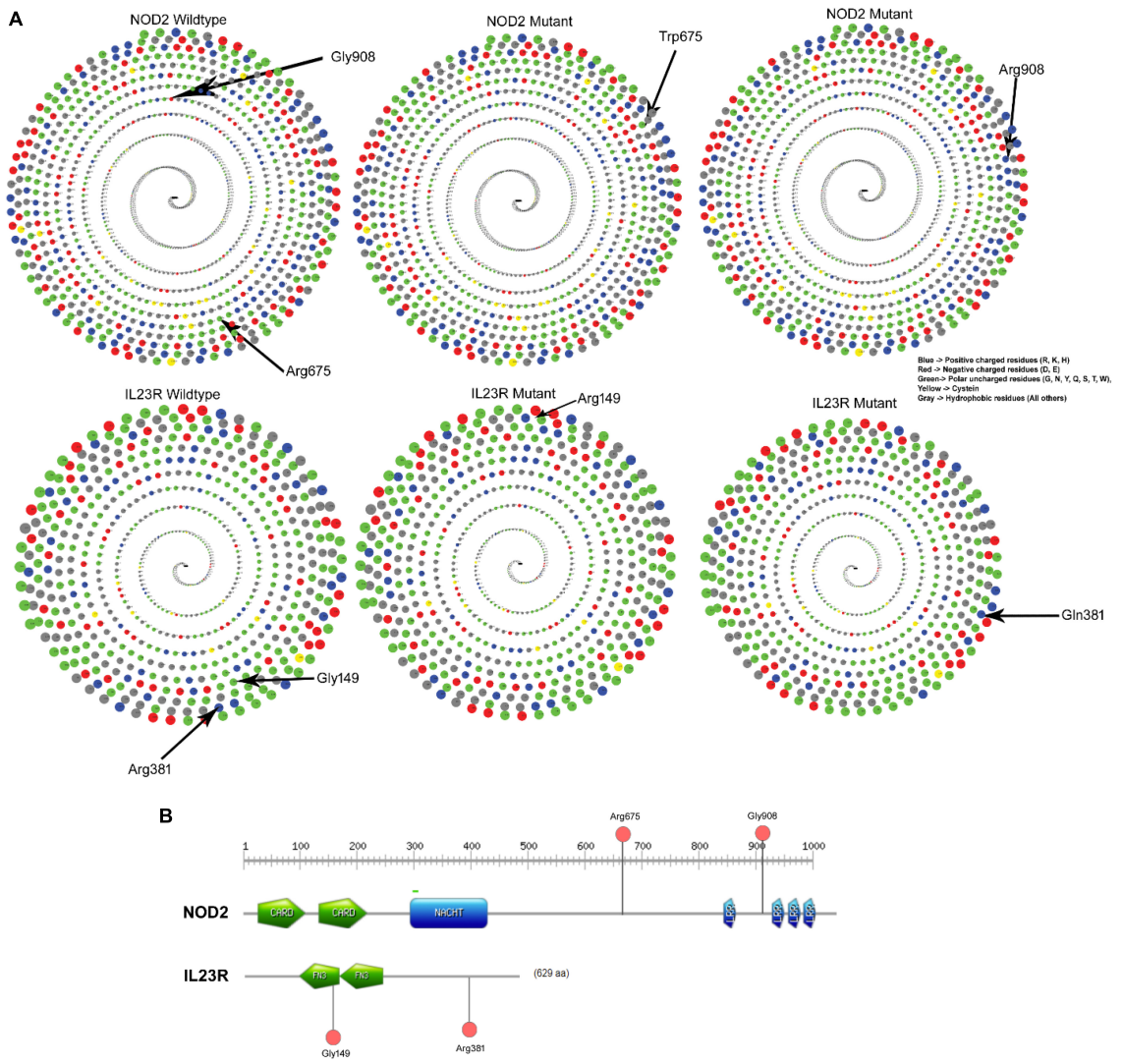


FIGURE 3
(A) Solvent accessibility surface area of wildtype and mutant NOD2 and IL23R. **(B)** Domain region in NOD2 and IL23R.

TABLE 3 Protein stability predictions for mutated and wild models of IL23R and NOD2.

Protein	Mutation	Stability predictions			
		mCSM*	DUET#	SDM$	Consequence
NOD2	Arg675Trp	−0.689	−0.33	−0.774	Destabilizing
	Gly908Arg	−0.895 Kcal/mol	−0.861 Kcal/mol	−0.77	Destabilizing
IL23R	Gly149Arg	−0.567 Kcal/mol	−0.304 Kcal/mol	−1.89	Destabilizing
	Arg381Gln	−0.058 Kcal/mol	−0.355 Kcal/mol	−0.5 Kcal/mol	Destabilizing

*mCSM: <-0 = destabilizing; >0 stabilizing.
#DUET = <-0 = destabilizing; >0 stabilizing.
$SDM: <-0 = destabilizing; >0 stabilizing.

stabilized after 6,000 ps, where they achieved RMSD values in the range of 0.55 to 0.75 nm (Figure 4A). For the *IL23R* wild type, stability in the graph was achieved after 4,300 ps at a RMSD value of 0.43 nm. For Gly149Arg and Arg381Gln of *IL23R* models, a change in stability was observed at 430 ps (RMSD value is 0.39 nm) and 4,000 ps (RMSD value is 0.45 nm) (Figure 4B). In addition to this, we have also assessed the radius of gyration (Rg) and solvent accessible surface area (SASA)

analyses to determine the tertiary structural features of proteins. The SASA identified the marginal exposure of Arg675Trp and Gly908Arg of *NOD2* and Gly149Arg and Arg381Gln of *IL23R* to solvent accessible areas (both hydrophilic and hydrophobic) in both native and mutant forms. However, they were found to be stable in the simulation phase. Our radius of gyration analysis showed that Rg values are different between *NOD2* wild (Rg value is 0.35 nm) and mutant (Rg value is 0.28 nm) types, suggesting the mutation induces conformational changes in the protein. The root mean square fluctuations analysis with *NOD2* and *IL23R* variants revealed flexible regions in the proteins' 3D structures. The ligand recognition region in Gly908 (wild type) *NOD2* is more flexible (RMSF score, which is 0.6 nm) than in 908Arg (mutant), which is more rigid (RMSF score is 0.32 nm). However, this change was not able to alter the overall domain flexibility but only the flexibility of surrounding amino acid residues (**Figure 4C**). For *IL23R*, the wild type model showed the fluctuations or flexibility of amino acid residues in the immunoglobulin like domain (60–80 amino acids) with a value higher than 0.7 nm. The 149 Arg mutant form (RMSF value of 0.45 nm) is located in the immunoglobulin region and affects the fluctuation nature of this region. The Arg381Gln mutation of *IL23R* is located near the immune globulin like domain, and its RMSF values showed more or less similar distribution in both native and mutant forms (**Figure 4D**).

We have also examined the secondary structural element features of both native and mutant *NOD2* and *IL23R* models during the simulation period. At 10 ns simulation time, the wild type *NOD2* conformation had 150–256 H-bonds, while the mutant (Gly908Arg) conformation had 173–252 H-bonds. The *NOD2* mutated model showed some distinct features of secondary structural elements, which suggests that the concerned amino acid residue disturbs its natural bonding with neighboring amino acids in the polypeptide chain. At 10ns simulation period, *IL23R*'s native conformation showed 185–196 H-bonds, while the mutants *IL23R* (Gly149Arg, Arg381Gln) showed fewer H-bonds that is ∼130–145 and ∼145–168 respectively. For *IL23R*, interestingly, both the two mutated models showed similar secondary structural elements compared to their wild type counterparts. So, it is clear that changes in the amino acid sequences of *NOD2* and *IL23R* genes affect the protein's structural stability.

3.5.1. Gene interaction network findings

Gene network analysis of *NOD2* and *IL23R* was performed with GeneMania to better understand their interacting gene partners. **Figure 5A** shows the physical interactions, co-expression, predicted interactions, pathways shared, co-localization, and shared protein domains network of *NOD2*. *NOD2* showed physical interaction with 18 genes, which play a very important role in many immune related pathways. *NOD2* showed co-expression with 3 genes, i.e., *RIPK2, TLR2* and *CARD9*. Interestingly, the *NOD2* interacting genes like *RIBK2*, *IKBG*, and *NKB1* are seen to share the nucleotide-binding domain leucine rich repeat receptor singling pathway, innate immune response pathway, intracellular signaling pathway, and inflammatory response pathways. Co-localization network analysis showed the interaction of *CASP4* and *TLR2* genes with *NOD2*.

NOD2 is also seen to share Leucine Rich Repeat and CARD Domain Containing 2 domains with *CASP1, CASP4, CASP12, CARD8, CARD9, NLRP1, NLRP4* and *RIPK2* genes. Out of all the genes involved in network, 7 genes i.e., *IKBKG, NLRC4, NFKB1, CARD9, RIPK2, XIAP* and *TLR2* plays important role in mediating the innate immune reactions. The other candidate gene *IL23R* shows direct physical interaction with IL23A and IL12RB1 genes in a network. The *IL23R* is co-expressed with IL18 and shares similar pathways with 19 genes. The gene partners which showed physical interaction, co-expression, and shared common pathways with *IL23R* gene, were all majorly involved in T-cell regulation function (**Figure 5B**).

3.5.2. Protein-protein docking studies

Based on our gene-gene network analysis, we predicted that *RIPK2* is the best interacting partner of *NOD2,* owing to its highest confidence score (0.999) (**Figure 5C**). Experimental studies have proved that in the presence of ligand peptidoglycan, *NOD2* interacts with *RIPK2* to perform different intracellular reactions. Therefore, we have employed advanced docking approaches to better understand the molecular interactions between *RIPK2* and *NOD2* (both wild and mutant types). The docking analysis showed that *RIPK2* interacts with wild type *NOD2* near to its signal recognition site and interacts with Trp93, Asp113, Lys118, Leu167, Tyr192, Asn276 and releases the energy of −467.8 Kj/Mol. The Arg675Trp and Gly908Arg mutant forms of *NOD2* interacts with *RIPK2* at few similar sites to that of the wild type, but they form H-bonds with different amino acid residues and releases the energy of >-400 Kj/Mol (**Figure 5C**). The network analysis of *IL23R* revealed that *JAK2*, is its strong interacting partner owing to its confidence score, i.e., 0.998. Our molecular docking analysis showed that *JAK2* interacts with *IL23R*, near C-terminal region amino acid residues Trp307, Asn405, Tyr476, Gln465 and Pro 478 and releases the binding energy of −635.6 KJ/Mol. The mutant models of *IL23R* (Trp307, His345, Phe441, Asp479, Leu310, Thr472) are shown to bind the similar cleft of *JAK2* as the wild type does and release the energy of −659.4 KJ/Mol and −652.5 KJ/Mol (**Table 4** and **Figure 5D**).

3.5.3. Identification of potential drugs against *NOD2* and *IL23R variants*

From the gene-drug interaction database[12] and from literature sources, we identified Tacrolimus and Celecoxib drugs which show specificity toward *NOD2* and *IL23R*, respectively.

12 http://dgidb.genome.wustl.edu/

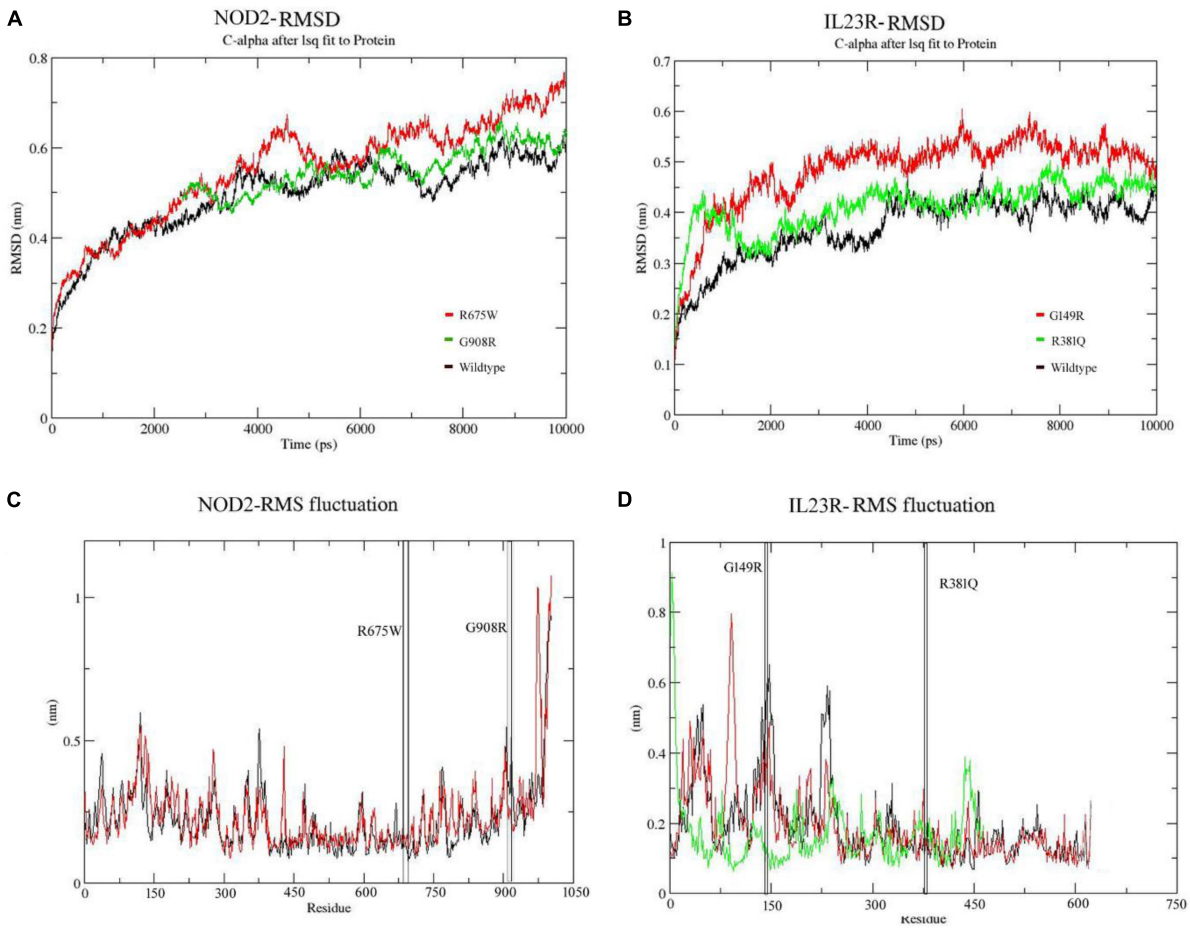


FIGURE 4
(A,B) Molecular dynamics RMSD of NOD2 and IL23R at 100ns. **(C,D)** MD simulation RMSF of NOD2 and IL23R at 100 ns.

Through the advanced molecular docking approaches, we identified that Tacrolimus interacts with both wild and mutant *NOD2* at the ligand binding sites (His720 and His734 amino acids) and releases the energy of −6.12K.Cal/Mol. However, the *NOD2* mutant forms (R675W and G908R) interact with Tacrolimus drug at the same ligand binding region and releases −7.21 K.cal/mol and −6.78 K.cal/mol energy, respectively. The hex docking analysis on the ligand muramyl peptide and *NOD2* interaction revealed that muramyl peptide binds more strongly to the mutant form (with an energy release of −68.2 K.Cal/Mol) compared to the wild (with an energy release of −62.5 K.Cal/Mol) form of *NOD2* (Figure 6A). For *IL23R*, both wild and mutant protein forms showed greater interaction with the Celecoxib drug, although their interacting poses are different. The Celecoxib interacts with the Thr261 amino acid residue in wild *IL23R* and releases the energy of −3.25 K.Cal/Mol. However, the same drug showed the highest interaction (in the form of H-bonds) with Thr261, Asn262 and Thr264 amino acid residues of mutant *IL23R* (G149R) with a binding energy of −10.42K.Cal/Mol. The second *IL23R* mutant (R381N) showed an interaction with Gln263 amino acid residue and released the energy of −4.89 K.Cal/Mol (Figure 6B and Table 5).

4. Discussion

The experimental elucidation of the genotype-protein phenotype relationship is an uphill task owing to the number of variant discoveries being added to the already existing huge IBD mutation data (33). In this context, computational prediction algorithms act as reliable tools for prioritizing candidate genetic mutations based on the nature of their impact (negative, neutral, or positive mutations) on proteins. In the current investigation, we have systemically applied diverse computational strategies to characterize the IBD variants based on their evolutionary constraints on coding regions. These strategies included algorithmic screening of genetic mutations based on the nucleotide sequence and protein structure conservation (integrated support vector machine

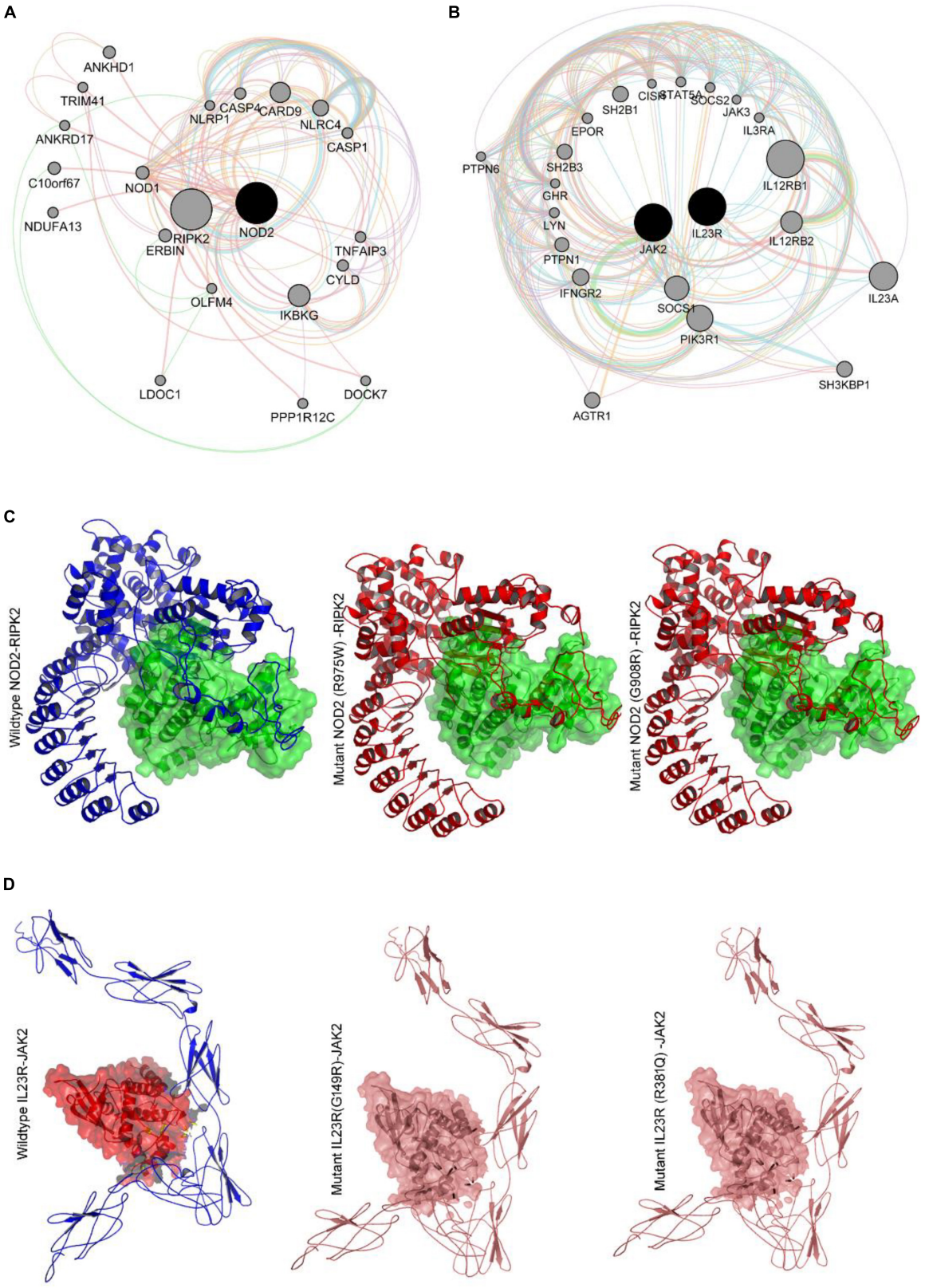


FIGURE 5

(A,B) Genemania protein interaction network of NOD2 and IL23R. **(C)** Molecular docking intearaction of wildtype and mutant NOD2 with RIPK2. **(D)** Molecular docking interaction of wildytype and mutant IL23R with JAK2.

learning algorithms) (ex: SIFT, PloyPhen2, GERP++, PhyloP and SiPhy) across different mammalian species (34). The rationality behind using multiple prediction methods is to generate consensual variant predictions.

The importance of comprehensive computational predictions of missense variants in CA2, LDLRAP1 and SQSTM1 genes has been well demonstrated (24, 35, 36). In the recent study, Polyphen-2, when compared with SIFT, M-CAP and CADD tools, can make better pathogenicity predictions for familial hypercholesterolemia (FH) causative LDLRAP1 mutations (24). Further verification of different computational tools like SIFT, PolyPhen, M-CAP, CADD in screening PCSK9 missense mutations causative to FH is also well demonstrated (37). Few other studies have also asserted the usefulness of various computational algorithms in predicting the damaging ability of nucleotide sequence variations belonging to human disease related genes (34, 38, 39).The quantitative measurement of each constrained element in GERP++ is according to the magnitude of the substitution deficit, measured as "rejected substitutions" (RS). Here, the negative and positive RS scores are inversely proportional to evolutionary selections, where in negative scores often are often considered to be strong signal of biological function. From our GERP++ analysis, we discovered that all the four SNPs fall in evolutionarily conserved regions (RS < 0) and are under strong negative selection. But, due to inherent differences of coding region with regards to the pattern of evolutionary constraints, analysis of population specific genetic variations in regulatory regions, which undergoes evolutionary remodeling, will be of greater assistance to better understand the human specific evolutionary selections (8).

The specific structural and functional implications of any genetic mutation (on its corresponding protein) can be predicted based on the information about the significance of amino acids it alter. In this context, amino acid residues which fall in evolutionarily conserved regions serves as important pointers in understanding the clear effects of genetic mutations of human diseases. Highly pathogenic mutations in a protein hotspot or active region may disrupt the activity of the protein (40). Additionally, studying the mutations at 3 dimensional structure level will help us in understanding the specific structural deformities a particular amino acid variant is likely to inflict on protein. The mapping of the mutation onto three-dimensional protein structures and analyzing these changes at the structural level will help to find the exact point where they loss function/alter interactions with proteins (41). As of today, the tertiary structural conformation of native and/or mutant *NOD2* (Arg675Trp, Gly908Arg) and *IL23R* (Gly149Arg, Arg381Gln) is not yet resolved through laboratory experimental x-ray crystallographic or NMR spectroscopic methods. So, we built the 3 dimensional structural models of *NOD2* (Arg675Trp, Gly908Arg) and *IL23R* (Gly149Arg, Arg381Gln) proteins by *ab initio* method, and analyzed for its biophysical characteristics like stability differences, structural deviations, solvent accessible surface areas and secondary structural features such as polypeptide folding (42).

The structural divergence of core proteins often correlates with amino acid sequence divergence in an exponential function manner (43). In our structural deviation analysis, the Arg675Trp and Gly908Arg mutations of NOD2 have indicated their significance by causing huge structural drift at amino acid residues but not at whole structure level deviations. The NOD2 mutation Arg675Trp variant is not directly localized in the domain region of the protein. However, the Arg675 amino acids form an H-bond network with the surrounding amino acids. Whereas in the mutated condition, the H-bond network is depleted, and this might cause structural alteration in the NOD2 protein (44). The second mutation G908R of NOD2 is actually located in Lucine Rich Domain (735–1,040a.a) (LRRD), which folds as horseshoe enabled by its rich content of hydrophobic amino acid leucine (45). Although, this domain is not directly involved in protein-protein interacting sites, but it assist in stabilizing the NOD2 polypeptide folds which have active site domains (46). Thus, it is explicable that Gly908Arg mutation is capable of altering the NOD2 interaction ability by changing its H-bond properties. In contrast, Gly149Arg and Arg381Gln mutations of IL23R are not seen to inflict any significant structural deviations at both amino acid and whole structure level. Single compared to multiple amino acid residue substitutions often fails to invoke compensatory effects (caused in case of multiple amino acid substitutions) on protein structure, they induce changes in side chain charge (47), active site conformations and polypeptide complementarity, which are essential for maintaining the protein structures. The two mutations (G149R and R381Q) of *IL23R* are located in extracellular domain and C-terminal cytoplasmic portions, respectively (48). Due to mutation G149 in *IL23R* structure the transmembrane domain the first beta barrel of *IL23R* increases its size from Ser251-Lys258 to Val251-Lys258; this structural change may alter the binding ability of extracellular domain of *IL23R* with its ligand (49). In the second mutated protein structure of *IL23R* (R381Q), helical structure (Leu468-Thr472) is converted into loop component in the extracellular domain portion, there by altering its binding ability with first intermediate molecules critical for inducing cascade of intracellular cellular signaling mechanisms underlying inflammatory bowel disease.

We used the molecular dynamics simulation approach to examine the natural and mutant NOD2 and IL23R structures at the atomic level to gain a better understanding about missense mutations induced impacts on protein structures. From the simulation trajectory values, basic metrics such as RMSD, RMSF, hydrogen bond numbers, and SASA were evaluated. Molecular stability and flexibility changes were estimated from

TABLE 4 Hex docking interaction scores of NOD2 and IL23R, wildtype and mutant models with their interaction partners.

Protein	Interaction partner	Hex binding energy (Kcal/Mol)	Difference in binding energy (Kcal/Mol)	Amino acid in interaction
NOD2 WILDTYPE	RIPK2	−467.8	–	Trp93, Asp113, Lys118, Leu167, Tyr192, Asn276
NOD2 mutant (R675W)		−418.2	49.6	Trp93, Lys118
NOD2 mutant (G908R)		−452.6	−15.2	Trp93, Asn94, Leu130, Asn276
IL23R wildtype	JAK2	−635.6	–	Trp307, Asn405, Tyr476, Gln465 and Pro 478
IL23R mutant (G149R)		−659.4	+23.8	Trp307, His345, Phe441, Asp479
IL23R mutant (R381Q)		−652.5	+16.9	Leu310,His 345, Thr472, Asp479

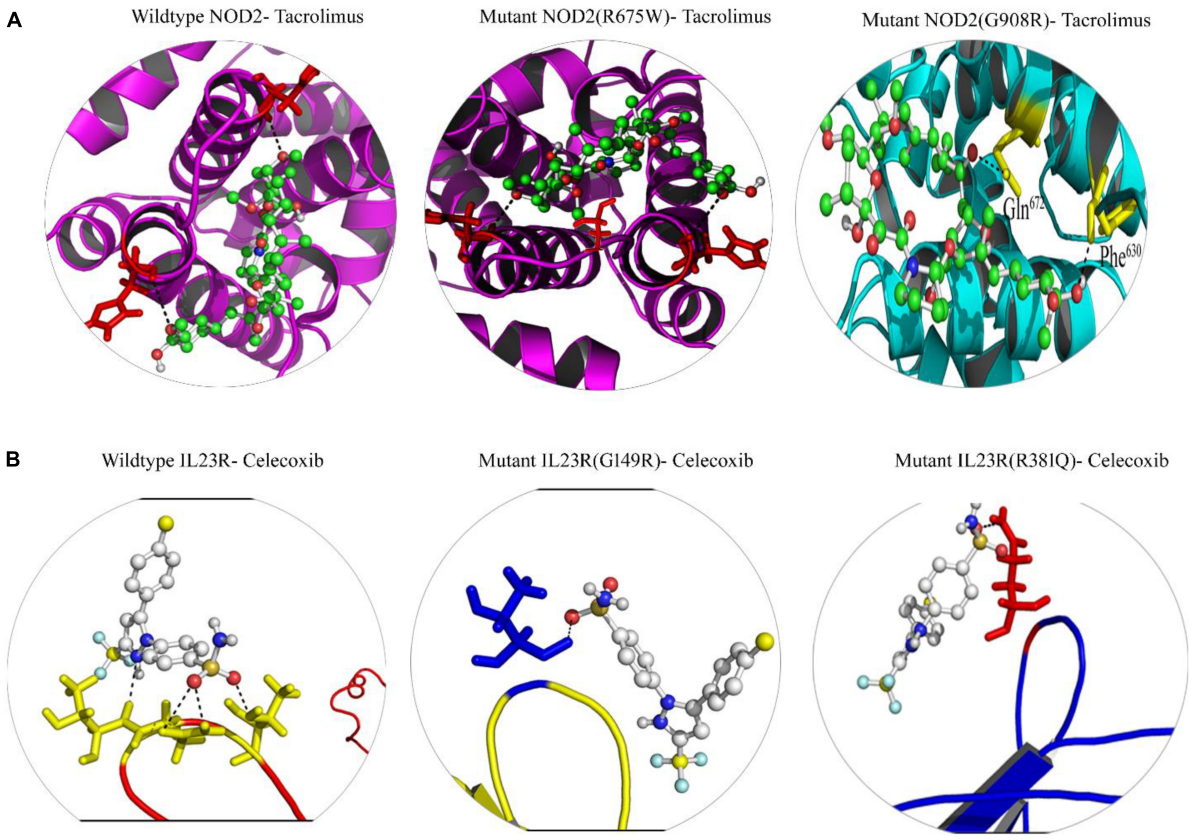


FIGURE 6
(A) Molecular docking interaction of wild and mutant NOD2 with drug Tacrolimus. **(B)** Molecular docking interaction of wild and mutant Il23R with drug Celecoxib.

RMSD and RMSF values. Stability is the fundamental property enhancing biomolecular function, activity, and regulation. In our study, the distinct change in the RMSD trajectories of mutated forms of *NOD2* and *IL23R*, indicate the differences in the route of alteration of structures from the starting conformation to their final states despite the initial structures being identical. This evidence clearly states the impact of amino acid substitutions on the dynamics of the proteins. The RMSF data also showed the mutated regions are highly flexible in both proteins (*NOD2* and *IL23R*) mutations state. A clear insight of stability loss was observed in both RMSD and RMSF parameters, which is further given the evidence by decreasing the number intermolecular hydrogen bonds in mutant structures. Intermolecular H-bonds are most important factors in maintain the protein conformation and creates stable interaction between the protein and its binding partner (50).

The exponential function between structural divergence of protein and amino acid sequence variation is variable based on the mutation rates of amino acid residues, which

TABLE 5 Docking energies of Drugs vs. NOD2 and IL23R (wild/mutant).

Protein	DRUG	Cluster[a]	RMSD[b]	Binding energy[c] (Kcal/mol)	Inhibition constant[d] (*Ki*)	No of H bonds	Amino acid involved in interaction
NOD2 wild type	Tacrolimus	38	0.458	−6.12	32μM	2	His720 and His734
NOD2 mutant (R675W)	Tacrolimus	18	1.43	−7.21	15 μM	2	Gln672, His720 and Phe630
NOD2 mutant (G908R)	Tacrolimus	32	1.256	−6.78	10 μM	2	Gln672 and Phe630
IL23R wildtype	Celecoxib	40	0.125	−3.25	53 μM	1	Thr261
IL23R mutant (G149R)	Celecoxib	44	0.225	−10.42	21.879nM	3	Thr261, Asn262 and Thr264
IL23R mutant (R381N)	Celecoxib	25	1.32	−4.89	42 μM	1	Gln263

[a]Indicative of the total number of binding modes produced.
[b]Heavy atoms root-mean-square deviation with respect to the experimental structure.
[c]The change in binding free energy is related to the inhibition constant using the equation: $\Delta G = RT \ln Ki$, where R is the gas constant 1.987 cal K-1 mol-1, and T is the absolute temperature assumed to be 298.15 K.
[d]Estimated inhibition constant at 298.15 K.

occupies either buried or accessible positions on protein surface (34). Following this principle, we identified that both Arg 675 (native) and 908 glycine (native) amino acids of *NOD2* protein is in buried position and portrays only 20, 40% surface accessible area to solvents, whereas the substitution of tryptophan and arginine amino acids, due to its physical conformation, portrays 80 and 60% surface accessible area to solvents. The Phe149 and Arg381 amino acid (native) residues of *IL23R* showed 80% surface accessible regions, out of which, only Arg381 showed the major drift (80–100%) in its solvent accessibility ability. An explanation in accordance with our observation is that amino acid residues in core portion of proteins is differentially conserved in terms of their sequence and structure, than those that solvent accessible (51). The good correlation of solvent accessibility and stability analysis suggests that *NOD2* and *IL23R*, further confirms that drift in solvent accessibility affects the protein stability.

The networking analysis of genes is a useful approach to understand the functional interactions and associated signaling cascades. The networks shown in form of arcs (relationships) and nodes (entity) are based up on their connectivity levels with other interacting proteins. The *NOD2* networking analysis suggested its strong role in immune mediated pathways. The *NOD2* showed physical interaction with 18 genes, which are playing very important role in many immune related pathways. The *NOD2* showed co-expression with 3 genes i.e., *RIPK2,* TLR2 and CARD9. Interestingly, the *NOD2* interacting genes like RIBK2, IKBG and NKB1 are seen to be sharing Nucleotide-binding domain leucine rich repeat receptor singling pathway, Innate immune response pathway, Intracellular signaling pathway, Inflammatory response pathways. Co-localization network analysis showed the interaction of CASP4 and TLR2 genes with *NOD2. NOD2* is also seen to share Leucine Rich Repeat And CARD Domain Containing 2 domains with CASP1, CASP4, CASP12, CARD8, CARD9, NLRP1, NLRP4 and *RIPK2* genes. Out of all the genes involved in network, 7 genes i.e., IKBKG, NLRC4, NFKB1, CARD9, *RIPK2,* XIAP and TLR2 plays important role in mediating the innate immune reactions. The genetic network *NOD2* showed that *RIPK2* is its highest interaction partner owingto the confidence string score of 0.999. *RIPK2* plays an important role in modulation of immune response (both adaptive and innate). The exposure of peptidoglycan content of foreign particles can activate both *NOD2* and NOD1, which further interacts with *RIPK2* through two caspase recruitment domains (CARD-CARD) leading to the tyrosine phosphorylation and activation of NF-Kappa B (52). Once NFKB is released and translocates into nucleus it activates hundreds of genes responsible for immune responses, growth control and apoptotic mechanisms. To better understand the interactions between *NOD2* with *RIPK2,* protein-protein docking study was performed, where we identified that *RIPK2*

binds at CARD domain (95–182AA) of *NOD2* (45). The NOD2 mutant form Arg675Trp forms weaker interactions with RIPK2 compared to wildtype conditions, indicating the mutation may destabilize the interaction of RIPK2 with NOD2 protein. The second mutant condition (G908R) state, *NOD2* interacts with *RIPK2* at the same CARD domain. However, the mutant *NOD2* (G908R, located in LRRD domain) shows differential binding conformation in terms of interacting amino acids, leading to energy differences between native and mutant forms *NOD2* protein against *RIPK2*.

Our multidimensional computation strategy (pathogenic prediction of nucleotide sequence variations in addition to molecular dynamics simulations) confirms that the R675W and G908R, mutation alters the structural conformation of *NOD2*, thus interaction with *RIPK2* and eventually dysregulate the *NOD2* −*RIPK2* signaling pathway. There have been several reports, which indicated the link of *NOD2* mutations with aberrant immune responses in terms of temporal and quantitative effects of activation of the TLR2-*NOD2* −*RIPK2* pathway on secretion of IL-10 further disturbing the between pro- and anti- inflammatory responses against gram-positive bacteria (53).

The other candidate gene *IL23R* shows direct physical interaction with IL23A and IL12RB1 genes in a network. The *IL23R* is co-expressed with IL18 and shares similar pathways with 19 genes. The gene partners showed physical interaction, co-expression and shared common pathway with *IL23R* gene are all majorly involved in T-cell regulation function (54). The selection of *JAK2*, best interacting partner of *IL23R* was based on the confidence string score of 0.093. Janus tyrosine kinase 2 (*JAK2*), a non-receptor type, class III protein is the intermediate molecule that binds to *IL23R*, whose activation by IL23, phosphorylates STAT and activates NFKB pathway that is essential for stimulating inflammatory reactions involving T-cells, NK cells and possibly certain macrophage/myeloid cells. Owing to the lack of data on *IL23R* and *JAK2* molecular binding characteristics, we performed *IL23R-JAK2* molecular docking. It was found that *JAK2* interacts with the cytosolic terminal of native *IL23R* (at Trp307, Asn405, Tyr476, Gln465 and Pro 478 amino acid residues). Interestingly, even in mutant state the *IL23R* also shows the samelevel interaction with *JAK2* but its binding affinity (+16.9 and +23.8 Kcal/Mol) is decreased when compared to wild type *IL23R* and *JAK2* conformation. A recent study G149R mutation of *IL23R*, observed the reduced expression of STAT3 (48). Cellular functional assays have also observed that R381Q mutation affects the constitutive association of *JAK2* with *IL23R*, with effects on subsequent STAT3 recruitment, phosphorylation, and transcriptional activation (55).

As of today, no specific drug or drug targets are established for treating IBD, except steroid medications, which just reduces the severity of inflammatory reactions in IBD patients (56). From our multidimensional computational approach, we propose that, *NOD2* and *IL23R* have the potential to act as molecular targets. The drug, Tacrolimus interacts with *NOD2* at the ligand binding site of *NOD2* and may positively upregulate different crucial pathways involved in immune suppressive mechanisms. On the other hand, Celecoxib, a non-steroidal anti-inflammatory drug shows strong interaction with mutant *IL23R* compared to its wild type, there by regulates the *IL23R* function. Our computational findings pave the way to test non-steroidal anti-inflammatory bowel disease drugs in experimental conditions.

In conclusion, our study found that SIFT, PolyPhen-2, GERP++, PhyloP, SiPhy and REVEL computational algorithms are very helpful in analyzing *NOD2* (R675W and G908R) and *IL23R* (G149R and R381N) variants. The secondary structure, tertiary structure, and stability prediction approaches have demonstrated how the loss-of-function variants induce minor structural drifts, shift free energy values, and reduce the conformation flexibility of the *NOD2* and *IL23R* protein molecules. Overall, our comprehensive computational approach adds a layer to estimate the deleterious potential of genetic variants associated with IBD. This study recommends implementing multidimensional genotype – protein phenotype assessment methods as a pre-laboratory approach in developing personalized medicine for IBD patients carrying NOD2 (R675W and G908R) and IL23R (G149R and R381N) variants.

Data availability statement

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

Author contributions

KN and TS: conceptualization, data curation, formal analysis, methodology, supervision, visualization, and writing original draft and editing. KN: funding acquisition, project administration, software, and validation. Both authors contributed to the article and approved the submitted version.

Funding

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, under Grant no. G:356-142-1441. The authors, therefore, acknowledge the DSR for technical and financial support.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2022.1090120/full#supplementary-material

References

1. Kaser A, Zeissig S, Blumberg R. Inflammatory bowel disease. *Ann Rev Immunol.* (2010) 28:573–621. doi: 10.1146/annurev-immunol-030409-101225

2. Khor B, Gardet A, Xavier R. Genetics and pathogenesis of inflammatory bowel disease. *Nature.* (2011) 474:307–17. doi: 10.1038/nature10209

3. Baumgart D, Carding S. Inflammatory bowel disease: cause and immunobiology. *Lancet.* (2007) 369:1627–40. doi: 10.1016/S0140-6736(07)60750-8

4. Jostins L, Ripke S, Weersma R, Duerr R, McGovern D, Hui K, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature.* (2012) 491:119–24. doi: 10.1038/nature11582

5. Liu J, van Sommeren S, Huang H, Ng S, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet.* (2015) 47:979–86. doi: 10.1038/ng.3359

6. Girardelli M, Loganes C, Pin A, Stacul E, Decleva E, Vozzi D, et al. Novel NOD2 mutation in early-onset inflammatory bowel phenotype. *Inflamm Bowel Dis.* (2018) 24:1204–12. doi: 10.1093/ibd/izy061

7. Alharbi R, Shaik N, Almahdi H, Jamalalail B, Mosli M, Alsufyani H, et al. Genetic association study of NOD2 and IL23R amino acid substitution polymorphisms in Saudi Inflammatory Bowel Disease patients. *J King Saud Univ Sci.* (2022) 34:101726. doi: 10.1016/j.jksus.2021.101726

8. Mesbah-Uddin M, Elango R, Banaganapalli B, Shaik N, Al-Abbasi F. In-silico analysis of inflammatory bowel disease (IBD) GWAS loci to novel connections. *PLoS One.* (2015) 10:e0119420. doi: 10.1371/journal.pone.0119420

9. Yamamoto S, Ma X. Role of Nod2 in the development of Crohn's disease. *Microb Infect.* (2009) 11:912–8. doi: 10.1016/j.micinf.2009.06.005

10. Cho J, Abraham C. Inflammatory bowel disease genetics: Nod2. *Annu Rev Med.* (2007) 58:401–16. doi: 10.1146/annurev.med.58.061705.145024

11. Sidiq T, Yoshihama S, Downs I, Kobayashi K. Nod2: A critical regulator of ileal microbiota and crohn's disease. *Front Immunol.* (2016) 7:367. doi: 10.3389/fimmu.2016.00367

12. Rochereau N, Roblin X, Michaud E, Gayet R, Chanut B, Jospin F, et al. NOD2 deficiency increases retrograde transport of secretory IgA complexes in Crohn's disease. *Nat Commun.* (2021) 12:261. doi: 10.1038/s41467-020-20348-0

13. Wilson N, Boniface K, Chan J, McKenzie B, Blumenschein W, Mattson J, et al. Development, cytokine profile and function of human interleukin 17-producing helper T cells. *Nat Immunol.* (2007) 8:950–7. doi: 10.1038/ni1497

14. Raychaudhuri S, Abria C, Raychaudhuri S. Regulatory role of the JAK STAT kinase signalling system on the IL-23/IL-17 cytokine axis in psoriatic arthritis. *Ann Rheum Dis.* (2017) 76:e36–36. doi: 10.1136/annrheumdis-2016-211046

15. Onodera K, Arimura Y, Isshiki H, Kawakami K, Nagaishi K, Yamashita K, et al. Low-Frequency IL23R coding variant associated with crohn's disease susceptibility in japanese subjects identified by personal genomics analysis. *PLoS One.* (2015) 10:e0137801. doi: 10.1371/journal.pone.0137801

16. Abdollahi E, Tavasolian F, Momtazi-Borojeni AA, Samadi M, Rafatpanah H. Protective role of R381Q (rs11209026) polymorphism in IL-23R gene in immune-mediated diseases: A comprehensive review. *J Immunotoxicol.* (2016) 13:286–300. doi: 10.3109/1547691x.2015.1115448

17. Ng P, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* (2003) 31:3812–4. doi: 10.1093/nar/gkg509

18. Adzhubei I, Jordan D, Sunyaev S. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet.* (2013) Chapter 7:Unit720. doi: 10.1002/0471142905.hg0720s76

19. Jagadeesh K, Wenger A, Berger M, Guturu H, Stenson P, Cooper D, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet.* (2016) 48:1581–6. doi: 10.1038/ng.3703

20. Rogers M, Shihab H, Mort M, Cooper D, Gaunt T, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics.* (2018) 34:511–3. doi: 10.1093/bioinformatics/btx536

21. Rentzsch P, Witten D, Cooper G, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* (2019) 47:D886–94. doi: 10.1093/nar/gky1016

22. Shaik N, Banaganapalli B. Computational Molecular Phenotypic Analysis of PTPN22 (W620R), IL6R (D358A), and TYK2 (P1104A) Gene Mutations of Rheumatoid Arthritis. *Front Genet.* (2019) 10:168. doi: 10.3389/fgene.2019.00168

23. Morad F, Rashidi O, Sadath S, Al-Allaf F, Athar M, Alama M, et al. In silico approach to investigate the structural and functional attributes of familial hypercholesterolemia variants reported in the saudi population. *J Comput Biol.* (2018) 25:170–81. doi: 10.1089/cmb.2017.0018

24. Shaik N, Al-Qahtani F, Nasser K, Jamil K, Alrayes N, Elango R, et al. Molecular insights into the coding region mutations of low-density lipoprotein receptor adaptor protein 1 (LDLRAP1) linked to familial hypercholesterolemia. *J Gene Med.* (2020) 22:e3176. doi: 10.1002/jgm.3176

25. Bokhari H, Shaik N, Banaganapalli B, Nasser K, Ageel H, Al Shamrani A, 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

26. Nasser K, Banaganapalli B, Shinawi T, Elango R, Shaik N. Molecular profiling of lamellar ichthyosis pathogenic missense mutations on the structural and stability aspects of TGM1 protein. *J Biomol Struct Dyn.* (2021) 39:4962–72. doi: 10.1080/07391102.2020.1782770

27. Papamichael K, Afif W, Drobne D, Dubinsky M, Ferrante M, Irving P, et al. Therapeutic drug monitoring of biologics in inflammatory bowel disease: unmet needs and future perspectives. *Lancet Gastroenterol Hepatol.* (2022) 7:171–85. doi: 10.1016/s2468-1253(21)00223-5

28. Chee D, Goodhand J, Ahmad T. Editorial: is pharmacogenetic testing for adverse effects to IBD treatments ready for roll-out? *Aliment Pharmacol Ther.* (2020) 52:1076–7. doi: 10.1111/apt.16025

29. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat.* (2011) 32:894–9. doi: 10.1002/humu.21517

30. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat.* (2013) 34:E2393–402. doi: 10.1002/humu.22376

31. Pronk S, Pall S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular

simulation toolkit. *Bioinformatics.* (2013) 29:845–54. doi: 10.1093/bioinformatics/btt055

32. Cotto K, Wagner A, Feng Y, Kiwala S, Coffman A, Spies G, et al. DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic Acids Res.* (2017) 46:D1068–73. doi: 10.1093/nar/gkx1143

33. Barral M, Dohan A, Allez M, Boudiaf M, Camus M, Laurent V, et al. Gastrointestinal cancers in inflammatory bowel disease: An update with emphasis on imaging findings. *Crit Rev Oncol Hematol.* (2016) 97:30–46. doi: 10.1016/j.critrevonc.2015.08.005

34. Shaik N, Kaleemuddin M, Banaganapalli B, Khan F, Shaik N, Ajabnoor G, et al. Structural and functional characterization of pathogenic non-synonymous genetic mutations of human insulin-degrading enzyme by in silico methods. *CNS Neurol Disord Drug Targets.* (2014) 13:517–32. doi: 10.2174/1871527311312660161

35. Shaik N, Bokhari H, Masoodi T, Shetty P, Ajabnoor G, 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

36. Shaik N, Nasser K, Alruwaili M, Alallasi S, 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

37. Awan Z, Bahattab R, Kutbi H, Jamal Noor A, Al-Nasser M, Ahmad Shaik N, 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.* (2020) 27:719–33.

38. Rajith B, Doss C. Disease-causing mutation in extracellular and intracellular domain of FGFR1 protein: computational approach. *Appl Biochem Biotechnol.* (2013) 169:1659–71. doi: 10.1007/s12010-012-0061-6

39. Banaganapalli B, Mohammed K, Khan I, Al-Aama J, Elango R, Shaik N. A computational protein phenotype prediction approach to analyze the deleterious mutations of human MED12 Gene. *J Cell Biochem.* (2016) 117:2023–35. doi: 10.1002/jcb.25499

40. Ma M, Wang C, Glicksberg B, Schadt E, Li S, Chen R. Identify cancer driver genes through shared mendelian disease pathogenic variants and cancer somatic mutations. *Pac Symp Biocomput.* (2016) 22:473–84.

41. Creighton T, Freedman R. A model catalyst of protein disulphide bond formation. *Curr Biol.* (1993) 3:790–3. doi: 10.1016/0960-9822(93)90034-l

42. Corridoni D, Rodriguez-Palacios A, Di Stefano G, Di Martino L, Antonopoulos D, Chang E, et al. Genetic deletion of the bacterial sensor NOD2 improves murine Crohn's disease-like ileitis independent of functional dysbiosis. *Mucosal Immunol.* (2016) 10:971–82. doi: 10.1038/mi.2016.98

43. Lesk A, Chothia C. Solvent accessibility, protein surfaces, and protein folding. *Biophys J.* (1980) 32:35–47.

44. Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. *BMC Genomics.* (2012) 13(Suppl 4):S4. doi: 10.1186/1471-2164-13-S4-S4

45. Maekawa S, Ohto U, Shibata T, Miyake K, Shimizu T. Crystal structure of NOD2 and its implications in human disease. *Nat Commun.* (2016) 7:11813. doi: 10.1038/ncomms11813

46. Freiberg A, Machner M, Pfeil W, Schubert W, Heinz D, Seckler R. Folding and stability of the leucine-rich repeat domain of internalin B from Listeri monocytogenes. *J Mol Biol.* (2004) 337:453–61. doi: 10.1016/j.jmb.2004.01.044

47. Fukami-Kobayashi K, Saito N. [How to make good use of CLUSTALW]. *Tanpakushitsu Kakusan Koso.* (2002) 47:1237–9.

48. Sivanesan D, Beauchamp C, Quinou C, Lee J, Lesage S, Chemtob S, et al. IL23R (Interleukin 23 Receptor) variants protective against inflammatory bowel diseases (IBD) display loss of function due to impaired protein stability and intracellular trafficking. *J Biol Chem.* (2016) 291:8673–85. doi: 10.1074/jbc.M116.715870

49. Huang J, Yang Y, Zhou F, Liang Z, Kang M, Kuang Y, et al. Meta-analysis of the IL23R and IL12B polymorphisms in multiple sclerosis. *Int J Neurosci.* (2016) 126:205–12. doi: 10.3109/00207454.2015.1007508

50. Abdul Samad F, Suliman B, Basha S, Manivasagam T, Essa M. A comprehensive in silico analysis on the structural and functional impact of SNPs in the congenital heart defects associated with NKX2-5 gene-a molecular dynamic simulation approach. *PLoS One.* (2016) 11:e0153999. doi: 10.1371/journal.pone.0153999

51. Hubbard T, Blundell T. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng.* (1987) 1:159–71. doi: 10.1093/protein/1.3.159

52. Franca R, Vieira S, Talbot J, Peres R, Pinto L, Zamboni D, et al. Expression and activity of NOD1 and NOD2/RIPK2 signalling in mononuclear cells from patients with rheumatoid arthritis. *Scand J Rheumatol.* (2015) 23:8–12. doi: 10.3109/03009742.2015.1047403

53. Kufer T, Banks D, Philpott D. Innate immune sensing of microbes by Nod proteins. *Ann N Y Acad Sci.* (2006) 1072:19–27.

54. Li Z, Wu F, Brant S, Kwon J. IL-23 receptor regulation by Let-7f in human CD4+ memory T cells. *J Immunol.* (2011) 186:6182–90. doi: 10.4049/jimmunol.1000917

55. Sarin R, Wu X, Abraham C. Inflammatory disease protective R381Q IL23 receptor polymorphism results in decreased primary CD4+ and CD8+ human T-cell functional responses. *Proc Natl Acad Sci USA.* (2011) 108:9560–5. doi: 10.1073/pnas.1017854108

56. Amiot A, Peyrin-Biroulet L. Current, new and future biological agents on the horizon for the treatment of inflammatory bowel diseases. *Therap Adv Gastroenterol.* (2015) 8:66–82. doi: 10.1177/1756283X14558193

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 26 January 2023
DOI 10.3389/fmed.2023.1052923

OPEN ACCESS

EDITED BY
C. George Priya Doss,
VIT University, India

REVIEWED BY
HaiHui Huang,
Shaoguan University, China
Dragos Horvath,
UMR 7140 Chimie de la Matière Complexe, France
Zhibin Lv,
Sichuan University, China

*CORRESPONDENCE
Hui Ding
✉ hding@uestc.edu.cn
Yang Zhang
✉ yangzhang@cdutcm.edu.cn
Ke-Jun Deng
✉ dengkj@uestc.edu.cn

†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 24 September 2022
ACCEPTED 05 January 2023
PUBLISHED 26 January 2023

CITATION
Zhang Y-F, Wang Y-H, Gu Z-F, Pan X-R, Li J, Ding H, Zhang Y and Deng K-J (2023) Bitter-RF: A random forest machine model for recognizing bitter peptides. *Front. Med.* 10:1052923. doi: 10.3389/fmed.2023.1052923

Bitter-RF: A random forest machine model for recognizing bitter peptides

Yu-Fei Zhang[1†], Yu-Hao Wang[1†], Zhi-Feng Gu[1], Xian-Run Pan[2], Jian Li[3], Hui Ding[1*], Yang Zhang[2*] and Ke-Jun Deng[1*]

[1]School of Life Science and Technology, Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China, [2]Innovative Institute of Chinese Medicine and Pharmacy, Academy for Interdiscipline, Chengdu University of Traditional Chinese Medicine, Chengdu, China, [3]School of Basic Medical Sciences, Chengdu University, Chengdu, China

Introduction: Bitter peptides are short peptides with potential medical applications. The huge potential behind its bitter taste remains to be tapped. To better explore the value of bitter peptides in practice, we need a more effective classification method for identifying bitter peptides.

Methods: In this study, we developed a Random forest (RF)-based model, called Bitter-RF, using sequence information of the bitter peptide. Bitter-RF covers more comprehensive and extensive information by integrating 10 features extracted from the bitter peptides and achieves better results than the latest generation model on independent validation set.

Results: The proposed model can improve the accurate classification of bitter peptides (AUROC = 0.98 on independent set test) and enrich the practical application of RF method in protein classification tasks which has not been used to build a prediction model for bitter peptides.

Discussion: We hope the Bitter-RF could provide more conveniences to scholars for bitter peptide research.

KEYWORDS

bitter peptide, sequence information, random forest, feature fusion, classification method

1. Introduction

The bitter peptides, often produced in fermented, aged, and spoiled foods, are oligopeptides with diverse structures. Studies have shown that hydrophobic amino acids and their positions are crucial determinants for bitter peptides to exhibit bitter taste (1, 2). Experiments have found that many toxins are bitter taste, so most mammals, including humans, avoid the intake of toxins by avoiding bitter substances (3). However, some bitter substances may have medicinal effects. In biomedical and clinical sciences, hormetic responses were of considerable importance. Many drugs displayed hormetic-like biphasic dose responses and showed opposite effects at low and high doses (4). In diabetic patients, the peptides in *Momordica charantia* (*M. charantia*) can significantly regulate blood glucose concentration. A 68-residue insulin receptor binding protein was isolated from *M. charantia*. McIRBP-19 in this protein can span the 50th-68th residues, enhance the binding of insulin and IR, stimulate the phosphorylation of PDK1 and Akt, and induce the expression of glucose transporter 4, thus promoting glucose clearance (5). And frequent consumption of *M. charantia* peptide is beneficial to multiple organs of human body (6). The active compound polypeptide K extracted from the seeds of *M. charantia* has gastroprotective effects in some gastric ulcer models (7). Hence, bitter peptides, previously avoided due to their potential toxicity, can be beneficial at the correct dosage. Consequently, the bitter peptides may be very useful in medicine, making their identification extremely important (8).

Experimental methods for identifying bitter peptides have a solid theoretical basis, but the operation is complex, time-consuming, and inaccurate. Biological methods often involve the extraction of bitter peptides from raw materials through gel separation, multiple rounds of liquid chromatography separation, and purification. Finally, Fourier transforms infrared spectroscopy (FTIR) was used to identify bitter peptides. Generally, spectroscopic-based methods have requirements for instruments, which are not universal (9, 10). Therefore, the bitterness evaluation stage requires the participation of human subjects, which may lead to inaccurate results (11, 12). Bioinformatics-based methods for predicting bitter peptides have the advantages of no professional instrument requirements, short time consumption, and high prediction accuracy. Therefore, it is imperative to develop a machine learning model for predicting bitter peptides.

At present, computational methods have been carried out to study peptides (13, 14). Models based on the quantitative structure of bitter taste relationship (QSBR), including multiple linear regression, the support vector machine (SVM), and artificial neural network (ANN), have been used to predict bitter peptides (2, 15–21). Specifically, based on 229 experimental bitterness values determined by human sensory evaluations, Dragon 5.4 software was designed to predict bitter peptides by extracting 1292 descriptors and reducing descriptors to 244 using a home-developed toolbox. Then, the GA-PLS method was used to select the six best-scoring descriptors for the QSAR model construction. The six descriptors, including SPAN, Mean square distance (MSD), E3s, G3p, Hats8U, and 3D-MoRSE, represent the dimension of the molecule, the numbers of atoms, weighted atomic electrical topological states, the 3rd-component symmetry directional WHIM index (weighed by polarizability), spatial autocorrelation-based descriptors and an indicator of size, mass, and volume of the molecules.

Further, to improve prediction accuracy, four generations of classification models based on bitter peptide sequences have been developed. The first-generation model used dipeptide propensity scores to predict bitter peptides by extracting a few characteristics of bitter peptides (22). The second-generation model utilized deep learning research methods. However, there may be problems with information redundancy and overfitting (23). The third-generation model integrated five peptide features to formulate bitter peptides, but the representativeness should be further optimized (24, 25). The fourth-generation model extracted feature extraction by deep learning pre-training, and then built a prediction model based on light gradient boosting machine (LGBM) (26).

Inspired by the previous four generations of models, we proposed Bitter-RF, a novel machine learning method for predicting bitter peptides. In total, ten kinds of feature information were extracted, consisting of 1,337 features in the feature set. By deleting all zero items, 1206 features were used for model learning. Here, we used five machine learning models to learn the features. After comparison, the RF method has the best classification effect. The schematic framework of Bitter-RF for bitter peptide prediction is shown in **Figure 1**.

2. Materials and methods

2.1. Dataset source

The fundamental for constructing a powerful model is to generate a high-quality benchmark dataset. To provide a reliable model and make a fair comparison, we used the same dataset as the previous four generation models (22–24), which can be obtained from http://pmlab.pythonanywhere.com/BERT4Bitter (accessed on 13 January 2022). This data was originally obtained by manually collecting experimentally validated bitter peptides from various literatures (22). The data contains 640 records, including 320 experimentally validated bitter peptides and 320 non-bitter peptides, which were randomly generated from BIOPEP. In order to objectively evaluate the model, we divided the data into training set and independent set at a ratio of 8:2. The training set contains 256 bitter peptides and 256 non-bitter peptides. The independent set contains 64 bitter peptides and 64 non-bitter peptides.

2.2. Feature extraction

In a computational model based on machine learning methods for biological sequence data, the coding methods of sequences, which can reveal as much sequence information as possible, are the most critical step (27–36). In the field of sequence analysis, scholars have done a lot of works, and various of sequence descriptors were proposed. Here, we used iLearnPlus to extract 10 types of features of bitter peptides (37). The specific information was described as follows.

2.2.1. Amino acid composition (AAC)

The AAC encoding calculates the frequencies of 20 natural amino acids in a peptide sequence (38–42). The equation was shown as follows.

$$f(t) = \frac{N(t)}{N},\ t \in \{A, C, \ldots, Y\} \tag{1}$$

where $N(t)$ means the number of amino acid type t, and N means the length of peptides.

2.2.2. Traditional pseudo-amino acid composition (TPAAC)

The TPAAC descriptor is proposed by Chou, which is also called the type1 pseudo-amino acid composition (43). Here, we use $H_1^0(i)$, $H_2^0(i)$, and $M^0(i)$ (i = 1, 2, 3,, 20) to respectively represent the original hydrophobicity values (44), original hydrophilicity values (45) and original side chain masses of 20 natural amino acids. We normalized these values based on the standard normal distribution, as follows.

$$H_1(i) = \frac{H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)}{\sqrt{\frac{\sum_{i=1}^{20}\left[H_1^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_1^o(i)\right]^2}{20}}} \tag{2}$$

$$H_2(i) = \frac{H_2^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_2^o(i)}{\sqrt{\frac{\sum_{i=1}^{20}\left[H_2^o(i) - \frac{1}{20}\sum_{i=1}^{20} H_2^o(i)\right]^2}{20}}} \tag{3}$$

$$M(i) = \frac{M^0(i) - \frac{1}{20}\sum_{i=1}^{20} M^0(i)}{\sqrt{\frac{\sum_{i=1}^{20}\left[M^0(i) - \frac{1}{20}\sum_{i=1}^{20} M^0(i)\right]^2}{20}}} \tag{4}$$

Then, the correlation function for residues R_i and R_j can be defined as:

$$\Theta\left(R_i, R_j\right) = \frac{1}{3}\left\{\left[H_1\left(R_i\right) - H_1\left(R_j\right)\right]^2 + \left[H_2\left(R_i\right) - H_2\left(R_j\right)\right]^2 + \left[M\left(R_i\right) - M\left(R_j\right)\right]^2\right\} \quad (5)$$

The correlation function contains the three amino acid properties mentioned above. By generalizing this function definition, an amino acid property (Eq. 6) and a set of amino acid properties (Eq.7) are defined.

$$\Theta\left(R_i, R_j\right) = \left[H_1\left(R_i\right) - H_1\left(R_j\right)\right]^2 \quad (6)$$

$$\Theta\left(R_i, R_j\right) = \frac{1}{n}\sum_{n=1}^{n}\left[H_k\left(R_i\right) - H_k\left(R_j\right)\right]^2 \quad (7)$$

where $H(R_i)$ is the amino acid property of amino acid R_i after standardization and $H_k(R_i)$ is the k-th attribute in the amino acid attribute set of amino acid R_i. And sequence order-correlated factors were defined as:

$$\theta_1 = \frac{1}{N-1}\sum_{i=1}^{N-1}\Theta\left(R_i, R_{i+1}\right) \quad (8)$$

$$\theta_2 = \frac{1}{N-2}\sum_{i=1}^{N-2}\Theta\left(R_i, R_{i+2}\right) \quad (9)$$

$$\cdots$$

$$\theta_\lambda = \frac{1}{N-\lambda}\sum_{i=1}^{N-\lambda}\Theta\left(R_i, R_{i+\lambda}\right) \quad (10)$$

where λ is a correlation parameter that can be adjusted, and λ should be less than N, we set $\lambda = 1$. And traditional pseudo-amino acid composition for a protein sequence can be defines as:

$$X_c = \frac{f_c}{\sum_{r=1}^{20} f_r + \omega\sum_{j=1}^{\lambda}\theta_j}, (1 < c < 20) \quad (11)$$

$$X_c = \frac{\omega\theta_{c-20}}{\sum_{r=1}^{20} f_r + \omega\sum_{j=1}^{\lambda}\theta_j}, (21 < c < 20+\lambda) \quad (12)$$

where ω is the weigthing factor and is set to 0.05 in this study.

2.2.3. Amphiphilic pseudo-amino acid composition (APAAC)

The APAAC is a kind of PseAAC. It contains 20+2λ discrete numbers: the first 20 numbers consist of conventional amino acids; the next 2λ numbers are a set of correlation factors that reflect different distribution patterns of hydrophobicity and hydrophilicity along the peptide chain (46). This feature was described as follows.

Firstly, using $H_1(i)$(Eq.2) and $H_2(i)$(Eq.3) which are defined in TPAAC to define hydrophobicity and hydrophilicity correlation functions:

$$H_{i,j}^1 = H_1(i)\,H_1(j) \quad (13)$$

$$H_{i,j}^2 = H_2(i)\,H_2(j) \quad (14)$$

Secondly, sequence order factors can be formulated as:

$$\tau_1 = \frac{1}{N-1}\sum_{i=1}^{N-1} H_{i,i+1}^1 \quad (15)$$

$$\tau_2 = \frac{1}{N-1}\sum_{i=1}^{N-1} H_{i,i+1}^2 \quad (16)$$

$$\tau_3 = \frac{1}{N-2}\sum_{i=1}^{N-2} H_{i,i+2}^1 \quad (17)$$

$$\tau_4 = \frac{1}{N-2}\sum_{i=1}^{N-2} H_{i,i+2}^2 \quad (18)$$

$$\cdots$$

$$\tau_{2\alpha-1} = \frac{1}{N-\alpha}\sum_{i=1}^{N-\alpha} H_{i,i+\alpha}^1 \quad (19)$$

$$\tau_{2\alpha} = \frac{1}{N-\alpha}\sum_{i=1}^{N-\alpha} H_{i,i+\alpha}^2 \quad (20)$$

Finally, the APAAC is defined as:

$$P_C = \frac{f_c}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda}\tau_j}, (1 < c < 20) \quad (21)$$

$$P_C = \frac{\omega\tau_u}{\sum_{r=1}^{20} f_r + w\sum_{j=1}^{2\lambda}\tau_j}, (21 < u < 20+2\lambda) \quad (22)$$

where w is the weighting factor, and it is set to 0.5 in this study. This value refers to Chou's work on protein cell property prediction using this feature (43). And we set λ 1 in this study.

2.2.4. Adaptive skip dinucleotide composition (ASDC)

ASDC is a modified dipeptide composition, which takes full account of the relevant information that exists between adjacent residues and between intervening residues. The feature vector for ASDC was defined as:

$$\text{ASDC} = (f_{v1}, f_{v2}, \ldots, f_{v400}),$$

$$f_{vi} = \frac{\sum_{g=1}^{L-1} O_i^g}{\sum_{i=1}^{400}\sum_{g=1}^{L-1} O_i^g} \quad (23)$$

where f_{vi} means the occurrence frequency of all possible dipeptide with ≤ *L-1* intervening peptides.

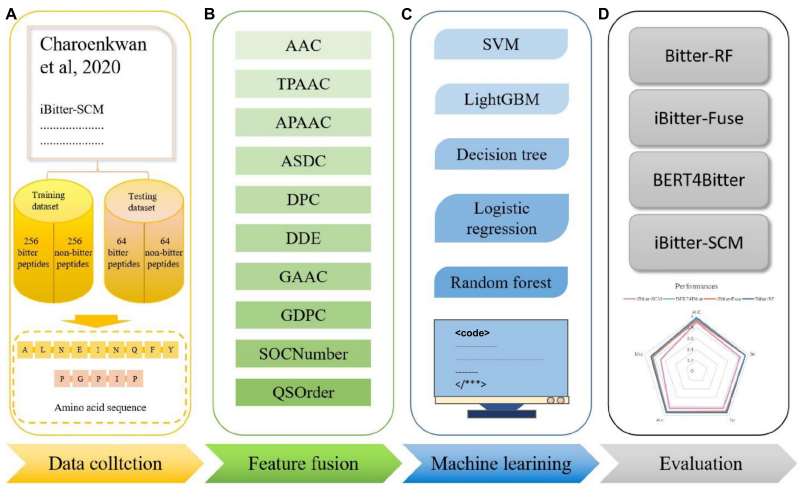


FIGURE 1
Schematic framework of the Bitter peptide prediction model (Bitter–RF). The main process of Bitter-RF design mainly includes the following steps: **(A)** dataset collection, **(B)** feature fusion, **(C)** modeling with multiple machine learning methods, **(D)** Bitter-RF performance evaluation.

2.2.5. Di-peptide composition (DPC)

The DPC encoding describes the frequencies of 400 dipeptide combination in peptide sequence (47). The calculation method was shown as follows.

$$D(r,s) = \frac{N_{rs}}{N-1},\ r,s \in \{A, C, D, \ldots, Y\} \tag{24}$$

where N_{rs} means the number of dipeptides combined by amino acid types r and amino acid types s and N is the length of peptide.

2.2.6. Dipeptide deviation from expected mean (DDE)

DDE includes three parameters: dipeptides composition (D_c), theoretical mean (T_m), and theoretical variance (T_v). D_c is the same as DPC's calculation method.T_m and T_v were calculated as follows:

$$T_m(r,s) = \frac{C_r}{C_N} \times \frac{C_s}{C_N} \tag{25}$$

$$T_v(r,s) = \frac{T_m(r,s)(1 - T_m(r,s))}{N-1} \tag{26}$$

where C_r means the number of codons for the amino acid types r, and C_s means the number of codons for the amino acid types s. C_N includes total possible codons, which means not including the three stop codons.

Using three parameters, DDE was calculated as follows:

$$DDE(r,s) = \frac{D_c(r,s) - T_m(r,s)}{T_v(r,s)} \tag{27}$$

2.2.7. Grouped amino acid composition (GAAC)

GAAC divides 20 amino acids into five groups based on their physicochemical properties that are the aliphatic group (*g1*: GAVLMI), aromatic group (*g2*: FYW), positive charge group (*g3*: KRH), negative charged group (*g4*: DE) and uncharged group (*g5*: STCPNQ). This feature describes the frequencies of these five groups of amino acids and can be calculated as follows:

$$f(g) = \frac{N(g)}{N},\ G \in \{g1, g2, g3, g4, g5\} \tag{28}$$

where $N(g)$ is the sum of the number of the amino acid which belongs to group g, and N is the length of peptide sequence.

2.2.8. Grouped dipeptide composition (GDPC)

GDPC is a variant of DPC based on the amino acid classification already mentioned in GAAC. The feature consists of 25 descriptors, calculated as follows:

$$f(r,s) = \frac{N_{rs}}{N-1},\ r,s \in \{g1, g2, g3, g4, g5\} \tag{29}$$

where N_{rs} is the number of dipeptides represented by amino acid type groups r and s, and N is the length of peptide sequence.

2.2.9. Sequence-order-coupling number (SOCNumber)

The d-th rank sequence-order-coupling number was calculated as follows:

$$\tau_d = \sum_{i=1}^{N-d} \left(d_{i,i+d}\right)^2,\ d = 1, 2, \ldots, nlag \tag{30}$$

where $d_{i,i+d}$ describes the distance between two amino acids at positions i and $i + d$ in a given distance matrix, *nlag* denotes the maximum value of the lag (default value: 30) and N is the length of the peptide sequence. The distance matrix used here from both Schneider–Wrede physicochemical distance matrix (48) and Grantham chemical distance matrix (49).

2.2.10. Quasi-sequence-order (QSOrder)

For each amino acid, defined QSOrder as follows:

$$X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag} \tau_d}, r = 1, 2, 3, \ldots, 20 \quad (31)$$

where f_r represent the normalized occurrence of amino acid which is r typed, and the weighting factor w is defined as 0.1, and $nlag$ denotes the maximum value of the lag (default value: 30). τ_d is the same as the definition in SOCNumber.

For other 30 quasi-sequence-order descriptors, defined QSOrder as follows:

$$X_d = \frac{w\tau_d - 20}{\sum_{r=1}^{20} f_r + w\sum_{d=1}^{nlag} \tau_d}, d = 21, 22, \ldots, 20 + nlag \quad (32)$$

2.3. Random forest

RF algorithm is an ensemble of decision trees and has been widely used for classification. Each tree depends on the value of a random vector that is sampled independently and has the same distribution for all trees in the forest. The introduction of randomness can reduce the possibility of overfitting, improve the ability to resist noise, and has strong adaptability to high-dimensional data.

RF algorithm has been applied to a variety of protein classification problems (50–54).

2.4. Model evaluation metrics

To evaluate the training effect and prediction ability of the model, we mainly used the Area Under the Receiver Operating Characteristic curve value (AUROC), supplemented by Sensitivity (Sn), Specificity (Sp), Matthew's correlation coefficient (MCC), accuracy (ACC) (55–72). These indexes can be formulated as follows:

$$Sn = \frac{TP}{(TP + FN)} \quad (33)$$

$$Sp = \frac{TN}{(TN + FP)} \quad (34)$$

$$MCC = \frac{(TN \times TP - FN \times FP)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (35)$$

$$ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (36)$$

where TP and FN represent the number that the bitter peptides are predicted as true bitter peptides and non-bitter peptides, respectively. On the contrary, TN and FP represent the number that the non-bitter peptides are predicted as true non-bitter peptides and bitter peptides, respectively. That is to say, bitter peptides were defined as positive samples, and non-bitter peptides were defined as negative samples in this work.

TABLE 1 Results of RF-based models using 10 single features.

Cross-validation	Feature	Dimension	AUROC	Sn	Sp	Acc	Mcc
10-fold cross-validation	AAC	20	**0.91**	0.85	**0.84**	**0.85**	**0.69**
	TPAAC	21	0.90	0.83	0.78	0.80	0.61
	APAAC	22	0.89	0.83	0.81	0.82	0.64
	ASDC	400	0.88	**0.89**	0.68	0.79	0.59
	DPC	400	0.86	0.87	0.64	0.76	0.53
	DDE	400	0.83	0.84	0.73	0.78	0.57
	GAAC	5	0.75	0.72	0.66	0.69	0.39
	GDPC	25	0.78	0.75	0.71	0.73	0.46
	SOCNumber	2	0.70	0.66	0.62	0.64	0.28
	QSOrder	42	0.89	0.82	0.82	0.82	0.64
Independent set validation	AAC	20	0.96	0.91	0.89	**0.90**	**0.80**
	TPAAC	21	0.94	0.83	0.86	0.84	0.69
	APAAC	22	**0.97**	0.89	**0.91**	**0.90**	0.80
	ASDC	400	0.92	0.89	0.75	0.82	0.65
	CKSAAGP	100	0.87	0.77	0.81	0.79	0.58
	DPC	400	0.89	0.88	0.70	0.79	0.59
	DDE	400	0.90	0.89	0.84	0.87	0.74
	GAAC	5	0.76	0.83	0.64	0.73	0.48
	GDPC	25	0.80	0.73	0.72	0.73	0.45
	SOCNumber	2	0.73	0.59	0.69	0.64	0.28
	QSOrder	42	0.95	**0.92**	0.84	0.88	0.77

Best performance metrics are shown in bold.

TABLE 2 Features after feature reduction operation.

Feature	Dimension	Dimension after operation
AAC	20	20
TPAAC	21	21
APAAC	22	22
ASDC	400	366
DPC	400	303
DDE	400	400
GAAC	5	5
GDPC	25	25
SOCNumber	2	2
QSOrder	42	42
Total of features	1,337	1,206

Sn is the model's sensitivity, representing the proportion of correctly predicted positive samples to the total number of actual positive samples (73–76). Sp is the model's specificity, representing the proportion of correctly predicted negative samples to the total number of actual negative samples (77, 78). Here ACC, MCC and AUROC are all comprehensive indicators. ACC represents the proportion of correct predicted samples to the total samples. And MCC is the correlation coefficient between the description classification and the predicted classification. Its range is [-1, 1]. If the value is 1, it means the model prediction performance is perfect. If the value is -1, it means the prediction is completely opposite to the actual. The AUROC indicator can be used as a standard for evaluating the quality of the binary classification model (79–82). The closer the value of AUROC is to 1, the better the classification effect.

3. Results and discussion

3.1. Single-feature-based results

Here, we used iLearnPlus to extract the above 10 features (AAC, TPAAC, APAAC, ASDC, DPC, DDE, GAAC, GDPC, SOCNumber, QSOrder) and then utilized them to train a RF-based predictive model for accurately identifying Bitter peptides (37). **Table 1** shows the results of 10-fold cross-validation and independent set.

As can be seen, AAC is the best among all single features, with AUROC of 0.91 and 0.96 in 10-fold cross-validation and independent data test, while the worst was SOCNumber, with AUROC of 0.70 and 0.73. This result should show that SOCNumber has only two dimensions, so this feature cannot afford enough information. Thus, this feature may be used to fuse other features to supplement additional information.

Amino acid composition is only a basic feature and does not burden physicochemical properties. Therefore, we think that there is still a large space for optimization. Previous studies have shown the relationship between bitter peptides and factors such as amino acid hydrophobicity and amino acid position. Some single features with poor performance have rich information that AAC does not have and can improve prediction performance. Therefore, we will study how to optimize the parameters of characteristics in following section.

3.2. Fusion feature processing

By fusing the 10 features mentioned above, we will get a 1,337-dimensional fusion feature. In this step, we de-zero the fusion feature. When a column contains only zero, it has no practical effect on the discrimination and is removed. After deleting all zero columns, 1206 features remain, as shown in detail in **Table 2**.

3.3. Fusion-feature–based results

In this study, we compared the prediction effect of the fusion features and the three features with the highest independent set validation AUROC value among the above 10 single features. It has been proved that using the RF method to deal with fused features does have more advantages in terms of predictive ability. **Table 3** and **Figure 2** show the results of 10-fold cross-validation and independent set validation.

It could be seen that, in 10-fold cross-validation and independent set validation, the prediction performance of fusion features was improved or remained unchanged compared with single feature prediction. That is to say, the fusion features have better predictive ability.

3.4. Comparison with other machine learning methods on fusion features

To further validate the prediction model of the RF method for bitter peptides, we compared it with some traditional machine

TABLE 3 Comparison between single-features and fusion feature using RF algorithm.

ML method	Cross-validation	Feature	Dimension	AUROC	Sn	Sp	Acc	Mcc
Random Forest	10-fold cross-validation	AAC	20	0.91	0.85	**0.84**	**0.85**	0.69
		APAAC	22	0.89	0.83	0.81	0.82	0.64
		QSOrder	42	0.89	0.82	0.82	0.82	0.64
		Fusion	1206	**0.93**	**0.86**	**0.84**	**0.85**	**0.70**
	Independent set validation	AAC	20	0.96	0.91	0.89	0.90	0.80
		APAAC	22	0.97	0.89	0.91	0.90	0.80
		QSOrder	42	0.95	0.92	0.84	0.88	0.77
		Fusion	1206	**0.98**	**0.94**	**0.94**	**0.94**	**0.88**

Best performance metrics are shown in bold.

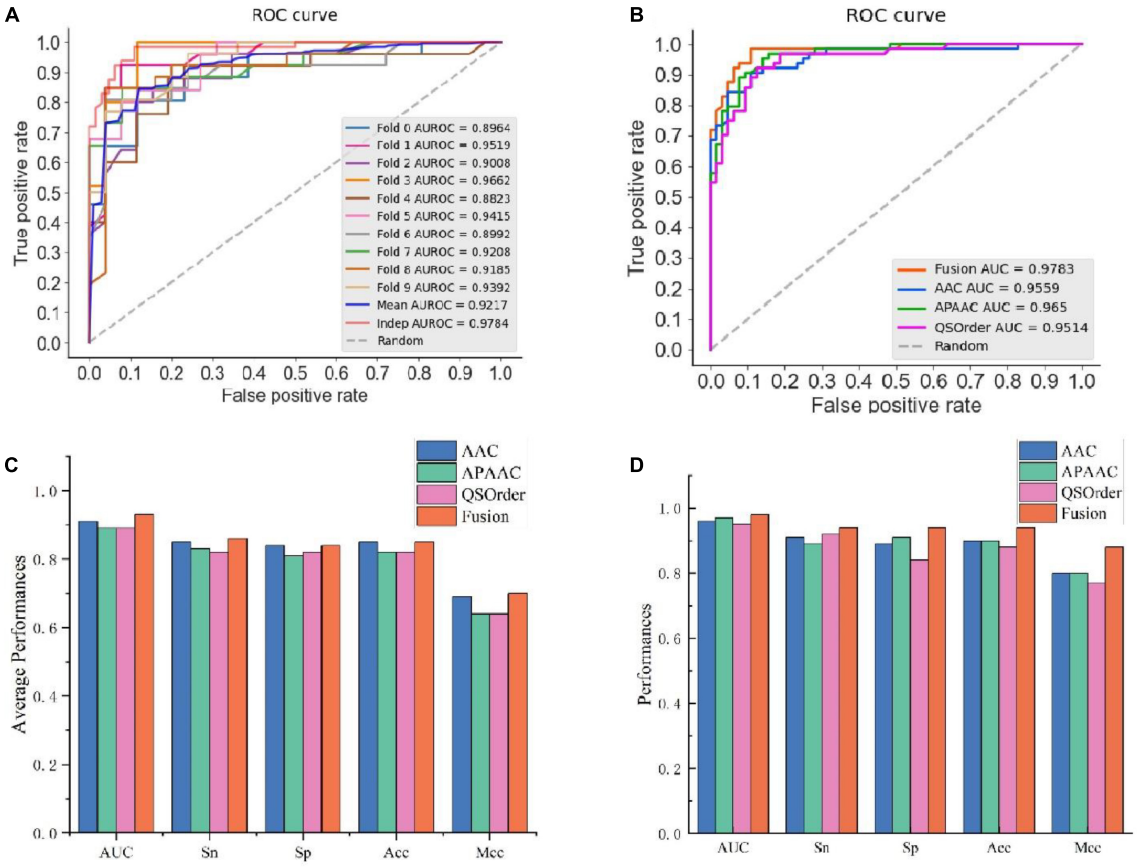


FIGURE 2
The prediction results using different features. **(A)** AUROC curves of fused features using RF; **(B)** AUROC curves of fusion features and three single-feature on independent data; **(C)** detailed results on training data using 10-fold cross-validation; **(D)** independent data validated results.

TABLE 4 Comparison of multiple machine learning methods using fusion features.

Cross-validation	Feature	ML method	AUROC	Sn	Sp	Acc	Mcc
10-fold cross-validation	Fusion	SVM	0.67	0.51	0.80	0.66	0.34
	Fusion	LightGBM	0.92	0.85	**0.85**	**0.85**	**0.70**
	Fusion	DT	0.80	0.83	0.77	0.80	0.60
	Fusion	LR	0.82	0.74	0.77	0.76	0.52
	Fusion	RF	**0.93**	**0.86**	0.84	**0.85**	**0.70**
Independent set validation	Fusion	SVM	0.74	0.61	0.78	0.70	0.40
	Fusion	LightGBM	0.97	0.92	0.91	0.91	0.83
	Fusion	DT	0.94	0.94	0.84	0.89	0.78
	Fusion	LR	0.89	0.80	0.84	0.82	0.64
	Fusion	RF	**0.98**	**0.94**	**0.94**	**0.94**	**0.88**

Best performance metrics are shown in bold.

learning methods. Here, Support Vector Machines (SVM), LightGBM, Decision Trees (DT), and Logistic Regression (LR) were selected to build models for comparison. The prediction results of each machine learning method are shown in Table 4 and Figure 3. It can be seen that the RF method is superior to or equal to other machine learning methods in various indicators, and has good learning effect and prediction ability. Therefore, according to the data characteristics provided by us, the RF method shows the best predictive ability.

3.5. Comparison with existed models

To evaluate the predictive ability of Bitter-RF, we compared it with the existing four sequence-based models. The first model is iBitter-SCM which was constructed based on the dipeptide propensity score, the second model is BERT4Bitter using deep learning method, the third model is iBitter-Fuse by combining fuses features with SVM, and the fourth model was iBitter-DRLF by selecting features through deep learning (22–24, 26). Here, Bitter-RF

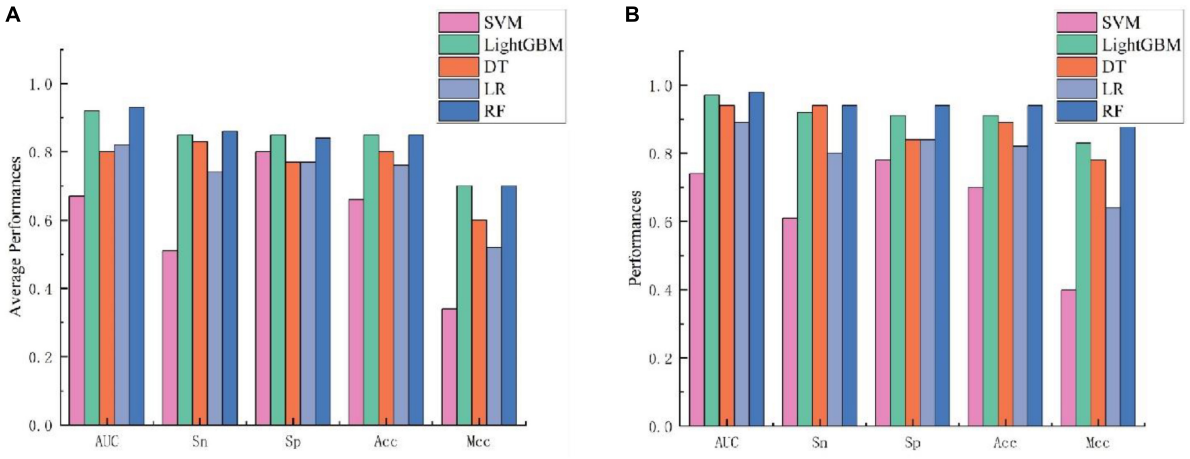


FIGURE 3
Performance evaluation of different machine learning models using **(A)** 10-fold cross-validation and **(B)** independent testing.

TABLE 5 Performance comparison of Bitter-RF with the existing methods.

Cross-validation	Classifier	AUROC	Sn	Sp	Acc	Mcc
10-fold cross-validation	iBitter-SCM	0.90	0.91	0.83	0.87	0.75
	BERT4Bitter	0.92	0.87	0.85	0.86	0.73
	iBitter-Fuse	0.94	**0.92**	**0.92**	**0.92**	**0.84**
	iBitter-DRLF	**0.95**	0.89	0.89	0.89	0.78
	Bitter-RF	0.93	0.86	0.84	0.85	0.70
Independent set validation	iBitter-SCM	0.90	0.84	0.84	0.84	0.69
	BERT4Bitter	0.96	**0.94**	0.91	0.92	0.84
	iBitter-Fuse	0.93	**0.94**	0.92	0.93	0.86
	iBitter-DRLF	**0.98**	0.92	**0.98**	**0.94**	**0.89**
	Bitter-RF	**0.98**	**0.94**	0.94	**0.94**	0.88

Best performance metrics are shown in bold.

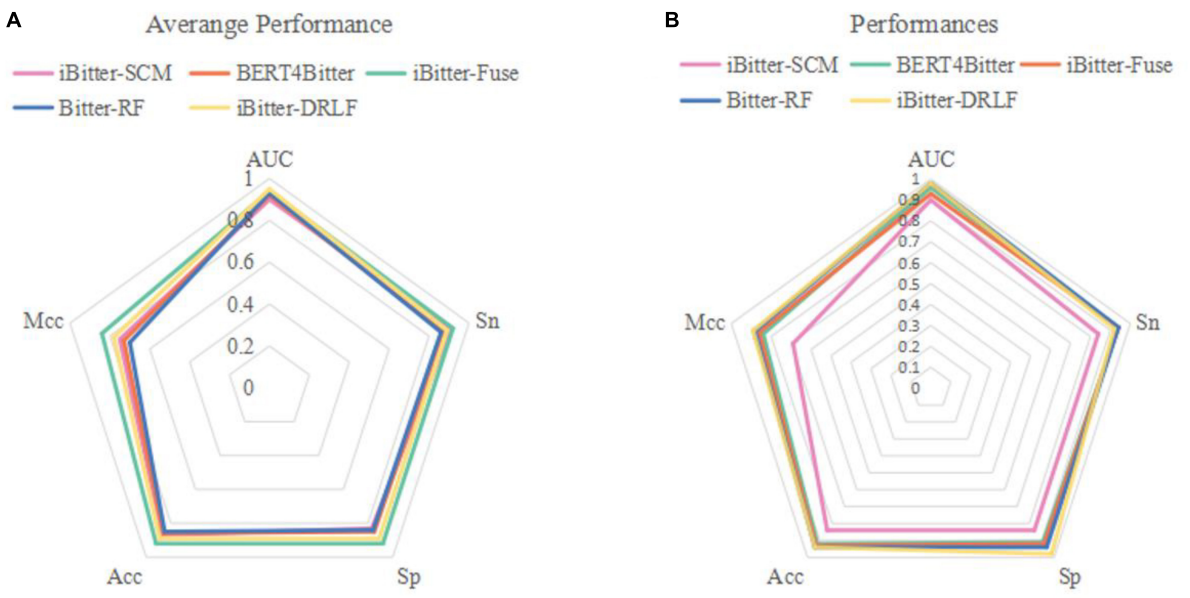


FIGURE 4
Radar plot for comparing Bitter-RF with other published models using **(A)** 10-fold cross-validation and **(B)** independent testing.

model used the same bitter peptide and non-bitter peptide sequences as the previous four models. We further extended the types of extracted features on the basis of the third model, and used the RF method for modeling. By referring to relevant literatures, we obtained the performance indicators of the four models. The comparison results have been shown in **Table 5** and **Figure 4**.

The performance comparison between Bitter-RF model and the four models showed that the results of Bitter-RF model in 10-fold cross-validation are similar to BERT4Bitter, and slightly lower than iBitter-Fuse. However, the results of Bitter-RF model on independent data are generally better than those of the first three models, and are comparable to those of the fourth model. Bitter-RF model has the same *Sn* index as the previous two generation models, which is superior to the first generation model. The indexes of *Sp*, *ACC* and *MCC* are better than those of the previous three generations. Furthermore, the AUROC of Bitter-RF model is 5% higher than that of iBitter-Fuse. Although the prediction performance of Bitter-RF is close to that of iBitter-DRLF, we used a traditional machine learning method, which consumes less computing resources. To sum up, Bitter-RF model shows stronger prediction performance and better practical application ability.

To our knowledge, we could not find any alternative bitterness classification studies allowing us to assess the intrinsic robustness of the bitter/non-bitter classification and therefore it cannot be excluded that the model may be affected by the inherent bias of training/test set data.

4. Conclusion

Compared with other proteins, there is still much room for related research on bitter peptides, and it has shown potential medical benefits. To better study bitter peptides, we developed a novel model Bitter-RF for predicting bitter peptides, which uses information from multiple perspectives, including sequence internal information and physicochemical properties. By comparison, we concluded that fused features could produce better performance than single features, RF is more suitable for bitter peptide prediction, and Bitter-RF has more application advantages than the four published models. Our research further enriches the application of RF method in the field of protein classification. And Bitter-RF model's better results also show that enrich physical and chemical properties, location information and other characteristics play an important role in the identification of bitter peptides, which can provide biologists with more directions for biological experiments to verify bitter peptides.

However, one may notice that the features were not optimized. In the future, we will use various of feature selection techniques (83–86) to pick out the best features for improving model's performance.

Based on the proposed method, a free and easy-to-use python package has been built and accessible at GitHub: https://github.com/ZhangYufei01/Bitter-RF.git, which can help scholars to identify bitter peptides.

Data availability statement

The original contributions presented in this study are included in this article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

HD, YZ, and K-JD conceived and designed the study. Y-FZ and Y-HW conducted the experiments and implemented the algorithms. Z-FG, X-RP, and JL performed the analysis. Y-FZ, JL, HD, YZ, and K-JD wrote the manuscript. JL, HD, YZ, and K-JD reviewed and edited the manuscript. HD and K-JD supervised the study. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the National Nature Scientific Foundation of China (81872957 and 62202069) and Natural Science Foundation of Sichuan Province (2022NSFSC1610).

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.

References

1. Xu B, Chung H. Quantitative structure-activity relationship study of bitter di-, tri- and tetrapeptides using integrated descriptors. *Molecules.* (2019) 24:2846. doi: 10.3390/molecules24152846

2. Kim H, Li-Chan E. Quantitative structure-activity relationship study of bitter peptides. *J Agric Food Chem.* (2006) 54:10102–11. doi: 10.1021/jf062422j

3. Maehashi K, Huang L. Bitter peptides and bitter taste receptors. *Cell Mol Life Sci.* (2009) 66:1661–71. doi: 10.1007/s00018-009-8755-9

4. Calabrese E, Baldwin L. Toxicology rethinks its central belief. *Nature.* (2003) 421:691–2. doi: 10.1038/421691a

5. Lo H, Li C, Ho T, Hsiang C. Identification of the bioactive and consensus peptide motif from *Momordica charantia* insulin receptor-binding protein. *Food Chem.* (2016) 204:298–305. doi: 10.1016/j.foodchem.2016.02.135

6. Hsu P, Pan F, Hsieh C. mcIRBP-19 of bitter melon peptide effectively regulates diabetes mellitus (dm) patients' blood sugar levels. *Nutrients.* (2020) 12:1252. doi: 10.3390/nu12051252

7. Abu Bakar N, Hakim Abdullah M, Lim V, Yong Y. Gastroprotective effect of polypeptide-K Isolated from *Momordica charantia's* seeds on multiple experimental gastric ulcer models in rats. *Evid Based Complement Alternat Med.* (2022) 2022:6098929. doi: 10.1155/2022/6098929

8. Ning L, Abagna H, Jiang Q, Liu S, Huang J. Development and application of therapeutic antibodies against covid-19. *Int J Biol Sci.* (2021) 17:1486–96. doi: 10.7150/ijbs.59149

9. Van Der Ven C, Muresan S, Gruppen H, De Bont D, Merck K, Voragen A. FTIR spectra of whey and casein hydrolysates in relation to their functional properties. *J Agric Food Chem.* (2002) 50:6943–50. doi: 10.1021/jf020387k

10. Kim H, Li-Chan E. Application of fourier transform Raman spectroscopy for prediction of bitterness of peptides. *Appl Spectrosc.* (2006) 60:1297–306. doi: 10.1366/000370206778998978

11. Karametsi K, Kokkinidou S, Ronningen I, Peterson D. Identification of bitter peptides in aged cheddar cheese. *J Agric Food Chem.* (2014) 62:8034–41. doi: 10.1021/jf5020654

12. Liu X, Jiang D, Peterson D. Identification of bitter peptides in whey protein hydrolysate. *J Agric Food Chem.* (2014) 62:5719–25. doi: 10.1021/jf4019728

13. Gauthaman A, Jacob R, Pasupati S, Rajadurai A, Doss C, Moorthy A. Novel peptide-based inhibitor for targeted inhibition of T cell function. *J Cell Commun Signal.* (2022) 16:349–59. doi: 10.1007/s12079-021-00660-0

14. Tayubi I, Kumar S, Doss C. Identification of potential inhibitors, conformational dynamics, and mechanistic insights into mutant Kirsten rat sarcoma virus (G13d) driven cancers. *J Cell Biochem.* (2022) 123:1467–80. doi: 10.1002/jcb.30305

15. Wu J, Aluko R. Quantitative structure-activity relationship study of bitter di- and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity. *J Pept Sci.* (2007) 13:63–9. doi: 10.1002/psc.800

16. Soltani S, Haghaei H, Shayanfar A, Vallipour J, Asadpour Zeynali K, Jouyban A. QSBR study of bitter taste of peptides: application of Ga-Pls in combination with Mlr, Svm, and Ann approaches. *Biomed Res Int.* (2013) 2013:501310. doi: 10.1155/2013/501310

17. Lv Z, Ao C, Zou Q. Protein function prediction: from traditional classifier to deep learning. *Proteomics.* (2019) 19:e1900119. doi: 10.1002/pmic.201900119

18. Lv Z, Cui F, Zou Q, Zhang L, Xu L. Anticancer peptides prediction with deep representation learning features. *Brief Bioinform.* (2021) 22:bbab008. doi: 10.1093/bib/bbab008

19. Ao C, Yu L, Zou Q. Prediction of bio-sequence modifications and the associations with diseases. *Brief Funct Genom.* (2021) 20:1–18. doi: 10.1093/bfgp/elaa023

20. Zhang Y, Liu T, Hu X, Wang M, Wang J, Zou B, et al. Cellcall: integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Res.* (2021) 49:8520–34. doi: 10.1093/nar/gkab638

21. Zhang Y, Liu T, Wang J, Zou B, Li L, Yao L, et al. Cellinker: a platform of ligand-receptor interactions for intercellular communication analysis. *Bioinformatics.* (2021) 37:2025–32. doi: 10.1093/bioinformatics/btab036

22. Charoenkwan P, Yana J, Schaduangrat N, Nantasenamat C, Hasan M, Shoombuatong W. iBITTER-SCM: identification and characterization of bitter peptides using a scoring card method with propensity scores of dipeptides. *Genomics.* (2020) 112:2813–22. doi: 10.1016/j.ygeno.2020.03.019

23. Charoenkwan P, Nantasenamat C, Hasan M, Manavalan B, Shoombuatong W. BERT4Bitter: a bidirectional encoder representations from transformers (bert)-based model for improving the prediction of bitter peptides. *Bioinformatics.* (2021) 37:2556–62. doi: 10.1093/bioinformatics/btab133

24. Charoenkwan P, Nantasenamat C, Hasan M, Moni M, Lio P, Shoombuatong W. iBitter-Fuse: a novel sequence-based bitter peptide predictor by fusing multi-view features. *Int J Mol Sci.* (2021) 22:8958. doi: 10.3390/ijms22168958

25. Yan N, Lv Z, Hong W, Xu X. Editorial: feature representation and learning methods with applications in protein secondary structure. *Front Bioeng Biotechnol.* (2021) 9:748722. doi: 10.3389/fbioe.2021.748722

26. Jiang J, Lin X, Jiang Y, Jiang L, Lv Z. Identify bitter peptides by using deep representation learning features. *Int J Mol Sci.* (2022) 23:7877. doi: 10.3390/ijms23147877

27. Zhao-Yue ZZ, Yu-He Y, Hao L. Towards a better prediction of subcellular location of long non-coding RNA. *Front Comput Sci.* (2022) 16:165903. doi: 10.1007/s11704-021-1015-3

28. Yang H, Luo Y, Ren X, Wu M, He X, Peng B, et al. Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Inform Fus.* (2021) 75:140–9. doi: 10.1016/j.inffus.2021.02.015

29. Hasan M, Basith S, Khatun M, Lee G, Manavalan B, Kurata H. Meta-I6ma: an interspecies predictor for identifying DNA N6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework. *Brief Bioinform.* (2021) 22:bbaa202. doi: 10.1093/bib/bbaa202

30. Wu X, Yu L. Epsol: sequence-based protein solubility prediction using multidimensional embedding. *Bioinformatics.* (2021) 37:4314–20. doi: 10.1093/bioinformatics/btab463

31. Jeon Y, Hasan M, Park H, Lee K, Manavalan B. Tacos: a novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization. *Brief Bioinform.* (2022) 23:bbac243. doi: 10.1093/bib/bbac243

32. Ao C, Zou Q, Yu L. NMRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences. *Brief Bioinform.* (2022) 23:bbab480. doi: 10.1093/bib/bbab480

33. Su R, Liu X, Wei L. MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy. *Brief Bioinform.* (2020) 21:687–98. doi: 10.1093/bib/bbz021

34. Wei L, Zhou C, Chen H, Song J, Su R. ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics.* (2018) 34:4007–16. doi: 10.1093/bioinformatics/bty451

35. Teng Z, Zhang Z, Tian Z, Li Y, Wang G. ReRF-Pred: predicting amyloidogenic regions of proteins based on their pseudo amino acid composition and tripeptide composition. *BMC Bioinform.* (2021) 22:545. doi: 10.1186/s12859-021-04446-4

36. Li H, Shi L, Gao W, Zhang Z, Zhang L, Zhao Y, et al. Dpromoter-Xgboost: detecting promoters and strength by combining multiple descriptors and feature selection using Xgboost. *Methods.* (2022) 204:215–22. doi: 10.1016/j.ymeth.2022.01.001

37. Chen Z, Zhao P, Li C, Li F, Xiang D, Chen Y, et al. Ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic Acids Res.* (2021) 49:e60. doi: 10.1093/nar/gkab122

38. Ahmed Z, Zulfiqar H, Tang L, Lin H. A statistical analysis of the sequence and structure of thermophilic and non-thermophilic proteins. *Int J Mol Sci.* (2022) 23:10116. doi: 10.3390/ijms231710116

39. Hasan M, Schaduangrat N, Basith S, Lee G, Shoombuatong W, Manavalan B. Hlppred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation. *Bioinformatics.* (2020) 36:3350–6. doi: 10.1093/bioinformatics/btaa160

40. Basith S, Lee G, Manavalan B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. *Brief Bioinform.* (2022) 23:bbab376. doi: 10.1093/bib/bbab376

41. Zhao X, Wang H, Li H, Wu Y, Wang G. Identifying plant pentatricopeptide repeat proteins using a variable selection method. *Front Plant Sci.* (2021) 12:506681. doi: 10.3389/fpls.2021.506681

42. Zhai Y, Chen Y, Teng Z, Zhao Y. Identifying antioxidant proteins by using amino acid composition and protein-protein interactions. *Front Cell Dev Biol.* (2020) 8:591487. doi: 10.3389/fcell.2020.591487

43. Chou K. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins.* (2001) 43:246–55. doi: 10.1002/prot.1035

44. Damborsky J. Quantitative structure-function and structure-stability relationships of purposely modified proteins. *Protein Eng.* (1998) 11:21–30. doi: 10.1093/protein/11.1.21

45. Hopp T, Woods K. Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci U.S.A.* (1981) 78:3824–8. doi: 10.1073/pnas.78.6.3824

46. Chou K. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics.* (2005) 21:10–9. doi: 10.1093/bioinformatics/bth466

47. Tang H, Zhao Y, Zou P, Zhang C, Chen R, Huang P, et al. HBPred: a tool to identify growth hormone-binding proteins. *Int J Biol Sci.* (2018) 14:957–64. doi: 10.7150/ijbs.24174

48. Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys J.* (1994) 66(2 Pt 1):335–44. doi: 10.1016/s0006-349580782-9

49. Grantham R. Amino acid difference formula to help explain protein evolution. *Science.* (1974) 185:862–4. doi: 10.1126/science.185.4154.862

50. Manavalan B, Patra M. MLCPP 2.0: an updated cell-penetrating peptides and their uptake efficiency predictor. *J Mol Biol.* (2022) 434:167604. doi: 10.1016/j.jmb.2022.167604

51. Thi Phan L, Woo Park H, Pitti T, Madhavan T, Jeon Y, Manavalan B. MLACP 2.0: an updated machine learning tool for anticancer peptide prediction. *Comput Struct Biotechnol J.* (2022) 20:4473–80. doi: 10.1016/j.csbj.2022.07.043

52. Lv Z, Jin S, Ding H, Zou Q. A random forest sub-golgi protein classifier optimized via dipeptide and amino acid composition features. *Front Bioeng Biotechnol.* (2019) 7:215. doi: 10.3389/fbioe.2019.00215

53. Lv Z, Zhang J, Ding H, Zou Q. RF-PseU: a random forest predictor for RNA pseudouridine sites. *Front Bioeng Biotechnol.* (2020) 8:134. doi: 10.3389/fbioe.2020.00134

54. Ao C, Zou Q, Yu L. RFhy-m2G: identification of RNA N2-methylguanosine modification sites based on random forest and hybrid features. *Methods.* (2022) 203:32–9. doi: 10.1016/j.ymeth.2021.05.016

55. Lv H, Dao F, Lin H. DeepKla: an attention mechanism-based deep neural network for protein lysine lactylation site prediction. *iMeta.* (2022) 1:e11. doi: 10.1002/imt2.11

56. Han Y, Yang H, Huang Q, Sun Z, Li M, Zhang J, et al. Risk prediction of diabetes and pre-diabetes based on physical examination data. *Math Biosci Eng.* (2022) 19:3597–608. doi: 10.3934/mbe.2022166

57. Akbar S, Ahmad A, Hayat M, Rehman A, Khan S, Ali F. iAtbP-Hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. *Comput Biol Med.* (2021) 137:104778. doi: 10.1016/j.compbiomed.2021.104778

58. Dong F, Zhao G, Tong H, Zhang Z, Lao X, Zheng H. The prospect of bioactive peptide research: a review on databases and tools. *Curr Bioinform.* (2021) 16:494–504. doi: 10.2174/1574893615999200813192148

59. Jagadeb M, Pattanaik K, Rath S, Sonawane A. Identification and evaluation of immunogenic Mhc-I and Mhc-Ii binding peptides from mycobacterium tuberculosis. *Comput Biol Med.* (2021) 130:104203. doi: 10.1016/j.compbiomed.2020.104203

60. Lin D, Yu J, Zhang J, He H, Guo X, Shi S. Predaip: computational prediction and analysis for anti-inflammatory peptide via a hybrid feature selection technique. *Curr Bioinform.* (2021) 16:1048–59. doi: 10.2174/1574893616666210601111157

61. Liu Y, Ouyang X, Xiao Z, Zhang L, Cao Y. A review on the methods of peptide-Mhc binding prediction. *Curr Bioinform.* (2020) 15:878–88. doi: 10.2174/1574893615999200429122801

62. Masoudi-Sobhanzadeh Y, Jafari B, Parvizpour S, Pourseif M, Omidi YA. Novel multi-objective metaheuristic algorithm for protein-peptide docking and benchmarking on the leads-pep dataset. *Comput Biol Med.* (2021) 138:104896. doi: 10.1016/j.compbiomed.2021.104896

63. Mulpuru V, Semwal R, Varadwaj P, Mishra N. Hamp: a knowledgebase of antimicrobial peptides from human microbiome. *Curr Bioinform.* (2021) 16:534–40. doi: 10.2174/1574893615999200802041228

64. Yu L, Wang M, Yang Y, Xu F, Zhang X, Xie F, et al. Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways. *PLoS Comput Biol.* (2021) 17:e1008696. doi: 10.1371/journal.pcbi.1008696

65. Wei L, Su R, Wang B, Li X, Zou Q, Gao X. Integration of deep feature representations and handcrafted features to improve the prediction of N-6-methyladenosine sites. *Neurocomputing.* (2019) 324:3–9. doi: 10.1016/j.neucom.2018.04.082

66. Wei L, Tang J, Zou Q. Local-Dpp: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform Sci.* (2017) 384:135–44. doi: 10.1016/j.ins.2016.06.026

67. Wang X, Yang Y, Liu J, Wang G. The stacking strategy-based hybrid framework for identifying non-coding RNAs. *Brief Bioinform.* (2021) 22:bbab023. doi: 10.1093/bib/bbab023

68. Tao Z, Li Y, Teng Z, Zhao Y. A method for identifying vesicle transport proteins based on Libsvm and Mrmd. *Comput Math Methods Med.* (2020) 2020:8926750. doi: 10.1155/2020/8926750

69. Guo Z, Wang P, Liu Z, Zhao Y. Discrimination of thermophilic proteins and non-thermophilic proteins using feature dimension reduction. *Front Bioeng Biotechnol.* (2020) 8:584807. doi: 10.3389/fbioe.2020.584807

70. Jiang Q, Wang G, Jin S, Li Y, Wang Y. Predicting human microrna-disease associations based on support vector machine. *Int J Data Min Bioinform.* (2013) 8:282–93.

71. Huang Y, Zhou D, Wang Y, Zhang X, Su M, Wang C, et al. Prediction of transcription factors binding events based on epigenetic modifications in different human cells. *Epigenomics.* (2020) 12:1443–56. doi: 10.2217/epi-2019-0321

72. Xu Z, Luo M, Lin W, Xue G, Wang P, Jin X, et al. DLpTCR: an ensemble deep learning framework for predicting immunogenic peptide recognized by T cell receptor. *Brief Bioinform.* (2021) 22:bbab335. doi: 10.1093/bib/bbab335

73. Lv H, Dao F, Guan Z, Yang H, Li Y, Lin H. Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Brief Bioinform.* (2021) 22:bbaa255. doi: 10.1093/bib/bbaa255

74. Dao F, Lv H, Zhang D, Zhang Z, Liu L, Lin H. Deepyy1: a deep learning approach to identify Yy1-mediated chromatin loops. *Brief Bioinform.* (2021) 22:bbaa356. doi: 10.1093/bib/bbaa356

75. Dao F, Lv H, Su W, Sun Z, Huang Q, Lin H. Idhs-Deep: an integrated tool for predicting Dnase I hypersensitive sites by deep neural network. *Brief Bioinform.* (2021) 22:bbab047. doi: 10.1093/bib/bbab047

76. Zhang D, Xu Z, Su W, Yang Y, Lv H, Yang H, et al. Icarps: a computational tool for identifying protein carbonylation sites by novel encoded features. *Bioinformatics.* (2020) 37:171–7. doi: 10.1093/bioinformatics/btaa702

77. Zhang L, Yang Y, Chai L, Li Q, Liu J, Lin H, et al. A deep learning model to identify gene expression level using cobinding transcription factor signals. *Brief Bioinform.* (2022) 23:bbab501. doi: 10.1093/bib/bbab501

78. Lv H, Zhang Y, Wang J, Yuan S, Sun Z, Dao F, et al. Irice-Ms: an integrated Xgboost model for detecting multitype post-translational modification sites in rice. *Brief Bioinform.* (2022) 23:bbab486. doi: 10.1093/bib/bbab486

79. Zhang Q, Li H, Liu Y, Li J, Wu C, Tang H. Exosomal non-coding RNAs: new insights into the biology of hepatocellular carcinoma. *Curr Oncol.* (2022) 29:5383–406.

80. Sun Z, Huang Q, Yang Y, Li S, Lv H, Zhang Y, et al. Psnod: identifying potential snorna-disease associations based on bounded nuclear norm regularization. *Brief Bioinform.* (2022) 23:bbac240. doi: 10.1093/bib/bbac240

81. Dao F, Lv H, Zhang Z, Lin H. Bdselect: a package for K-Mer selection based on the binomial distribution. *Curr Bioinform.* (2022) 17:238–44. doi: 10.2174/1574893616666211007102747

82. Yu L, Xia M, An Q. A network embedding framework based on integrating multiplex network for drug combination prediction. *Brief Bioinform.* (2022) 23:bbab364. doi: 10.1093/bib/bbab364

83. Huang H, Wu N, Liang Y, Peng X, Shu J. Slnl: a novel method for gene selection and phenotype classification. *Int J Intell Syst.* (2022) 37:6283–304. doi: 10.1002/int.22844

84. Huang H, Liang Y. A novel cox proportional hazards model for high-dimensional genomic data in cancer prognosis. *IEEE/ACM Trans Comput Biol Bioinform.* (2021) 18:1821–30. doi: 10.1109/TCBB.2019.2961667

85. Huang H, Peng X, Liang Y. Splsn: an efficient tool for survival analysis and biomarker selection. *Int J Intell Syst.* (2021) 36:5845–65. doi: 10.1002/int.22532

86. Huang H, Rao H, Miao R, Liang Y. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. *BMC Bioinform.* (2022) 23(Suppl. 10):353. doi: 10.1186/s12859-022-04887-5

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 01 February 2023
DOI 10.3389/fmed.2023.1051039

OPEN ACCESS

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

REVIEWED BY
Aliyah Almomen,
King Saud University,
Saudi Arabia
Mahran Abdel-Rahman,
Assiut University,
Egypt
Aisha El-Turki,
Brighton and Sussex Medical School,
United Kingdom

*CORRESPONDENCE
Noor B Almandil
✉ nbalmandil@iau.edu.sa

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

RECEIVED 22 September 2022
ACCEPTED 13 January 2023
PUBLISHED 01 February 2023

CITATION
Almandil NB, Alismail MA, Alsuwat HS, AlSulaiman A, AbdulAzeez S and Borgio JF (2023) Exome-wide analysis identify multiple variations in olfactory receptor genes (*OR12D2* and *OR5V1*) associated with autism spectrum disorder in Saudi females. *Front. Med.* 10:1051039.
doi: 10.3389/fmed.2023.1051039

Exome-wide analysis identify multiple variations in olfactory receptor genes (*OR12D2* and *OR5V1*) associated with autism spectrum disorder in Saudi females

Noor B. Almandil[1]*, Maram Adnan Alismail[2,3], Hind Saleh Alsuwat[2], Abdulla AlSulaiman[4], Sayed AbdulAzeez[2] and J. Francis Borgio[2]

[1]Department of Clinical Pharmacy Research, Institute for Research and Medical Consultations (IRMC), Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, [2]Department of Genetic Research, Institute for Research and Medical Consultations (IRMC), Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, [3]College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia, [4]Department of Neurology, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia

Background: Autism Spectrum Disorder (ASD) is a multifactorial, neurodevelopmental disorder, characterized by deficits in communication, restricted and repetitive behaviors. ASD is highly heritable in Saudi Arabia; indecencies of affected individuals are increasing.

Objectives: To identify the most significant genes and SNPs associated with the increased risk of ASD in Saudi females to give an insight for early diagnosis.

Methods: Pilot case–control study mostly emphasized on the significant SNPs and haplotypes contributing to Saudi females with ASD patients (n=22) compared to controls (n=51) without ASD. With the use of allelic association analysis tools, 243,345 SNPs were studied systematically and classified according to their significant association. The significant SNPs and their genes were selected for further investigation for mapping of ASD candidate causal variants and functional impact.

Results: In females, five risk SNPs at $p \leq 2.32\times10^{-05}$ was identified in association with autism. The most significant exonic variants at chromosome 6p22.1 with olfactory receptor genes (*OR12D2* and *OR5V1*) clustered with high linkage disequilibrium through haplotyping analysis. Comparison between highly associated genes (56 genes) of male and female autistic patients with female autistic samples revealed that 39 genes are unique biomarkers for Saudi females with ASD.

Conclusion: Multiple variations in olfactory receptor genes (*OR5V1* and *OR12D2*) and single variations on *SPHK1, PLCL2, AKAP9* and *LOC107984893* genes are contributing to ASD in females of Arab origin. Accumulation of these multiple predisposed coding SNPs can increase the possibility of developing ASD in Saudi females.

KEYWORDS

autism spectrum disorder, Saudi females, coding variants, single nucleotide polymorphism, haplotyping

1. Introduction

Autism Spectrum Disorder (ASD) is a range of neurodevelopmental and neuropsychiatric disorders that start appearing from early childhood and lasts throughout the person's life (1, 2). Autism is among the most heritable and severe form of ASD (3), characterized by deficits in communication as well as repetitive and restricted behaviors as reported in *the Diagnostic and Statistical Manual, Fifth Edition (DSM-5)* (4). The World Health Organization (WHO) revealed statistics of 1 in 160 children to be diagnosed with ASD (5). In 2017, the latest ASD statistics in Saudi Arabia revealed that one per 167 individual is affected by autism (6). ASD is commonly multifactorial and many studies suggested interactions between immunological, neurological, environmental and genetic factors (7, 8). Tremendous sex bias of ASD shows that males are more affected than females with a male to female ratio of 3:1 (9). Several studies investigated the genetic risk factors against ASD in females, yet the key factors remain unknown (10–12). This paper identifies some genetic variables susceptible to cause autism in Saudi females, addresses the correlation between some diseases and pathways and genetic variants in Saudi females.

2. Methodology

2.1. Sample collection

The present study is conducted in accordance with the Declaration of Helsinki and received approval from the Institutional Review Board (IRB) of Imam Abdulrahman Bin Faisal University (IRB-2016-13-152). Out of 73 female age matched samples were included, 22 were cases and 51 were controls (Table 1). The present study sheds light on potential genetic contributors to autism in Saudi female subjects. Buccal cell samples were collected from the study subjects upon receiving the signed informed consent. All the samples were collected from the King Fahad Hospital of the University, Al Khobar, Saudi Arabia.

2.2. DNA extraction and genotyping

To extract DNA from buccal cell samples, the Gentra Puregene Buccal Cell Kit (Qiagen, Hilden, Germany) was used. The buccal cells were collected by scraping the inside of the mouth 10 times with given sterile brush. DNA was extracted within 3 h from the collection, 300 µl Cell lysis solution was dispensed in a 1.5 ml tube, incubated at 65°C for 15 min, and 1.5 µl of proteinase K was added and incubated at 55°C for 60 min. Then we added 100 µl protein precipitation reagent and incubated for 5 min on ice and centrifuged (13,000–16,000 ×*g* for 3 min). Supernatant was mixed with 300 µl isopropanol and 0.5 µl glycogen, centrifuged for 5 min at 13,000–16,000 ×*g*. The supernatant was discarded, the DNA pellet was washed with 70% ethanol and suspended the DNA in TE buffer. The Human Exome Bead Chip Kit v1.0 and v1.1 Illumina (San Diego, CA, United States), which is constituted of 243,345 putative functional exonic markers, was used with Illumina iScan for the microarray genotyping. DNA processing was performed in accordance with the manufacturer's protocol and all genotyping data were obtained from iScan control software (Illumina). DNA extraction and microarray genotyping and analysis took place in the genetics research laboratory of the Institute for Research and Medical Consultation, Imam Abdulrahman Bin Faial University, Dammam, Saudi Arabia. The procedures were executed between 2016 and 2019. The Infinium HTS workflow is a rapid 3 days work flow, in brief: The PicoGreen dsDNA quantification reagent was used to quantify double-stranded DNA samples. The quantified DNA samples were processed in 96 well plates. The quantified DNA samples were denatured and neutralized to prepare them for amplification. All the DNA samples were incubated uniformly to amplify, to generate a sufficient quantity of each individual DNA sample to be used in the Infinium HTS Assay. We Incubated the MSA3 plate with amplified DNA in the Illumina hybridization oven for 20–24 h at 37°C. Then to fragment the DNA, an endpoint fragmentation was used. A 100% 2-propanol and precipitation reagents were used to precipitate the DNA. Then, re-suspended the precipitated and fragmented DNA. The re-suspended DNA was dispensed onto bead chips and incubated for hybridization of each DNA sample to specific section of the bead chip. Afterwards, the bead chips were prepared for the staining process. Then the un-hybridized and non-specifically hybridized DNA samples were washed from the bead chips, added labeled nucleotides to extend primers hybridized to the sample, and stains the primers. For imaging the bead chip we followed the instructions in the System Guide for instrument to scan. Intensity files from iScan of the individual DNA samples from the exome chip were to perform the genotyping. Sample sheets with sample information, such as plate ID, cell ID, gender and so on were used for fetching the data from intensity files to perform the genotyping using GenomeStudio 2,0 software (Illumina.

TABLE 1 Characteristics of Saudi female patients with autism and controls without autism.

Parameter	Control group n =51	Case group n =22	Value of p
Age (year)	7.73 ± 3.13	7.09 ± 3.93	0.2251
Weight (kg)	30.01 ± 13.61	25.11 ± 11.92	0.1034
Height (cm)	121.38 ± 23.33	119.2 ± 20.69	0.3707
Body mass index	19.95 ± 4.97	16.70 ± 2.03	0.0033*

The data are presented as the mean values ± standard deviations. *Significant at $p \leq 0.05$.

2.3. Statistical and functional analysis

Initial quality check of call rate was fulfilled using GenomeStudio 2,0 software (Illumina). Only one control was eliminated from the analysis due to a call rate of < 0.98% and remaining samples were re-clustered. Using the Chi-square test with 1 degree of freedom (df), Hardy–Weinberg equilibrium (HWE) was tested individually for all the variants. Reference SNP ID numbers and gene names were acquired from SNP-Nexus (13) and Kaviar (14). To assess the outcomes of different alleles and haplotypes, 95% confidence interval, odds ratios and case–control association analyses were calculated using gPlink version 2.050 (15) and Haploview version 4.2 (16). The p values < 0.001 were regarded as significant. DAVID 6.7 (17) and Enricher (18) were utilized to annotate the highly significant remarkable ($p < 1 \times 10^{-05}$) genes for functional implications.

3. Results

Genotyping (Illumina) data were submitted to the NCBI (National Center for Biotechnology Information) Gene Expression Omnibus (GEO) repository [GEO accession number: GSE221098; BioProject accession numbers: PRJNA912746; GEO accession numbers for

10.3389/fmed.2023.1051039

TABLE 2 The most significant SNPs associated with autism in Saudi females.

S.NO	CHR	SNP ID	BP	MA	MAF	Gene	AA	Value of *p*	CHISQ	OR (L95-U95)	Case, control frequencies	HWpval
1	17	rs2247856	74,381,555	A	0.247	*SPHK1*	A	3.07×10^{-06}	21.77	6.28(2.77–14.23)	0.500, 0.137	0.0017
2	16	rs386789496	17,988,303	A	0.473	*LOC107984893*	A	1.04×10^{-05}	19.44	5.5(2.48–12.17)	0.750, 0.353	0.0117
3	3	rs4602367	17,053,499	A	0.336	*PLCL2*	A	1.78×10^{-05}	18.41	4.96(2.32–10.6)	0.591, 0.225	0.2088
4	7	rs6960867	91,712,698	G	0.397	*AKAP9*	G	2.17×10^{-05}	18.03	4.86(2.28–10.38)	0.659, 0.284	0.588
5	1	rs12035482	195,738,953	A	0.493	none	G	2.32×10^{-05}	17.91	0.18(0.08–0.42)	0.773, 0.390	0.0717
6	19	rs7507442	53,278,953	G	0.486	*ZNF600*	G	2.83×10^{-05}	17.53	5.05(2.28–11.15)	0.750, 0.373	0.0396
7	7	rs6964587	91,630,620	A	0.403	*AKAP9*	T	3.43×10^{-05}	17.17	4.8(2.22–10.37)	0.667, 0.294	0.3397
8	5	rs160632	96,503,523	G	0.445	*RIOK2*	C	3.46×10^{-05}	17.15	4.76(2.21–10.27)	0.705, 0.333	0.1332
9	3	rs9854207	27,614,316	C	0.363	none	C	3.53×10^{-05}	17.11	4.64(2.18–9.85)	0.614, 0.255	0.1288
10	19	rs142920057	334,472	C	0.121	*MIER2*	G	4.29×10^{-05}	16.74	8.143(2.64–25.09)	0.300, 0.050	0.6628
11	6	rs2073149	29,365,423	A	0.493	*OR5V1*	A	4.30×10^{-05}	16.74	4.89(2.21–10.82)	0.750, 0.380	0.3153
12	4	rs1339	154,631,563	G	0.197	*RNF175*	C	5.60×10^{-05}	16.23	5.50(2.28–13.25)	0.405, 0.110	0.5161
13	7	rs10488360	4,411,209	A	0.452	none	A	5.67×10^{-05}	16.21	4.56(2.12–9.81)	0.705, 0.343	0.4184
14	5	rs409045	34,628,627	G	0.37	none	C	6.14×10^{-05}	16.06	4.41(2.08–9.33)	0.614, 0.265	0.4098
15	7	rs1063243	91,726,927	C	0.411	*AKAP9*	C	6.27×10^{-05}	16.02	4.42(2.08–9.39)	0.659, 0.304	0.5296
16	19	rs57088011	53,454,387	G	0.062	*ZNF816*	C	7.33×10^{-05}	15.72	22.44(2.71–185.8)	0.182, 0.010	0.4609
17	5	rs11556045	73,985,215	G	0.233	*HEXB*	A	7.95×10^{-05}	15.57	0.048(0.00–0.36)	0.977, 0.676	0.282
18	1	rs669408	232,519,150	C	0.35	none	C	8.77×10^{-05}	15.38	4.5(2.06–9.79)	0.600, 0.250	1
19	3	rs2642926	27,615,419	A	0.459	none	T	9.15×10^{-05}	15.3	4.37(2.03–9.38)	0.705, 0.353	0.0125
20	19	rs7248104	7,224,431	A	0.459	*INSR*	A	9.15×10^{-05}	15.3	4.37(2.03–9.38)	0.705, 0.353	0.9049
21	6	rs2073153	29,364,835	C	0.472	*OR12D2*	T	9.17×10^{-05}	15.3	0.21(0.09–0.47)	0.773, 0.418	0.3848
22	3	rs17272796	17,077,268	G	0.336	*PLCL2*	C	9.29×10^{-05}	15.28	4.27(2.01–9.06)	0.568, 0.235	0.2088
23	7	rs10260011	84,709,356	A	0.226	*SEMA3D*	T	9.44×10^{-05}	15.25	4.77(2.10–10.86)	0.432, 0.137	0.5474

S.NO: serial number; CHR: chromosome; SNP ID: single nucleotide polymorphism ID; BP: base pair position at the respective chromosome as per GRCh37.p13; MA: minor allele name; MAF: frequency of minor allele in controls; AA: associated allele; *p*: value of *p*; ChisQ: basic allelic test Chi-square; *p*: value of p; OR: odd ratio; L95: lower bound of 95% confidence interval for odds ratio; U95: upper bound of 95% confidence interval for odds ratio.; CCF: case, control frequencies; HWpval: value of *p* of Hardy–Weinberg equilibrium.

individual samples: GSM6845201-GSM6845273].[1] After filtering 243,345 SNPs according to their *p* values, 280 SNPs with $p<0.0001$ were selected as significant ($p<9.44\times10^{-05}$; Table 2). The most significant SNPs suggesting a correlation with autism were rs2247856 ($p=3.069\times10^{-06}$ at *SPHK1*), rs386789496 ($p=1.036\times10^{-05}$ at *LOC107984893*), rs4602367 ($p=1.783\times10^{-05}$ at *PLCL2*), rs6960867 ($p=2.17\times10^{-05}$ at *AKAP9*) and rs12035482 ($p=2.32\times10^{-05}$) located on chromosome 17, 16, 3, 7 and 1, respectively, (Figure 1). All the significant SNPs of Saudi females autistic patients with $p<0.00018$ are listed in Supplementary Table 1 in which all obey the hardy–Weinberg equilibrium.

All association tests were screened using minor alleles' frequency in controls, value of *p* of Hardy–Weinberg equilibrium and type 1 error rate to achieve the strongest genetic predisposition and imputed for linkage disequilibrium in HapMap SNPs in multiple chromosomes (Figure 2). The haplotype analysis implemented on SNPs with significance of $p<0.0001$ were classified into protective (less probable to cause autism) and risk (more probable to cause autism; Table 3). Risk alleles are listed as the following: in Chromosome 1: *IL24*-rs1150258C; rs1507765C (*value of* $p=2.3773\times10^{-5}$); Chromosome 3: *PLCL2*-rs4602367A; *PLCL2*-rs17272796C (*value of* $p=3.2579\times10^{-5}$); rs9854207; rs2642926 (*value of* $p=9.399\times10^{-6}$); Chromosome 6: *OR5V1*-rs9257819A; *OR5V1*-rs2022077A; *OR12D2*-rs9257834G; *OR12D2*-rs4987411T; *OR12D2*-rs2073154C; *OR12D2*-rs2073153T; *OR12D2*-rs2073151G; *OR5V1*-rs2073149A; *OR5V1*-rs1028411T; *OR5V1*-rs2394607T (*value of* $p=4.5015\times10^{-5}$); Chromosome 7: *AKAP9*-rs6964587T; *AKAP9*-rs6960867G; *AKAP9*-rs1063243C (*value of* $p=2.1723\times10^{-5}$) and Chromosome 19; *ZNF600*-rs7507442G; *ZNF816*-rs57088011C (*value of* $p=7.3276\times10^{-5}$; Table 3; Figure 2). Whereas the alleles of protective haplotypes are: in Chromosome 1: *IL24*-rs1150258T; rs1507765A (*value of* $p=5.9759\times10^{-5}$); Chromosome 3: *PLCL2*-rs4602367G; *PLCL2*-rs272796T (*value of* $p=3.2579\times10^{-5}$); rs9854207A; rs2642926C (*value of* $p=2\times10^{-4}$); Chromosome 6: *OR5V1*-rs9257819C; *OR5V1*-rs2022077T; *OR12D2*-rs9257834T; *OR12D2*-rs4987411C; *OR12D2*-rs2073154G; *OR12D2*-rs2073153G; *OR12D2*-rs2073151A; *OR5V1*-rs2073149T; *OR5V1*-rs1028411G; *OR5V1*-rs2394607C (*value of* $p=1\times10^{-4}$); Chromosome 7: *AKAP9*-rs6964587G; *AKAP9*-rs6960867A; *AKAP9*-rs1063243A (*value of* $p=6.272\times10^{-5}$) and Chromosome 19; *ZNF600*-rs7507442A; *ZNF816*-rs57088011G (*value of* $p=2.8266\times10^{-5}$; Tables 3; Figure 2). Surprisingly, olfactory receptor family 23 subfamily D member 2 (*OR12D2*) and

1 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE221098

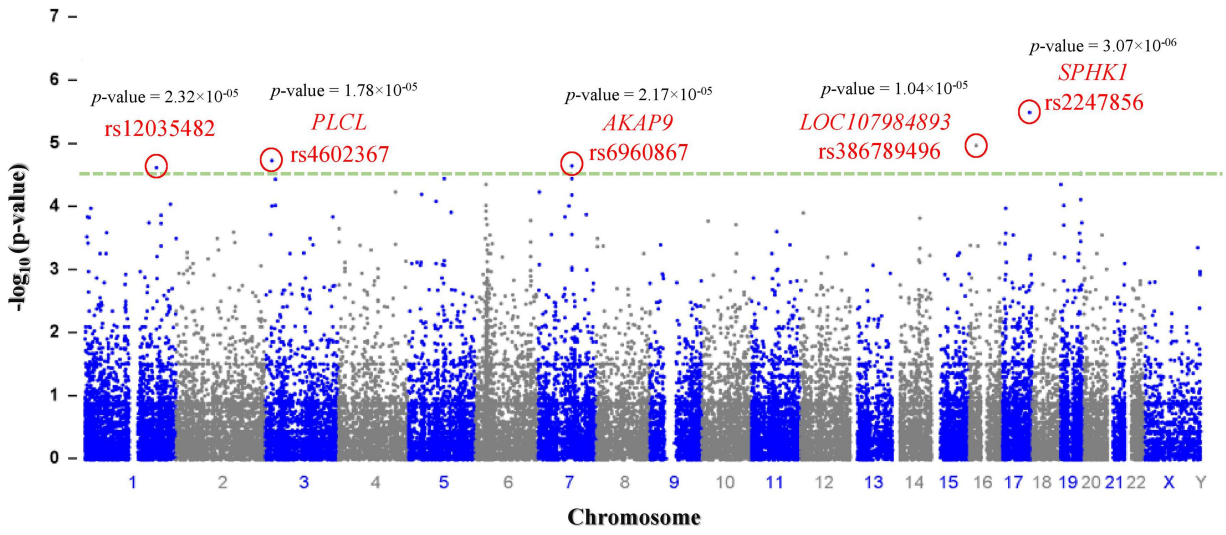


FIGURE 1
Manhattan plot: a total of (n=243,345) SNPs are plotted according to p-values (y-axis) and their position in the genome (x-axis). The most significant candidate nucleotide variants rs2247856 (*SPHK1*), rs386789496 (*LOC107984893*), rs4602367 (*PLCL*), rs6960867 (*AKAP9*) and rs12035482 on chromosome 17, 16, 3, 7and 1 respectively, exceed the significance threshold line (p=1.00×10$^{-4.5}$ - green dash line) indicating a statistically significant correlation with autism.

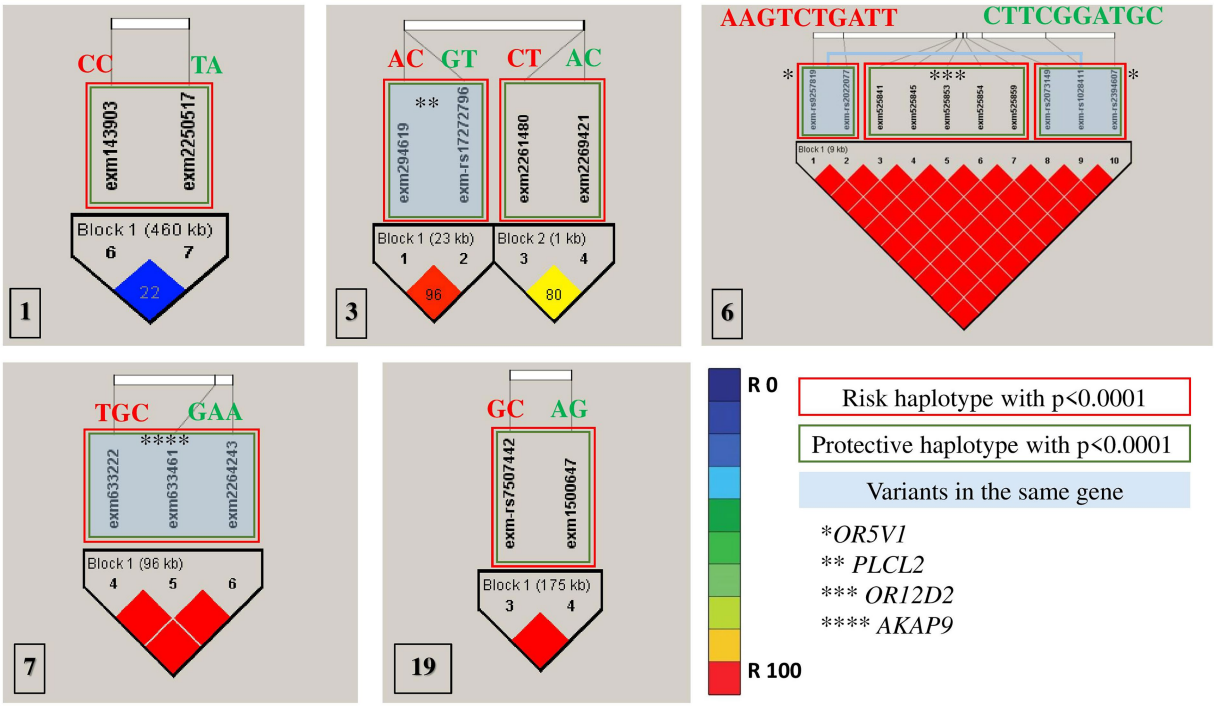


FIGURE 2
Haplotype blocks representing the linkage disequilibrium of chosen SNPs on chromosomes of autistic females in Saudi Arabia. The numbers at the bottom left of each picture correspond to the chromosome number. * refers to the genes' names. Red rectangles are the most risk haplotypes, green rectangles are the most protective haplotypes and light blue rectangles highlights the SNPs which are located in the same gene. Further details are in Table 3.

olfactory receptor family 5 subfamily V member 1 (*OR5V1*) located on chromosome 6 had multiple significant nucleotide variants in Saudi autistic females (Figure 2).

After conducting functional enrichment analysis of females' gene list, genes with SNPs $p<0.00018$ have shown a link to certain diseases including systemic lupus erythematosus disease (SLE; 5 Genes; $p=0.004400769$; *BRD2*, *OR12D2*, *CR2*, *OR5V1*, *HLA-DOA*), Amyotrophic Lateral Sclerosis (3 Genes; $p=0.046868501$; *CUBN*, *SIPA1L2*, *COMMD10*), as well as some pathways like regulation of complement cascade (2 Genes; $p=0.$ 0.0018; *CD55*, *CR2*) vitamin B12

10.3389/fmed.2023.1051039

TABLE 3 Haplotype blocks of SNPs with significant $p<0.0001$ in Saudi autistic females.

Chr	Block	Haplotype	Freq.	Case, control ratio counts	Case, control frequencies	Chi square	Value of *p*	Haplotypes	Risk/ protective
1	Block 1	CC	0.303	24.1: 19.9, 20.2: 81.8	0.548, 0.198	17.86	2.38×10^{-05}	rs1150258C; rs1507765C	Risk
1		TA	0.303	3.1: 40.9, 41.2: 60.8	0.071, 0.404	16.11	5.98×10^{-05}	rs1150258T; rs1507765A	Protective
1		TC	0.21	8.9: 35.1, 21.8: 80.2	0.202, 0.214	0.027	0.8688	rs1150258T; rs1507765C	
1		CA	0.183	7.9: 36.1, 18.8: 83.2	0.179, 0.185	0.006	0.9376	rs1150258C; rs1507765A	
3	Block 1	GT	0.657	18.0: 26.0, 78.0: 24.0	0.409, 0.765	17.261	3.26×10^{-05}	rs4602367G; rs272796T	Protective
3		AC	0.329	25.0: 19.0, 23.0: 79.0	0.568, 0.225	16.361	5.24×10^{-05}	rs4602367A; rs272796C	Risk
3	Block 2	AC	0.503	11.8: 32.2, 61.7: 40.3	0.269, 0.605	13.88	2.00×10^{-04}	rs9854207A; rs2642926C	Protective
3		CT	0.325	25.8: 18.2, 21.7: 80.3	0.587, 0.212	19.63	9.40×10^{-06}	rs9854207C; rs2642926T	Risk
3		AT	0.134	5.2: 38.8, 14.3: 87.7	0.118, 0.140	0.138	0.7106	rs9854207A; rs2642926T	
3		CC	0.038	1.2: 42.8, 4.3: 97.7	0.027, 0.042	0.207	0.6491	rs9854207C; rs2642926C	
6	Block 1	AAGTCTGATT	0.493	33.0: 11.0, 39.0: 63.0	0.750, 0.382	16.647	4.50×10^{-05}	rs9257819A; rs2022077A; rs9257834G; rs4987411T; rs2073154C; rs2073153T; rs2073151G; rs2073149A; rs1028411T; rs2394607T	Risk
6		CTTCGGATGC	0.466	10.0: 34.0, 58.0: 44.0	0.227, 0.569	14.395	1.00×10^{-04}	rs9257819C; rs2022077T; rs9257834T; rs4987411C; rs2073154G; rs2073153G; rs2073151A; rs2073149T; rs1028411G; rs2394607C	Protective
6		AAGTCTGTTC	0.027	1.0: 43.0, 3.0: 99.0	0.023, 0.029	0.052	0.8204	rs9257819A; rs2022077A; rs9257834G; rs4987411T; rs2073154C; rs2073153T; rs2073151G; rs2073149T; rs1028411T; rs2394607C	
6		AAGTCTGTTT	0.014	0.0: 44.0, 2.0: 100.0	0.000, 0.020	0.887	0.3463	rs9257819A; rs2022077A; rs9257834G; rs4987411T; rs2073154C; rs2073153T; rs2073151G; rs2073149T; rs1028411T; rs2394607T	
7	Block 1	GAA	0.589	15.0: 29.0, 71.0: 31.0	0.341, 0.696	16.019	6.27×10^{-05}	rs6964587G; rs6960867A; rs1063243A	Protective
7		TGC	0.397	29.0: 15.0, 29.0: 73.0	0.659, 0.284	18.032	2.17×10^{-05}	rs6964587T; rs6960867G; rs1063243C	Risk
19	Block 1	AG	0.514	11.0: 33.0, 64.0: 38.0	0.250, 0.627	17.531	2.83×10^{-05}	rs7507442A; rs57088011G	Protective
19		GG	0.425	25.0: 19.0, 37.0: 65.0	0.568, 0.363	5.31	0.0212	rs7507442G; rs57088011G	
19		GC	0.062	8.0: 36.0, 1.0: 101.0	0.182, 0.010	15.724	7.33×10^{-05}	rs7507442G; rs57088011C	Risk

Chr: chromosome; Freq.: frequency.

metabolism (2 Genes; $p=0.006683$; *CUBN, INSR*), and female preferences for male odors (2 Genes; $p=0.008301262$; *OR12D2*, *OR5V1*). The most significant SNPs with $p<0.0001$ (Supplementary Table 1) associated with autism in Saudi females were subjected for the functional annotation, there were 27 DAVID IDs. Gene ontology enrichment analysis indicated the significant (p value $=0.0051$; involved 6 genes *MLXIPL*, *ZNF816*, *YEATS2*, *INSR*, *PROX2*, and *ZNF600*) biological process, regulation of DNA-templated transcription (GO:0006355).

4. Discussion

This study evaluates the risk of genetic variation in ASD Saudi female subjects. The most significant *SPHK1* SNP rs2247856 was reported recently as a significantly associated variant with Parkinson's disease in both genders (19). GWAS catalog[2] of rs2247856 reported the observed risk allele (rs2247856-A) of the present study with reticulocyte count, mean corpuscular volume, lymphocyte count, and reticulocyte fraction of red cells. However, no earlier reports on autism. No previous association was reported on rs386789496, rs6960867 and rs12035482. GWAS catalog of *PLCL2* SNP rs4602367-A reported the association with rheumatoid arthritis (20). Even though, *PLCL2* SNP rs4602367-A was not reported on autism, a recent study revealed the association of *PLCL2* SNPs (rs6800583 and rs73139272) with autism (21).

Beginning with significant genes plotted in Manhattan plot, *SPHK1* has the highest *value of p* of 3.069×10^{-6}. *SPHK1* is a key enzyme of sphingolipid metabolism which modulates cellular proliferation and pro-survival function (22). Since *SPHK1* and *SPHK2* phosphorylate sphingosine to sphingosine-1-phosphate (S1P) (23) (Figure 3), the presence of a high concentration of *SPHK1* increases the production of S1P which when elevated can lead to autism according to Wu et al. (24). In addition, based on multiple logistic regression analysis, S1P alterations were considered significant biomarker predictor for autism (23).

2 https://www.ebi.ac.uk/gwas/variants/rs2247856

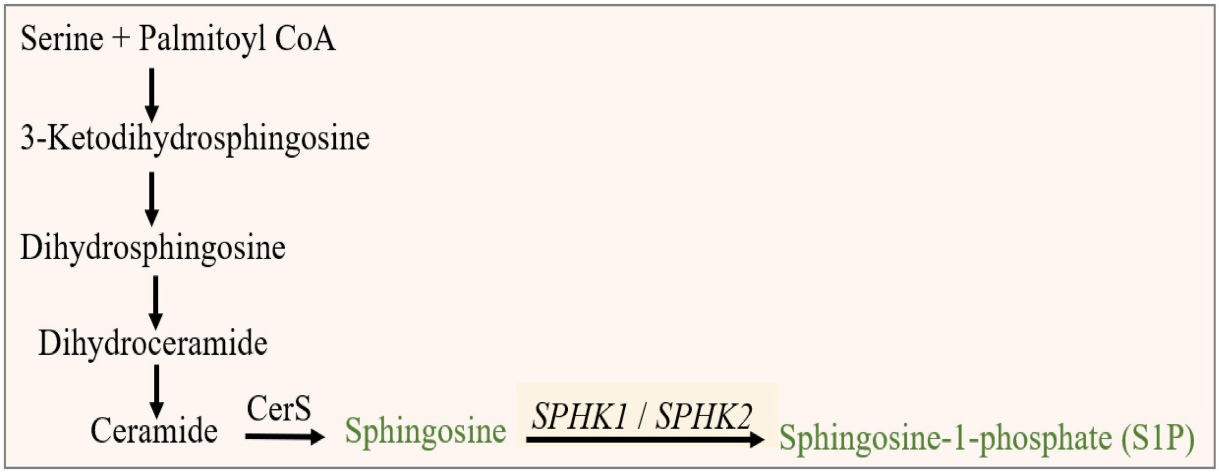


FIGURE 3
Sphingolipid metabolism pathway illustrates the processes of sphingosine-1-phosphate (S1P) production. The last step is catalyzed by *SPHK1*, a significant protein-coding gene associated with autism in Saudi females.

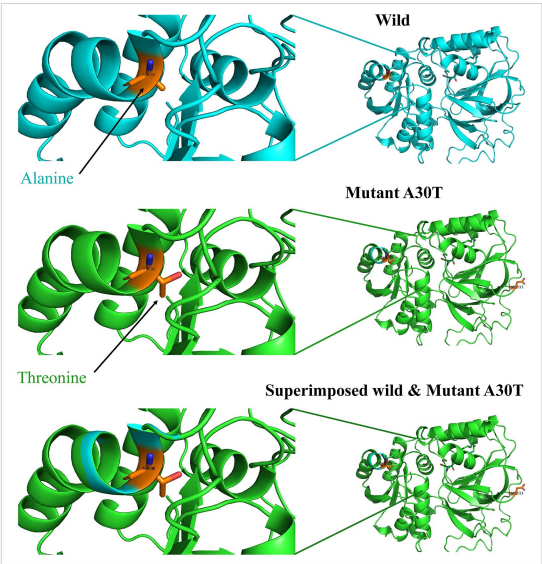


FIGURE 4
Protein models of *SPHK1* wild-type and mutated proteins. Root mean square deviation (RMSD) for the superimposed: 0.001.

Similarly, dysregulation of S1P triggers the manifestation of psychiatric and neurological diseases such as Alzheimer's disease (25), schizophrenia (26) Parkinson's disease (27) and anxiety disorder (28). However, an experiment done on valproic acid rat model found that protein expression of *SPHK1* wasn't significant as it did not reach the significance level (24). Protein modeling of mutated SPHK1 denotes the damaging changes, which indicates the mutated SPHK1 protein can affect the Sphingolipid metabolism pathway (Figure 4). Some findings reported proteins encoded by *AKAP9, another significant gene in the allelic association study*, to be highly expressed in autism subjects (29). *Yet, the mechanism of the association is still unknown.*

Significant SNP candidates for ASD etiology in females were perceived to be located at *OR12D2* and *OR5V1* genes on chromosome 6 having high linkage disequilibrium (Figure 2). Several studies have perceived sensory abnormalities in autism subjects including unusual odor perception (30, 31). Indeed, olfactory genes influence the neurodevelopment status in terms of social, emotional and behavioral functioning (30, 32). A study reported a link between a cluster of SNPs located within the olfactory receptor genes on chromosome 6p22.1 and social defects in ASD ((33); Figure 5). Interestingly, Systemic Lupus Erythematosus (SLE), which is an autoimmune disease closely related to autism, is accompanied by variations in olfactory receptor genes (Figure 5). A research conducted in Egypt revealed that 7 out of 38 autoimmune ASD patients had a family history of SLE (34). The largest cohort study done on 719 SLE offspring reported a strong association between the two disorders (35). Further evidence supporting the relationship between the two disorders is found through the functional enrichment analysis suggesting that *OR12D2* and *OR5V1* are commonly affected genes in both SLE and female ASD patients. Large cluster of olfactory receptor genes on chromosome 6 is located in proximity to class 1 histocompatibility complex genes which mediate immunity (36). Another factor that attributes to the development of ASD in SLE's offspring is the presence of the autoimmune antibodies in patients with SLE which attack the Ro60 protein bound to YRNA (37, 38). All these factors justify the reason behind the doubled risk of ASD in SLE patients' offspring, giving that 21.4–26% of SLE offspring have autism (39). Current studies have emphasized on the potential role for the immune system in ASD, with immune-genetic abnormalities and the inappropriate response of the immune system to environmental challenges. A meta-analysis of 7 observational studies (25,005 ASD cases and 4,543,321 participants) was conducted assessing the relationships between maternal systemic lupus erythematosus (SLE) or rheumatoid arthritis (RA) and risk for ASD in offspring. The results showed that maternal RA was associated with an increased risk for ASDs, whereas maternal SLE was associated with an increased risk for ASD only in western population (40, 41).

Another medical condition that shares common genes with ASD is vitamin B12 (cobalamin) deficiency, which causes many neurological and psychiatric disorders. Cobalamin catalyzes the conversion of

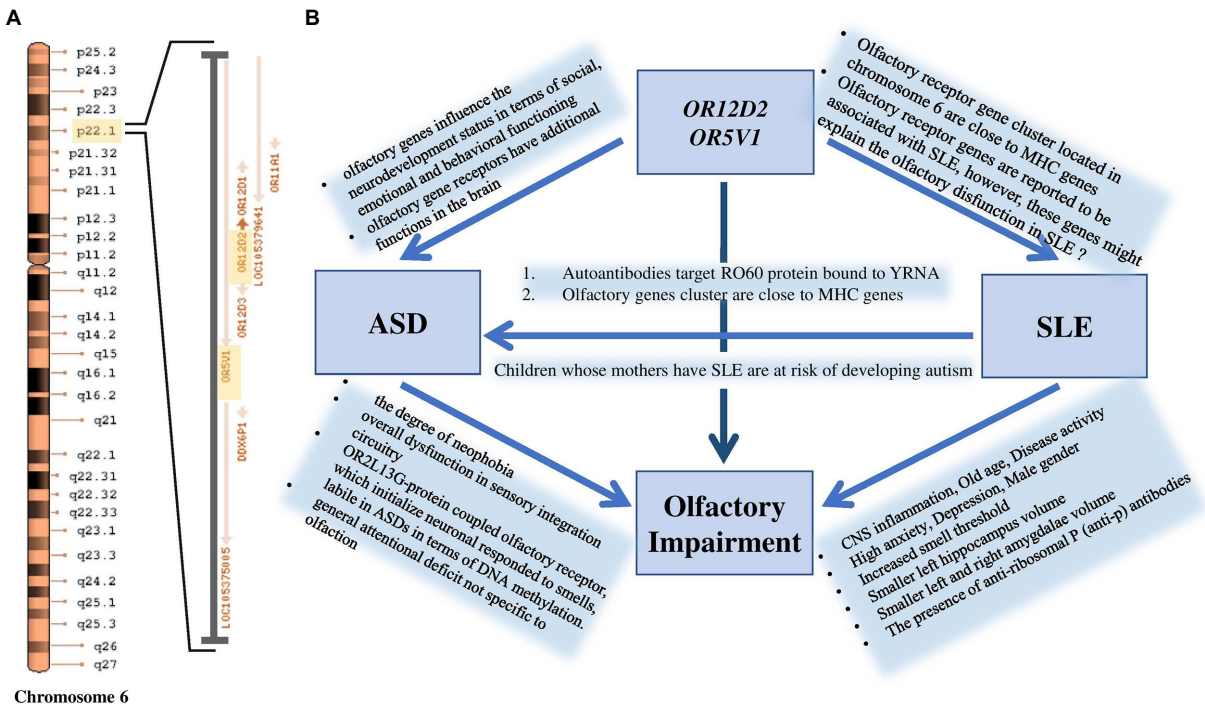


FIGURE 5
(A) Significant ASD variation at *OR12D2* and *OR5V1* genes are located within a cluster of olfactory receptor genes at the chromosomal region 6p22.1.
(B) The figure shows that maternal SLE influences the manifestation of ASD in their offspring. Smell deficits and variations in olfactory receptor genes, particularly in *OR12D2* and *OR5V1*, are common in both ASD and SLE patients.

homocysteine to methionine (1, 42). A study conducted in Oman on 80 participants half of which are cases revealed an accumulation of homocysteine and reduced levels of methionine due to vitamin B12 insufficiency (43). Another study attributed the association between Vitamin B12 and autism to the role of vitamin B12 in the methylation cycle and genetic material biosynthesis (44). Biochemical abnormalities related with ASD consist of impaired methylation and sulphation capacities beside low glutathione (GSH) redox capacity. Possible managements for these abnormalities comprise cobalamin (B12). A systematic review of a total 17 studies was identified studies using vitamin B12 to manage ASD. The study found that generally; vitamin B12 seems to have evidence for efficacy in patients with ASD, especially in individuals who have been identified with unfavorable biochemical profiles. Initial clinical evidence proposes that vitamin B12, mostly subcutaneously injected, improves metabolic abnormalities in ASD alongside with clinical symptoms. Cobalamin is a promising supplement used in the management of ASD (45). The limitations of the current study should be acknowledged. First, the pilot study design nature and its relatively small sample size.

5. Conclusion

In summary, the findings of this study provide the first evidence for female-based genetic analysis in Saudi Arabia and assess the relationship between olfactory receptor genes and ASD. Furthermore, variations on olfactory receptor genes elucidate the impact of SLE in females and the inheritance of ASD. Future investigations with more representative samples that include experiments on rat models are needed to practically prove the association and enhance ASD managing choices.

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 at: GEO database, under accession GSE221098.

Ethics statement

The studies involving human participants were reviewed and approved by Institutional Review Board (IRB) of Imam Abdulrahman Bin Faisal University (IRB-2016-13-152). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

NA, AA, SA, JB: conceptualization, data curation, and investigation. MA, HA, SA, JB, and NA: formal analysis. NA: funding acquisition. NA, SA, and JB: methodology, project administration, resources, software, supervision, validation, and visualization. MA, HA, AA, SA, JB, and NA: writing—original draft. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University (Grant No: 2016-057-IRMC).

Acknowledgments

We express our sincere gratitude to the Dean of the Institute for Research and Medical Consultations (IRMC), Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia for her continuous encouragement and support. We also appreciate the technical assistance from Ranilo M. Tumbaga, Horace T. Pacifico, and Jee E. Aquino. We are grateful for all patients who participated in this study.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1051039/full#supplementary-material

References

1 Almandil, NB, Alkuroud, D, AbdulAzeez, S, AlSulaiman, A, Elaissar, A, and Borgio, JF. Environmental and genetic factors in autism Spectrum disorders: special emphasis on data from Arabian studies. *Int. J. Environ. Res. Public Health*. (2019) 16:658. doi: 10.3390/ijerph16040658

2 Werling, DM. The role of sex-differential biology in risk for autism spectrum disorder. *Biol. Sex Differ.* (2016) 7:58. doi: 10.1186/s13293-016-0112-8

3 El-Fishawy, P. The genetics of autism: key issues, recent findings, and clinical implications. *Psychiatr. Clin.* (2010) 33:83–105. doi: 10.1016/j.psc.2009.12.002

4 Grzadzinski, R, Huerta, M, and Lord, C. DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes. *Mol. Autism.* (2013) 4:12. doi: 10.1186/2040-2392-4-12

5 Who.int. (2018). Autism spectrum disorders. Available at: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders (Accessed July 14, 2019).

6 Alnemary, FM, Aldhalaan, HM, Simon-Cereijido, G, and Alnemary, FM. Services for children with autism in the Kingdom of Saudi Arabia. *Autism*. (2017) 21:592–602. doi: 10.1177/1362361316664868

7 Gottfried, C, Bambini-Junior, V, Francis, F, Riesgo, R, and Savino, W. The impact of neuroimmune alterations in autism spectrum disorder. *Front. Psychol.* (2015) 6:121. doi: 10.3389/fpsyt.2015.00121

8 Lyall, K, Croen, L, Daniels, J, Fallin, MD, Ladd-Acosta, C, Lee, BK, et al. The changing epidemiology of autism spectrum disorders. *Annu. Rev. Public Health*. (2017) 38:81–102. doi: 10.1146/annurev-publhealth-031816-044318

9 Loomes, R, Hull, L, and Mandy, WPL. What is the male-to-female ratio in autism spectrum disorder? A systematic review and meta-analysis. *J. Am. Acad. Child Adolesc. Psychiatry*. (2017) 56:466–74. doi: 10.1016/j.jaac.2017.03.013

10 Carney, RM, Wolpert, CM, Ravan, SA, Shahbazian, M, Ashley-Koch, A, Cuccaro, ML, et al. Identification of MeCP2 mutations in a series of females with autistic disorder. *Pediatr. Neurol.* (2003) 28:205–11. doi: 10.1016/S0887-8994(02)00624-0

11 Hu, VW, Sarachana, T, Sherrard, RM, and Kocher, KM. Investigation of sex differences in the expression of RORA and its transcriptional targets in the brain as a potential contributor to the sex bias in autism. *Mol. Autism.* (2015) 6:7. doi: 10.1186/2040-2392-6-7

12 Woodbury-Smith, M, Deneault, E, Yuen, RK, Walker, S, Zarrei, M, Pellecchia, G, et al. Mutations in RAB39B in individuals with intellectual disability, autism spectrum disorder, and macrocephaly. *Mol. Autism.* (2017) 8:59. doi: 10.1186/s13229-017-0175-3

13 Dayem Ullah, AZ, Lemoine, NR, and Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.* (2012) 40:W65–70. doi: 10.1093/nar/gks364

14 Glusman, G, Caballero, J, Mauldin, DE, Hood, L, and Roach, JC. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics*. (2011) 27:3216–7. doi: 10.1093/bioinformatics/btr540

15 Purcell, S, Neale, B, Todd-Brown, K, Thomas, L, Ferreira, MA, Bender, D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* (2007) 81:559–75. doi: 10.1086/519795

16 Barrett, JC, Fry, B, Maller, JDMJ, and Daly, MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. (2004) 21:263–5. doi: 10.1093/bioinformatics/bth457

17 Huang, DW, Sherman, BT, and Lempicki, RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* (2009) 4:44–57. doi: 10.1038/nprot.2008.211

18 Chen, EY, Tan, CM, Kou, Y, Duan, Q, Wang, Z, Meirelles, GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* (2013) 14:128. doi: 10.1186/1471-2105-14-128

19 Jo, S, Park, KW, Hwang, YS, Lee, SH, Ryu, H-S, and Chung, SJ. Microarray genotyping identifies new loci associated with dementia in Parkinson's disease. *Genes*. (2021) 2021:1975. doi: 10.3390/genes12121975

20 Ha, E, Bae, SC, and Kim, K. Large-scale meta-analysis across east Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. *Ann. Rheum. Dis.* (2021) 80:558–65. doi: 10.1136/annrheumdis-2020-219065

21 Rao, S, Baranova, A, Yao, Y, Wang, J, and Zhang, F. Genetic relationships between attention-deficit/hyperactivity disorder, autism Spectrum disorder, and intelligence. *Neuropsychobiology*. (2022) 81:484–96. doi: 10.1159/000525411

22 Maceyka, M, Sankala, H, Hait, NC, Le Stunff, H, Liu, H, Toman, R, et al. SphK1 and SphK2, sphingosine kinase isoenzymes with opposing functions in sphingolipid metabolism. *J. Biol. Chem.* (2005) 280:37118–29. doi: 10.1074/jbc.M502207200

23 Wang, H, Liang, S, Wang, M, Gao, J, Sun, C, Wang, J, et al. Potential serum biomarkers from a metabolomics study of autism. *J. Psychiatry Neurosci.* (2016) 41:27–37. doi: 10.1503/jpn.140009

24 Wu, H, Zhang, Q, Gao, J, Sun, C, Wang, J, Xia, W, et al. Modulation of sphingosine 1-phosphate (S1P) attenuates spatial learning and memory impairments in the valproic acid rat model of autism. *Psychopharmacology*. (2018) 235:873–86. doi: 10.1007/s00213-017-4805-4

25 Asle-Rousta, M, Kolahdooz, Z, Oryan, S, Ahmadiani, A, and Dargahi, L. FTY720 (Fingolimod) attenuates Beta-amyloid peptide (Aβ 42)-induced impairment of spatial learning and memory in rats. *J. Mol. Neurosci.* (2013) 50:524–32. doi: 10.1007/s12031-013-9979-6

26 Kucharska-Mazur, J, Tarnowski, M, Dołęgowska, B, Budkowska, M, Pędziwiatr, D, Jabłoński, M, et al. Novel evidence for enhanced stem cell trafficking in antipsychotic-naïve subjects during their first psychotic episode. *J. Psychiatr. Res.* (2014) 49:18–24. doi: 10.1016/j.jpsychires.2013.10.016

27 Sivasubramanian, M, Kanagaraj, N, Dheen, ST, and Tay, SSW. Sphingosine kinase 2 and sphingosine-1-phosphate promotes mitochondrial function in dopaminergic neurons of mouse model of Parkinson's disease and in MPP+-treated MN9D cells *in vitro*. *Neuroscience*. (2015) 290:636–48. doi: 10.1016/j.neuroscience.2015.01.032

28 Jang, S, Kim, D, Lee, Y, Moon, S, and Oh, S. Modulation of sphingosine 1-phosphate and tyrosine hydroxylase in the stress-induced anxiety. *Neurochem. Res.* (2011) 36:258–67. doi: 10.1007/s11064-010-0313-1

29 Poelmans, G, Franke, B, Pauls, DL, Glennon, JC, and Buitelaar, JK. AKAPs integrate genetic findings for autism spectrum disorders. *Transl. Psychiatry*. (2013) 3:e270. doi: 10.1038/tp.2013.48

30 Almandil, NB, AlSulaiman, A, Aldakeel, SA, Alkuroud, DN, Aljofi, HE, Alzahrani, S, et al. Integration of Transcriptome and exome genotyping identifies significant variants with autism Spectrum disorder. *Pharmaceuticals*. (2022) 15:158. doi: 10.3390/ph15020158

31 Boudjarane, MA, Grandgeorge, M, Marianowski, R, Misery, L, and Lemonnier, É. Perception of odors and tastes in autism spectrum disorders: a systematic review of assessments. *Autism Res.* (2017) 10:1045–57. doi: 10.1002/aur.1760

32 Tonacci, A, Billeci, L, Tartarisco, G, Ruta, L, Muratori, F, Pioggia, G, et al. Olfaction in autism spectrum disorders: a systematic review. *Child Neuropsychol.* (2017) 23:1–25. doi: 10.1080/09297049.2015.1081678

33 St Pourcain, B, Whitehouse, AO, Ang, WQ, Warrington, NM, Glessner, JT, Wang, K, et al. Common variation contributes to the genetic architecture of social communication traits. *Mol. Autism.* (2013) 4:34. doi: 10.1186/2040-2392-4-34

34 Mostafa, GA, and Kitchener, N. Serum anti-nuclear antibodies as a marker of autoimmunity in Egyptian autistic children. *Pediatr. Neurol.* (2009) 40:107–12. doi: 10.1016/j.pediatrneurol.2008.10.017

35 Vinet, É, Pineau, CA, Clarke, AE, Scott, S, Fombonne, É, Joseph, L, et al. Increased risk of autism spectrum disorders in children born to women with systemic lupus erythematosus: results from a large population-based cohort. *Arthritis Rheum.* (2015) 67:3201–8. doi: 10.1002/art.39320

36 Ortega-Hernandez, OD, Kivity, S, and Shoenfeld, Y. Olfaction, psychiatric disorders and autoimmunity: is there a common genetic association? *Autoimmunity*. (2009) 42:80–8. doi: 10.1080/08916930802366140

37 Kowalski, MP, and Krude, T. Functional roles of non-coding Y RNAs. *Int. J. Biochem. Cell Biol.* (2015) 66:20–9. doi: 10.1016/j.biocel.2015.07.003

38 Wolin, SL, Belair, C, Boccitto, M, Chen, X, Sim, S, Taylor, DW, et al. Non-coding Y RNAs as tethers and gates: insights from bacteria. *RNA Biol.* (2013) 10:1602–8. doi: 10.4161/rna.26166

39 Yengej, FAY, van Royen-Kerkhof, A, Derksen, RH, and Fritsch-Stork, RD. The development of offspring from mothers with systemic lupus erythematosus. A systematic review. *Autoimmun. Rev.* (2017) 16:701–11. doi: 10.1016/j.autrev.2017.05.005

40 Enstrom, AM, Van de Water, JA, and Ashwood, P. Autoimmunity in autism. *Curr. Opin. Investig. Drugs*. (2009) 10:463–73. PMID: 19431079

41 Zhu, Z, Tang, S, Deng, X, and Wang, Y. Maternal systemic lupus Erythematosus, rheumatoid arthritis, and risk for autism Spectrum disorders in offspring: a meta-analysis. *J. Autism Dev. Disord.* (2020) 50:2852–9. doi: 10.1007/s10803-020-04400-y

42 Li, YJ, Ou, JJ, Li, YM, and Xiang, DX. Dietary supplement for core symptoms of autism spectrum disorder: where are we now and where should we go? *Front. Psychol.* (2017) 8:155. doi: 10.3389/fpsyt.2017.00155

43 Al-Farsi, YM, Waly, MI, Deth, RC, Al-Sharbati, MM, Al-Shafaee, M, Al-Farsi, O, et al. Low folate and vitamin B12 nourishment is common in Omani children with newly diagnosed autism. *Nutrition*. (2013) 29:537–41. doi: 10.1016/j.nut.2012.09.014

44 Zhang, Z, Yu, L, Li, S, and Liu, J. Association study of polymorphisms in genes relevant to vitamin B12 and Folate metabolism with childhood autism Spectrum disorder in a Han Chinese population. *Med. Sci. Monit.* (2018) 24:370–6. doi: 10.12659/MSM.905567

45 Rossignol, DA, and Frye, RE. The effectiveness of Cobalamin (B12) treatment for autism Spectrum disorder: a systematic review and meta-analysis. *J. Pers. Med.* (2021) 11:784. doi: 10.3390/jpm11080784

frontiers | Frontiers in Molecular Biosciences

TYPE Original Research
PUBLISHED 23 February 2023
DOI 10.3389/fmolb.2023.1106128

OPEN ACCESS

EDITED BY
Huiyong Sun,
China Pharmaceutical University, China

REVIEWED BY
Soumendranath Bhakat,
University of Pennsylvania, United States
Isman Kurniawan,
Telkom University, Indonesia

*CORRESPONDENCE
T. P. Krishna Murthy,
tpk@live.in
Manikanta Murahari,
manikanta.murahari@gmail.com

SPECIALTY SECTION
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 26 November 2022
ACCEPTED 07 February 2023
PUBLISHED 23 February 2023

CITATION
Poola AA, Prabhu PS, Murthy TPK, Murahari M, Krishna S, Samantaray M and Ramaswamy A (2023), Ligand-based pharmacophore modeling and QSAR approach to identify potential dengue protease inhibitors.
Front. Mol. Biosci. 10:1106128.
doi: 10.3389/fmolb.2023.1106128

Ligand-based pharmacophore modeling and QSAR approach to identify potential dengue protease inhibitors

Anushka A. Poola[1], Prithvi S. Prabhu[1], T. P. Krishna Murthy[1]*, Manikanta Murahari[2]*, Swati Krishna[1], Mahesh Samantaray[3] and Amutha Ramaswamy[3]

[1]Department of Biotechnology, M. S. Ramaiah Institute of Technology, Bengaluru, Karnataka, India, [2]Department of Pharmacy, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India, [3]Department of Bioinformatics, Pondicherry University, Pondicherry, India

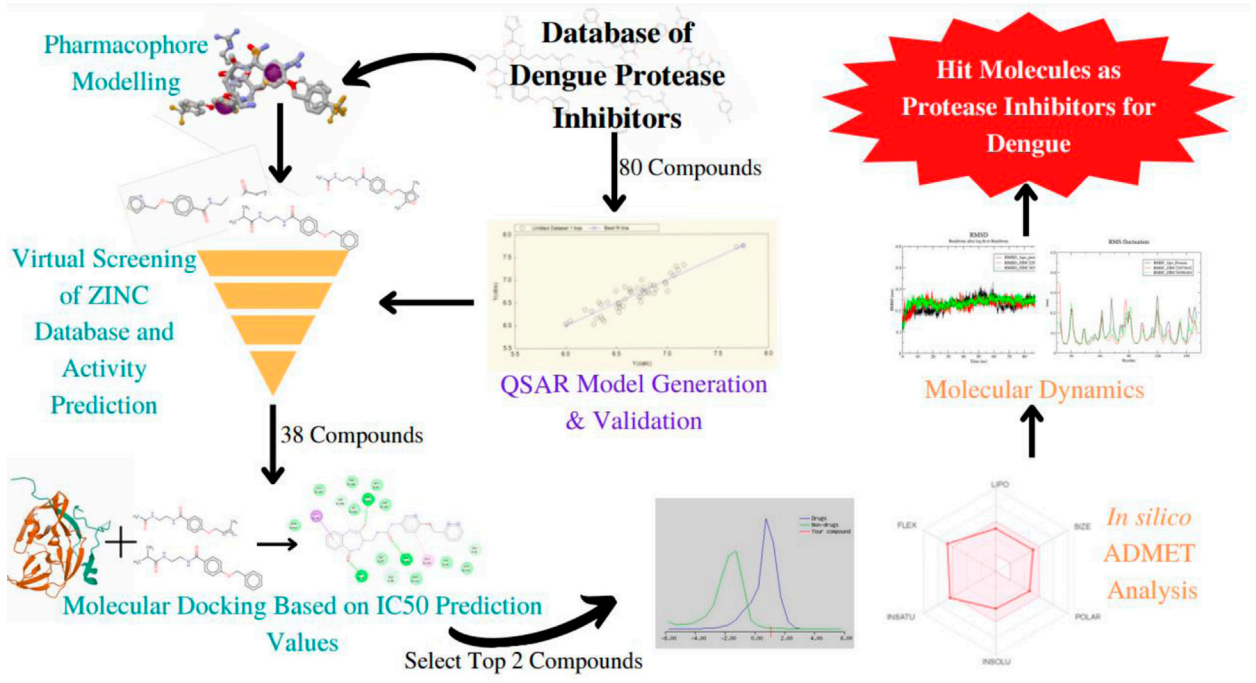
The viral disease dengue is transmitted by the Aedes mosquito and is commonly seen to occur in the tropical and subtropical regions of the world. It is a growing public health concern. To date, other than supportive treatments, there are no specific antiviral treatments to combat the infection. Therefore, finding potential compounds that have antiviral activity against the dengue virus is essential. The NS2B-NS3 dengue protease plays a vital role in the replication and viral assembly. If the functioning of this protease were to be obstructed then viral replication would be halted. As a result, this NS2B-NS3 proves to be a promising target in the process of anti-viral drug design. Through this study, we aim to provide suggestions for compounds that may serve as potent inhibitors of the dengue NS2B-NS3 protein. Here, a ligand-based pharmacophore model was generated and the ZINC database was screened through ZINCPharmer to identify molecules with similar features. 2D QSAR model was developed and validated using reported 4-Benzyloxy Phenyl Glycine derivatives and was utilized to predict the IC50 values of unknown compounds. Further, the study is extended to molecular docking to investigate interactions at the active pocket of the target protein. ZINC36596404 and ZINC22973642 showed a predicted pIC50 of 6.477 and 7.872, respectively. They also showed excellent binding with NS3 protease as is evident from their binding energy of −8.3and −8.1 kcal/mol, respectively. ADMET predictionsofcompounds have shown high drug-likeness. Finally, the molecular dynamic simulations integrated with MM-PBSA binding energy calculations confirmedboth identified ZINC compounds as potential hit moleculeswith good stability.

KEYWORDS

Dengue, QSAR, pharmacophore modeling, docking, molecular dynamics

Introduction

Dengue, a viral disease caused by members of the Flaviviridae family, is a leading public health concern, affecting most Asian and Latin American countries, and becoming a major cause of hospitalization and death in these regions (WHO, 2022). The disease spreads among humans through infected female *Aedes aegypti* or *Aedes albopictus* (Adawara et al., 2020). There are four serotypes of Dengue virus (DENV), namely, DEN-1, DEN-2, DEN-3, and DEN-4, of which DEN-2 is considered



GRAPHICAL ABSTRACT

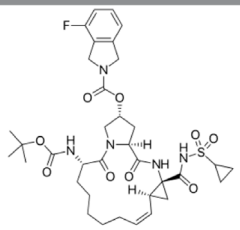
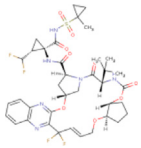
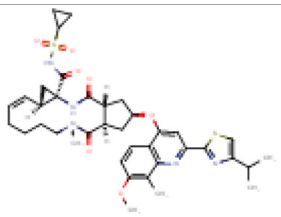
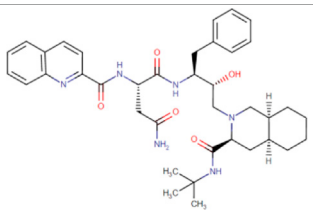
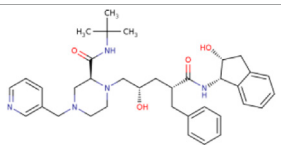
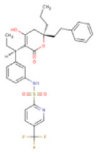
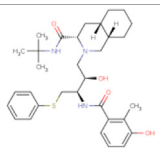
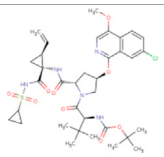
the most virulent strain (Adawara et al., 2020; Dwivedi et al., 2021). Up to date, other than supportive, no specific antiviral treatment exists to treat the illness, thus finding potential compounds that have an anti-dengue activity that can be developed into efficient drugs with the least toxic effects on human beings is the need of the hour (Wellekens et al., 2022). *In vitro* testing of inhibitory activities of various compounds is a time-consuming procedure and is also expensive, pointing toward the usage of quantitative structure-activity relationship (QSAR) models which is a promising way to predict the biological activity of new compounds (Kurniawan et al., 2020).

The viral genome encodes for three structural proteins and seven non-structural proteins, of which NS3 is a non-structural protein that is essential for RNA replication and viral assembly (Dwivedi et al., 2021). This protein contains a serine protease domain, whose activity depends on the formation of a non-covalent complex with the NS2B protein as a cofactor, thus making the NS3 protein an attractive target that can be used to develop dual-acting drugs that are effective against DENV (Behnam et al., 2015). It has been reported that structure-based drug design may not be suitable for developing NS3–NS2B inhibitors due to the specific structure of the protease which is slightly smooth in 3D space, and to date, ligand interaction mechanism and QSAR information are very limited (Luo et al., 2017).

Various *in silico* studies aiming to identify NS2B/NS3 inhibitors have been performed, for example, a study by Qamar et al., in 2017 pointed out that plant flavonoids have the potential to inhibit the dengue protease enzyme and could stop replication of DENV(Qamar et al., 2017). Other studies focusing on phytocompounds as novel dengue protease inhibitors have also been reported isolated phytochemicals belonging to different groups including fatty acids, glucosides, terpenes and terpenoids, flavonoids, phenolics, chalcones, acetamides, and peptides. Curcumin, quercetin, and myricetin were found to act as non-competitive inhibitors for the NS2b/NS3 protease enzyme (Saqallah et al., 2022). Though various *in silico* experiments have been performed to identify NS2b/NS3 inhibitors, most of these studies are molecular docking based, and studies based on QSAR are few.

In 2015, Behnam et al. performed a study that presents an extensive biological evaluation of NS3 inhibitors containing benzyl ethers of 4-hydroxyphenylglycine that function as non-natural peptide building blocks synthesized *via* a copper-complex intermediate. In this study, we make use of these inhibitors to develop a ligand-based pharmacophore model as well as a QSAR model, in order to identify lead compounds having anti-dengue activity. This study also elaborates on the ligand interactions and toxicity analysis of the inhibitors based on *in silico* predictions. These findings can then be utilized and integrated into *in vitro* studies in order to

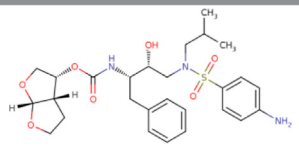
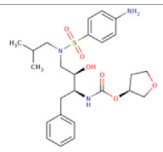
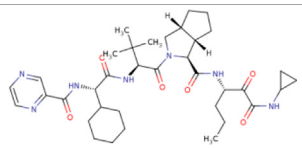
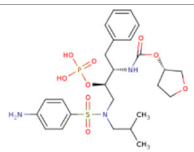
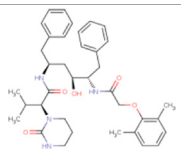
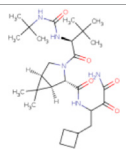
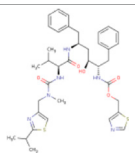
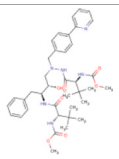
TABLE 1 Structures of the selected FDA approved drugs and their docking scores.

S. No.	Standard drug	Structure	Binding energy (kcal/mol)
1	Danoprevir		−13.5
2	Glecaprevir		−13
3	Simeprevir		−12.1
4	Saquinavir		−10.5
5	Indinavir		−10.5
6	Tipranavir		−10.3
7	Nelfinavir		−10.2
8	Asunaprevir		−9.9

(Continued on following page)

10.3389/fmolb.2023.1106128

TABLE 1 (*Continued*) Structures of the selected FDA approved drugs and their docking scores.

S. No.	Standard drug	Structure	Binding energy (kcal/mol)
9	Darunavir		−9.4
10	Amprenavir		−9.3
11	Telaprevir		−9.2
12	Fosamprenavir		−9.2
13	Lopinavir		−9.1
14	Boceprevir		−8.8
15	Ritonavir		−8.6
16	Atazanavir		−8

further confirm the possibility of developing these inhibitors into effective drugs.

Methodology

Identification of inhibitor compounds

An extensive survey of literature revealed the DenvInD-Database of inhibitors of Dengue virus (https://webs.iiitd.edu.in/raghava/denvind/), a curated database of Dengue virus inhibitors for clinical and molecular research (Dwivedi et al., 2021). This database contains detailed information about the SMILES, PubChem IDs, EC_{50}, CC_{50}, IC_{50}, and K_i values of 484 compounds which have been validated as inhibitors against various drug targets of dengue virus using *in vitro* studies. From this database, the specific set of inhibitors against NS3 protease was selected for further studies. Out of the 365 NS3 protease inhibitors reported in the database, 104 compounds containing 4-Benzyloxy Phenyl Glycine residues were selected, whose biological assays were performed using fluorometric assay HPLC-based DENV-protease assay in order to eliminate false positives (Behnam et al., 2015). The IC_{50} value is a measure of the effectiveness of a drug in bringing about the inhibition of its respective target. Therefore, based on the availability of IC_{50} values, 80 compounds were further selected for the pharmacophore modeling and QSAR study as is presented in the supplementary information. The IC_{50} values were converted to pIC_{50} values in order to normalize the variation in concentration units. The structures of these 80 compounds were drawn using ChemSketch, a software developed by Advanced Chemistry Development, Inc. (Li et al., 2004).

Identification of standard drugs

There is presently no standard treatment for dengue infection and therefore there is a need to explore all avenues that will lead us to potential drugs. In order to carry out a comparative analysis between the compounds obtained from DenvInD and standard drugs used to treat other similar viruses, as well as to check the possibility of drug repurposing, a set of 15FDA-approved standard antiviral drugs have been reported to inhibit protease in Hepatitis C Virus (HCV) and Human Immunodeficiency Virus (HIV) was identified, as shown in Table 1. The SDF files of these compounds were downloaded from DrugBank for further analysis (Wishart et al., 2018).

Pharmacophore-based screening of ZINC database

The top 3 compounds with the highest pIC50 values were selected and their energies were minimized using Avogadro, using the steepest descent algorithm and MMFF94 force field (Hanwell et al., 2012). These molecules were converted to mol2 format and were provided as input to PharmaGist with the maximum number of output pharmacophores as 5, in order to develop the pharmacophore model. The pharmacophore feature output file was then used as input to ZINCPharmer, an open web server used to screen the ZINC database to identify compounds with similar pharmacophore features (Koes and Camacho, 2012). The resultant compound hits were then downloaded as SDF files for molecular docking analysis.

Quantitative structure-activity studies (QSAR) studies

Creating training and test set

The 80 final compounds chosen from DenvInD were split into training set and test set. The range of pIC50 values for the training set and test set was 5.42–7.74 and 5.01–7.55, respectively. Based on a randomized process, 64 compounds were considered in the training set, and the remaining 16 compounds were considered in the test set. The training set was used to build the QSAR model.

Generation of descriptor

Molecular descriptors refer to structural and physicochemical properties that define a molecule and usually include properties like steric parameters, hydrophobic properties, electrostatic properties, *etc.*, as well as constitutional properties of the molecule. The descriptors for the 64 compounds in the training set were calculated using PaDEL software (Yap, 2011). Significant descriptors were selected for further analysis based on their correlation with the pIC_{50} values of the training compounds.

Building QSAR model-generation and validation

The BuildQSAR tool was used to build the QSAR model using the 64 training compounds (Singh et al., 2022). A QSAR study performed First, a systematic search was performed to select a set of descriptors (maximum 3) on the basis of user-given correlation criteria with respect to activity (pIC_{50}). Further, the Multiple Linear Regression (MLR) method was used to build the QSAR model using multiple combinations of the selected descriptors (Murahari et al., 2017). The descriptors were selected based on various statistical parameters like high correlation coefficient (R), high Fischer's value (F-Test), low Standard error of estimate (s), statistical significance (p), high cross-validated square of correlation coefficient (Q^2), low sum of squared error of prediction (SPRESS) and low standard deviation of error of prediction (SDEP). The models that showed significant statistical parameters were tested using the 16 compounds in the test set, to check the fitness of the QSAR model.

Activity prediction of screened ZINC compounds

The pIC_{50} values of ZINC database compounds obtained as a result of ZINCPharmerscreening were predicted using the validated QSAR model that showed highly significant statistical parameters. The compounds with good pIC_{50} values in comparison with

TABLE 2 PharmaGist results.

S. No.	Score	Spatial features	Aromatic	Hydrophobic	Donor	Acceptor	Molecules
1	29.394	6	2	0	3	1	DenvInD_285, DenvInD_266, DenvInD_265
2	22.780	6	1	1	3	1	DenvInD_285, DenvInD_266, DenvInD_265
3	22.045	4	2	0	1	1	DenvInD_285, DenvInD_266, DenvInD_265

compounds obtained from DenvInD were used for further computational studies.

Molecular docking studies

Preparation of protein

The structure of Dengue Virus NS2B/NS3 Protease was obtained from RCSB PDB (PDB ID: 2FOM) (Sarwar et al., 2018). SWISS-MODEL was used to repair the missing atoms (Waterhouse et al., 2018). Further, the ligands from the protein structure were removed using BIOVIA Discovery Studio and the protein was prepared for docking in AutoDock Vina, a part of MGL tools 1.5.7 (Seeliger and De Groot, 2010; Pawar and Rohane, 2021). Water molecules were deleted, polar hydrogen atoms and Kollman charges were added. The prepared protein was saved as a pdbqt file and further used for docking analysis. The binding site coordinates were obtained as x = −3.243 y= −9.193 and z = 16.143 based on key amino acid residues (His 51, Asp 75, and Ser 135) using PyMol version 4.4, a molecular visualization software (Yuan et al., 2017). The grid box size of 40 A^0 was used for docking.

Docking with ZINC database compounds and standard drugs

The compounds obtained from the ZINC database after the pharmacophore-based screening, as well as the 15FDA-approved antiviral protease inhibitors were converted to pdbqt format and their energy was minimized using the MMFF94 force field. AutoDock Vina was used for docking. Docking was performed using exhaustiveness parameter as 10. Docking scores and binding interactions at the active pocket of target protein for respective ligands were inspected and recorded carefully. The output complexes with high binding affinity and pIC_{50} were further used to perform molecular dynamics simulation studies.

Molecular dynamic simulations

The top 2 compounds obtained after docking and QSAR activity predictions of the selected ZINC database compounds were further subjected to molecular dynamic simulations using GROMACS version 2018.1 (Van Der Spoel et al., 2005). The receptor topology was obtained by the "pdb2gmx" script, while the ligand topologies were obtained by the PRODRG server (Schüttelkopf and Van Aalten, 2004). Each of the generated ligand topologies was rejoined to the processed receptor structure to construct the ligand-protein complex. GROMOS96 54a7 force field was used to obtain the energy minimized conformations of all the processed complexes (Schmid et al., 2011). Next, a solvation step was performed wherein the structures were solvated in a cubic periodic box (90 Å, 90 Å, 90 Å) with water extended simple point charge (SPC) model. In order to neutralise the system, 4 Na ions added. Subsequently, energy minimization of the system was carried out for 50,000 steps using the steepest descent algorithm with <10.0 kJ/mol force. Upon energy minimization, equilibration of the system was performed with two consecutive steps. The NVT ensemble followed by NPT ensemble was done for 50,000 steps each. A constant temperature of 300 K and constant pressure of 1 atm were maintained through the entire MD simulation. The long-range electrostatic interactions were obtained by the particle mesh Eshwald method with a 12 Å cut-off and 12 Å Fourier spacing. Finally, the three well-equilibrated systems (one apo protein and two protein-ligand complexes) was subjected to a final 100 ns simulation. Root mean square deviation (RMSD), Root Mean Square Fluctuation (RMSF), Radius of Gyration (R g), Solvent Accessible Surface Area (SASA) and Number of Hydrogen bonds of the protein and complxes were calculated using gmx_rms, gmx_rmsf, gmx_gyrate, gmx_sasa and gmx_hbond tools, respectively. The MM/PBSA study using g_mmpbsa version 5.1.2 utility was used to analyze the binding free energy (ΔG binding) of the ligands with protein over the whole 100 ns simulation time.

Prediction of drug-likeness and ADMET properties of ZINC compounds

The hit molecules were then studied further investigated for drug-likeness, toxicity, and ADME properties. Molsoft Drug-Likeness and molecular property prediction tool were used to predict drug-likeness (Elsherif et al., 2020) Other chemical properties like the number of hydrogen bond donors, hydrogen bond acceptors, BBB score, $pK_{a,}$ etc., were also analyzed during this step. It is extremely important to understand the toxicity levels of compounds before considering it further as a potential drug lead. Hence to predict the toxicity class of compounds, ProTox-II was used (Drwal et al., 2014). Further, to elucidate the physicochemical descriptors, pharmacokinetic properties, ADME parameters, and drug-like nature, SwissADME tool was used (Daina et al., 2017).

Results and discussion

Ligand-based pharmacophore modeling

Top 3 compounds with highest pIC50, i.e., DenvInD_285, DenvInD_265 and DenvInD_266, were submitted to PharmaGistwebserver to generate the pharmacophore model. This web server predicts a ligand-based pharmacophore model

TABLE 3 Details of the descriptors chosen to build the QSAR model (Karthikeyan et al., 2021).

S. No.	Descriptor	Description	Descriptor class
1	GATS6e	Geary autocorrelation–lag 6/weighted by Sanderson electronegativities	Autocorrelation descriptor
2	GATS5i	Geary autocorrelation–lag 5/weighted by first ionization potential	
3	VE1_DzZ	Coefficient sum of the last eigenvector from Barysz matrix/weighted by atomic number	Barysz Matrix descriptor
4	VE2_DzZ	Average coefficient sum of the last eigenvector from Barysz matrix/weighted by atomic number	
5	VE3_DzZ	Logarithmic coefficient sum of the last eigenvector from Barysz matrix/weighted by atomic number	
6	SpMAD_Dzp	Spectral mean absolute deviation from Barysz matrix/weighted by polarizabilities	
7	SpMax3_Bhp	Largest absolute eigenvalue of Burden modified matrix–n 3/weighted by relative polarizabilities	Burden Modified Eigen values descriptor
8	ETA_Epsilon_5	A measure of electronegative atom count	Extended Topochemical Atom descriptor
9	IC1	Information content index (neighborhood symmetry of 1-order)	Information Content descriptor
10	IC2	Information content index (neighborhood symmetry of 2-order)	
11	TIC0	Total information content index (neighborhood symmetry of 0-order)	
12	MIC1	Modified information content index (neighborhood symmetry of 1-order)	
13	WTPT-3	Sum of path lengths starting from heteroatoms	PaDEL Weighted Path descriptor

based on the best alignment of maximum features between the submitted molecules. Considering a perfect alignment of all the 3 molecules submitted, a pharmacophore model was obtained with a PharmaGist score of 29.394 having six spatial features. The pharmacophore model generated includes a total of 6 features-spatial features, aromatic 2), donors 3), acceptor 1), and the results of other pharmacophores identified were presented in Table 2.

Pharmacophore-based screening of ZINC database

The pharmacophore features obtained from PharmaGist were downloaded and used to screen the ZINC database through ZINCPharmer webserver in order to find ligands with similar pharmacophore features with an assumption of having similarity in pharmacological properties. The query led to 38 hits from the ZINC database with optimization of low RMSD and molecular weight. The structures of these compounds were presented in the Supplementary Material.

Building QSAR model and activity prediction of ZINC database compounds

Using PaDEL software 1,444 descriptors were generated for the training set of 64 compounds. Based on the correlation coefficient calculated with respect to pIC_{50} values of the respective compounds, 13 descriptors were identified for further analysis. The training set of 64 compounds was given as input to the BuildQSAR tool to generate the QSAR models. A variable selection search was performed using "systematic search" mode using correlation criteria limits of 0.6–0.78 and the variable limit of 3. The influencing parameters were found to be GATS6e (X1), GATS5i(X2), VE1_DzZ (X3), VE2_DzZ (X4), VE3_DzZ (X5), SpMAD_Dzp (X6), SpMax3_Bhp(X7), ETA_Epsilon_5 (X8), IC1(X9), IC2(X10), TIC0(X11), MIC1(X12), WTPT-3 (X13) and they are further described in Table 3. GATS6e and GATS5i are autocorrelation descriptors which are essentially molecular descriptors that encode molecular structure as well as the physicochemical properties attributed to the atoms in the form of vectors (Hollas, 2003). VE1_DzZ, VE2_DzZ, VE3_DzZ and SpMAD_Dzp are Barysz Matrix descriptors. Barysz matrix is a weighted distance matrix that accounts for the presence of multiple bonds and heteroatoms in the molecule under consideration. SpMax3_Bhp is a Burden Modified Eigenvalues descriptor that reflects the topology of the molecule. ETA_Epsilon_5 is an Extended Topochemical Atom descriptor that determines the contributions of specific positions within common substructures of molecular graphs towards total functionality (Roy and Ghosh, 2003). IC1, IC2, TIC0, and MIC1 are Information Content descriptors, and WTPT-3 is a PaDEL Weighted Path descriptor. The QSAR model was generated using a trial-and-error method to find the best fitting model that has a high R, R^2, F-test, and Q^2 and low s values, SPRESS, and SDEP statistical values. The top six models were shown in Table 4. These models were further tested using the test set to verify whether the pIC_{50} value predicted by these models was comparable to experimental values. Upon graphical analysis, it was seen that model 1 exhibited the highest R^2 value of 0.703 between observed and predicted pIC_{50} values. Hence model 1 was chosen for further studies. The pIC_{50} predicted using Model 1 ranged from 4.507 to 8.164. Further information about the model is given in the supplementary file. The pIC_{50} of the library compounds ranged from 5.013 to 7.744. This shows that the validated QSAR model could identify compounds with better predicted pIC_{50} values, for which the objective was partially fulfilled. As the compounds need to be tested experimentally. The predicted activity for the ZINC

TABLE 4 QSAR models and their statistical parameters.

Model no.	Descriptor 1	Descriptor 2	Descriptor 3	R Value	R^2 values	S	Q2	F	*p*	SPRESS	SDEP	*n*	QSAR equation
1	X1	X7	X10	0.92	0.85	0.1537	0.8197	89.5853	0	0.1663	0.1614	53	Y1 = - 2.5236 (±0.4389) X1 - 0.0599 (±0.0592) X7 + 1.0876 (±0.3812) X10 + 4.5352 (±2.1729)
2	X1	X8	X10	0.92	0.85	0.1530	0.8215	90.5701	0	0.1654	0.1606	53	Y1 = - 2.5530 (±0.4378) X1 - 0.3144 (±0.2939) X8 + 1.0895 (±0.3795) X10 + 4.5728 (±2.1648)
3	X1	X5	X9	0.92	0.84	0.1521	0.8129	85.77733	0	0.1659	0.1610	52	Y1 = - 2.3301 (±0.4819) X1 + 0.0140 (±0.0127) X5 + 0.8225 (±0.4152) X9 + 6.5410 (±1.9935)
4	X1	X6	X10	0.91	0.83	0.1596	0.8053	82.5516	0	0.1718	0.1669	54	Y1 = - 2.5308 (±0.4532) X1 - 0.0097 (±0.0106) X6 + 1.0874 (±0.3936) X10 + 4.5341 (±2.2422)
5	X1	X5	X10	0.91	0.83	0.1607	0.8016	81.2388	0	0.1735	0.1685	54	Y1 = - 2.5155 (±0.4559) X1 + 0.0110 (±0.0136) X5 + 0.9397 (±0.4266) X10 + 5.1655 (±2.4390)

database compounds were presented in Table 5. These compounds were then analyzed using docking studies to identify the binding patterns and interactions at the active pocket of the target protein.

Molecular docking studies

Docking of ZINC database compounds

The selected set of ZINC database compounds was subjected to docking against dengue protease as stated in the protocol. The binding energies ranged from −9 kcal/mol to −7.3 kcal/mol as shown in Table 5. The top 2 compounds identified were ZINC36596404 and ZINC22973642 with binding energies −9 kcal/mol and −8.9 kcal/mol, respectively. The interactions between the protein and the ligand were summarized in Table 6. Upon observing the interaction between dengue protease and ZINC36596404, conventional hydrogen bond, carbon-hydrogen bond, Pi-donor hydrogen bond, pi-sigma, and pi-alkyl were found to be significant. Lys74, Trp83 and Trp89 were involved in a conventional hydrogen bond, Gly148, Glu88 and Glu91 were involved in carbon-hydrogen bond and pi-donor hydrogen bond, Leu76 was involved in pi-sigma bond and Ala166 in pi-alkyl bond. Next, the interaction between dengue protease and ZINC22973642 was analyzed, revealing that van der Waals, conventional hydrogen bond, carbon hydrogen bond, alkyl, and pi-alkyl were noteworthy. The amino acid interactions for these bonds were seen to involve Thr118. Thr120, Trp89, Glu88, Asn152, Lys73, Ile165 for van der Waals bonds; Asn167, Leu149, Val47 contributed to conventional hydrogen bonding; Gly148, Leu76, Trp83, Gly87, Leu85 for hydrogen bonds; Val154, Ile123, Ala166, Ala164, Lys74 for alkyl and pi-alkyl. The interactions are represented in Figure 1.

Docking of standard drugs

The results obtained when the 15 chosen standard drugs were docked against the Dengue protease were presented in Table 1. The binding energies fall in the range of −13.5 kcal/mol to −8 kcal/mol. From this, we can observe that Danoprevir, Glecaprevir, Simeprevir, Indinavir, Tipranavir, Nelfinavir, Asunaprevir, Darunavir, and Amprenavir have a better binding affinity with the Dengue protease compared to the ZINC database compounds screened in this study. This directs us to conduct an experimental study in order to formulate a drug that works against dengue protease. Danoprevir interacts with the receptor using van derWaals forces contributed by Asn167, Ala166, Ala164, Ile165, Asn152, Leu76, Met49, Leu149, Gly148 and Val147. Conventional hydrogen bonds made by Lys74 and carbon hydrogen bonds made by Leu85, Val146 and Gly87 also take part in the interactions. Glecaprevir interacted with the receptor through attractive charges of Glu88, conventional hydrogen bond of Trp83, carbon hydrogen bond of Gly148, halogen bond by Val147 and pi-cation bond by Glu88. Amino acids in Simeprevir that interacted with the receptor include Lys74, Asn167, Lys73, Ala164, Asn152, Ile123, Gly153, Val154, Thr120, Thr118, Asn119 and Val155 that contribute to van der waals forces, and Asp71 that is involved in attractive charges. Indinavir was seen to interact with the receptor through mainly alkyl and pi-alkyl bonds formed by Trp83, Leu149, Leu76 and Leu85, attractive charges of

10.3389/fmolb.2023.1106128

TABLE 5 Results of docking ZINC database compounds against NS3 protease.

S. No.	ZINC compound	Binding energy (kcal/mol)	Predicted pIC50
1	ZINC36596404	−9	6.477
2	ZINC22973642	−8.9	7.872
3	ZINC09789323	−8.7	6.399
4	ZINC16699623	−8.7	4.507
5	ZINC19143967	−8.5	7.047
6	ZINC09833225	−8.3	6.907
7	ZINC02458390	−8.2	6.189
8	ZINC06148003	−8.2	6.869
9	ZINC27672080	−8.2	7.086
10	ZINC14028064	−8.1	6.700
11	ZINC14037170	−8.1	7.188
12	ZINC35025967	−8	6.584
13	ZINC14036276	−8	6.860
14	ZINC67678868	−7.9	6.473
15	ZINC36656172	−7.9	7.474
16	ZINC02563681	−7.8	6.434
17	ZINC01155209	−7.8	6.743
18	ZINC15634648	−7.7	6.628
19	ZINC17795,206	−7.7	7.198
20	ZINC23080510	−7.7	8.050
21	ZINC32477936	−7.6	7.121
22	ZINC23327308	−7.6	7.414
23	ZINC32042479	−7.6	7.702
24	ZINC32908224	−7.6	7.391
25	ZINC14664807	−7.5	7.170
26	ZINC33242299	−7.5	7.713
27	ZINC69504947	−7.5	6.964
28	ZINC09826328	−7.5	6.728
29	ZINC23114768	−7.5	7.242
30	ZINC06445998	−7.4	5.955
31	ZINC37514943	−7.4	6.332
32	ZINC22755327	−7.4	7.666
33	ZINC32485749	−7.4	7.466
34	ZINC78464608	−7.4	8.164
35	ZINC32908634	−7.3	7.931
36	ZINC64718088	−7.3	6.723
37	ZINC23114770	−7.3	7.242
38	ZINC93765844	−7.3	6.838

TABLE 6 Summary of protein-ligand interactions.

S. No.	Compound	Residues involved in protein-ligand interactions						
		Conventional hydrogen bond	Carbon hydrogen bond	Pi-donor hydrogen bond	Pi-sigma	Pi-alkyl	Alkyl bonds	Van der waals	Hydrogen bond
1	ZINC36596404	Lys74, Trp83 and Trp 89	Gly148 and Glu88 and Glu91		Leu76	Ala166	-	-	-
2	ZINC22973642	Asn167,Leu149,Val47	-	-	-	Val154, Ile123, Ala166, Ala164, Lys74		Thr118, Thr120, Trp89, Glu88, Asn152, Lys73, Ile165	Gly148, Leu76, Trp 83, Gly87, Leu85
3	ZINC09789323	ASN152	-	-	-	Lys90, Ala166, Leu76		Leu115, Glu91, Trp89, Lys74, Ala164, Leu149, gly148, Met49	-

Glu88, and carbon hydrogen bond formed by Gly148 and Ala164. The amino acid interactions seen among other standard drugs studies are elaborated in the Supplementary Information. The binding interactions of Danoprevir and Glecaprevir, the top 2 compounds were further examined and compared with the binding interactions of the top hit ZINC compounds ZINC36596404 and ZINC22973642. Comparing the amino acid interaction of ZINC compounds and standard drugs with the receptor, we get interesting inferences. The results show that Ala166, Leu76, and Gly148 seem to play an important role in interaction with the receptor as they are involved in interactions with the receptor in Danoprevir, ZINC36596404, and ZINC22973642. While Ala166 is involved in van der Waals forces in Danoprevir interaction, it is involved in pi-alkyl and alkyl bonding in ZINC36596404 and ZINC22973642 interactions, but we can conclude that they are important residues in hydrophobic interactions. Leu76 and Gly148 seem to be contributing significantly to different types of hydrogen bonding. Glu88 and Trp83 were identified as another set of important amino acid residues interacting with the receptor in Glecaprevir, ZINC36596404, and ZINC22973642. Glu88 can be said to be necessary for hydrophobic interactions like pi-cation interaction and van der Waals interactions as well as hydrogen bonding. Trp83 has shown to be contributing to various hydrogen bonds in Glecaprevir, ZINC36596404, and ZINC22973642. Gly148 can be pointed out as a major key residue as it is involved in hydrogen bonding in all the compounds discussed above. From this, we can understand that by preserving these key interactions in the ZINC compounds and modifying other groups, we can develop the identified ZINC compounds into effective inhibitors of Dengue Protease.

Molecular dynamic simulation

Root mean square deviation analysis

ZINC36596404 and ZINC22973642 with the lowest binding energies were subjected to molecular dynamics simulation in order to analyze the flexibility and stability of the protein-ligand complexes in a cellular atmosphere. The changes in the complex structure and conformation were assessed for a simulation time frame of 100 ns through MD simulations. Differentparameters like RMSD, RMSF, R_g, SASA were determined to understand the stability of the molecular trajectory, flexibility, ligand-receptor affinity and the extent of compactness and folding behavior. Figure 2 shows the pose of respective ligand during MD simulations in the active pocket at 25, 50, 75 and 100 ns, respectively. Supplementary Figure S3 summarizes the results obtained. RMSD evaluates whether the complex system has equilibrated and attained stability over the time duration of the simulation. In the case of apo-protein, the RMSD values showed a general increasing trend from 0 to 1.6 ns with RMSD values from 0 to 0.194 nm. Thereafter, the values showed slight variations of small magnitude. Towards the end of the simulation, particularly after 50 ns, a fairly constant value that remained between 0.2 and 0.24 nm was obtained. Considering the ZINC22973642 compound, the RMSD values showed a general increasing trend till 19.68 ns, with RMSD

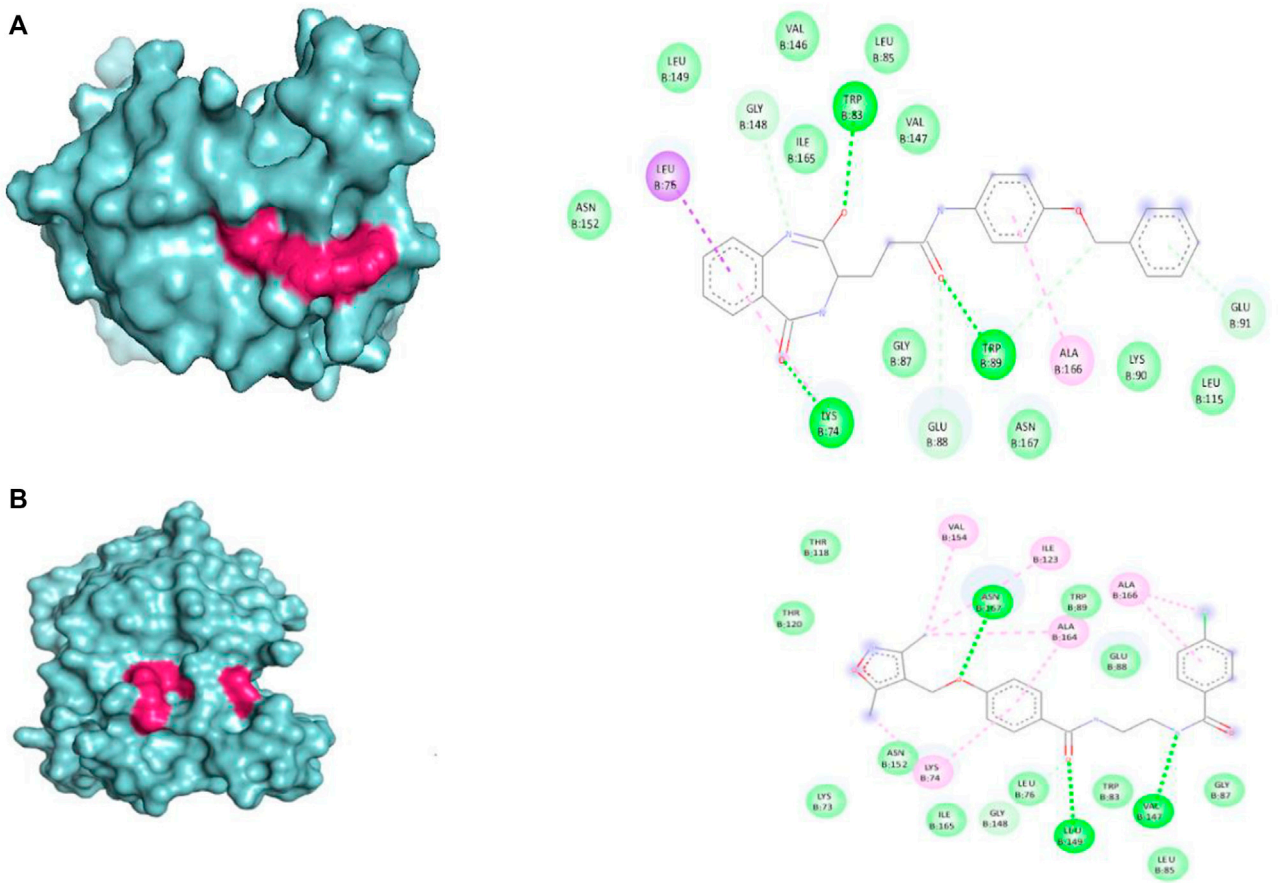


FIGURE 1
Visual representation of the docked complexes and the amino acid interactions of **(A)** ZINC36596404 **(B)** ZINC22973642.

values ranging from 0 to 0.27 nm. From this point ahead, the values remained fairly constant in the range between 0.2 and 0.24 nm. The compound ZINC36596404 showed relatively better stability, as the results show an increase followed by decreasing trend until around 30 ns and thereafter remains at an almost constant value of 0.23 nm with only slight variations.

Root mean square fluctuation analysis

RMSF values for C_α atoms were calculated and comparatively analyzed for the ligand-bound complexes along with that of the apo-protein in order to look into the mean residual fluctuations, motion, and flexibility of the amino acid residues of particular regions of the ligand binding during the simulation time. Supplementary Figure S4 shows the results obtained. It was observed that about seven amino acids (Gly62, Val72, Lys104, Gly114, Gly121, Pro132, Gly153) are directly involving in the complex formation *via* interactions like conventional hydrogen bonds, carbon hydrogen bonds, Pi-donor hydrogen bond,Pi-sigma,Pi-alkyl, Van der Waals, *etc.* From the figure we can see that these residues are decreased in the complex due to the ligand binding properties when compared to their free dynamics in the apoprotein. From this, it is understood that the apo-protein, ZINC22973642, and ZINC36596404 show a very similar pattern where maximum residues show fluctuations, however, the vacillation was less than 0.3 nm for a majority of these residues.

Radius of gyration (R_g)

The radius of gyration refers to the root mean square distance of the atoms from their rotational axis. It helps to gatherdetails about the compactness, rigidity, and folding behavior of the receptor during the time frame of the simulation. Lower R_g valuesshow that minimal fluctuations indicate a stable protein-ligand complex. Higher R_g values along with variation suggests instability of the complex. The values of R_g obtained are pictorially represented in Supplementary Figure S5. ZINC22973642 and ZINC36596404 happen to show a similar R_g pattern where the value remains fairly constant at 1.65 nm with very minor variations. From these results, we can conclude that the protein attained a compact state and does not show abrupt fluctuations indicating that a stably folded protein is formed upon binding of ligands to the ZINC database compounds.

Solvent accessible surface area

The binding of small molecules to receptor protein induces certain structural and conformational changes which have an impact on the protein volume. This change can indirectly give an insight

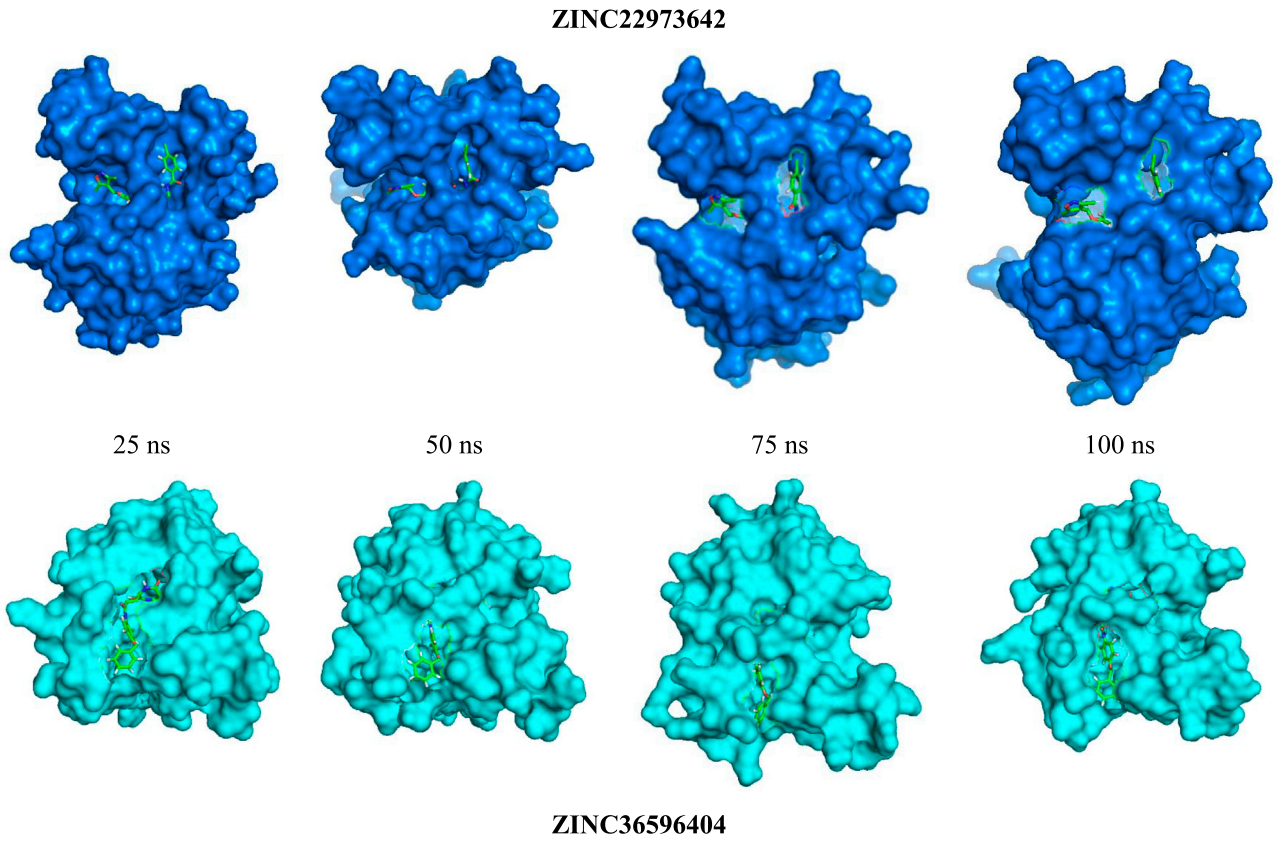


FIGURE 2
RMSD study of top 2 ligands for 100 ns MD Simulation.

into the protein-ligand complex during the simulation. SASA was calculated to look into the solvent behavior of the dengue protease upon binding to the ligands and it was comparatively analyzed to the changes in surface area of the apo-protein. Hydrophobic residues contribute to SASA values. The exposure of these residues from their hydrophobic core region leads to complex instability by decompressing the receptor. Similar to R_g, lower and minimal fluctuations in the values indicated stabilized, compressed and correctly folded target protein. The SASA values were calculated and plotted against time in Supplementary Figure S6. The apo-protein exhibited minimal fluctuations in SASA values until around 50,000 ps from where it started increasing up until 60,000 ps and further decreased until the values stabilized. Both the ZINC database compounds showed a closely similar pattern of minimal fluctuations in the SASA values throughout the simulation period.

Hydrogen bonds

The binding affinity of identified small molecules with the target protein can be ascertained by hydrogen bond formation. The number of hydrogen bonds formed between ligand and dengue protease revealed the binding affinity. Graphical results were presented in Supplementary Figure S7. ZINC22973642 showed an average binding affinity with the protein and formed a maximum of 7 hydrogen bonds throughout the simulation period. ZINC36596404 had higher binding energy with the protein and this is clearly explained by the consistent hydrogen bond formation with the protein. From the figure, we can see that the ZINC compounds consistently maintain at least 5 hydrogen bonds throughout the simulation period. The residues involved in hydrogen bonding in ZINC36596404 were Lys74, Trp83 and Trp89 which were involved in a conventional hydrogen bond, Gly148, Glu88 and Glu91 which were involved in carbon-hydrogen bond and pi-donor hydrogen bond. Similarly, for ZINC22973642, Asn167, Leu149, Val47 contributed to conventional hydrogen bonding, and, Gly148, Leu76, Trp83, Gly87, Leu85 for hydrogen bonds. The complexes eventually stabilized, as it can be interpreted from the structural parameters.

MM-PBSA binding free energy

One of the widely accepted methods for estimation of binding free energy of small ligands with biological macromolecules is Molecular Mechanics Poisson Boltzmann Surface Area continuum solvation (MM-PBSA). The energy values obtained were summarized in Table 7. For both the ZINC database compounds, SASA energy contributed more significantly towards the binding as compared to Electrostatic energy and van der Waal energy. In both cases, polar solvation energy seems to be positively influencing the binding and hence we can say that it does not favorably benefit the binding. In conclusion, the results of the

TABLE 7 MM-PBSA values of the two complexes after 100 ns simulation.

S. No.	Energy terms (KJ/mol)	ZINC22973642	ZINC36596404
1	Van der Waal	−241.848 ± 0.791	−250.309 ± 1.106
2	Electrostatic	−87.760 ± 1.045	−104.692 ± 1.163
3	Polar solvation	220.611 ± 16.207	261.191 ± 22.538
4	SASA	−23.200 ± 0.055	−23.788 ± 0.072
5	Binding energy	−132.196 ± 16.764	−116.651 ± 21.635

TABLE 8 Drug-likeness and ADMET properties of top 2 compounds.

S.No.	Parameter	ZINC22973642	ZINC36596404
1	Number of Hydrogen Bond Acceptors	5	5
2	Number of Hydrogen Bond Donors	2	3
3	BBB Score	2.85	2.22
4	Drug-likeness model score	1.1	0.43
5	Solubility	3.44e-05	4.09e-05
6	GI absorption	High	High
7	CYP1A2 inhibitor	Yes	No
8	CYP2C19 inhibitor	Yes	Yes
9	CYP2C9 inhibitor	Yes	Yes
10	CYP2D6 inhibitor	Yes	Yes
11	CYP3A4 inhibitor	Yes	Yes
12	Log Kp (skin permeation)	−6.42 cm/s	−6.67 cm/s
13	Bioavailability score	0.55	0.55
14	LD50	586 mg/Kg	3000 mg/kg
15	Toxicity class	4	5

molecular dynamics simulation show that both ZINC36596404 and ZINC22973642 have a good affinity and binding stability towards the targeted dengue protease.

Prediction of drug likeliness and ADMET properties

The drug-likeness of ZINC36596404 predicted using Molsoft showed a score of 0.43. From the results, 5 hydrogen bond acceptors and 3 hydrogen bond donors were also identified. The BBB score was reported as 2.22 which is on the lower side. The drug-likeness of ZINC22973642 analyzed by Molsoft had a score of 1.10. This drug-likeness score is predicted Molsoft's chemical fingerprints made using a dataset containing 5,000 marketed drugs and 10,000 non-drug compounds. The drug-likeness value ranges from −1 to +1, where values equal to or less than 0 indicates that the compound does not seem to be a likely drug, whereas values greater than 0 indicate good drug-likeness of the compound. Since both the compounds discussed here have positive drug-likeness scores, we can say that they seem to be drug-like. The results also identified 5 hydrogen bond acceptors and 2 hydrogen bond donors. The BBB score was 2.85 and is on the lower side, similar to the previous compound. ZINC36596404 belongs to toxicity class 5 indicating that it may be harmful if swallowed (2000 < LD50 ≤ 5,000) and ZINC22973642 to class 4 signifying that it may be harmful if swallowed (300 < LD50 ≤ 2000) as per predictions made by ProtoxII. The ADME results obtained from SwissADME are shown in Table 8. ZINC22973642 shows no violation of Lipinski's rule of five. It is seen to have good GI absorption, good solubility, and low BBB permeability indicating that it does not cross the blood-brain barrier. It is seen to inhibit CYP1A2, CYP2C19, CYP2C9, CYP2D6, and CYP3A4 which are cytochrome enzymes involved in the detoxification and metabolism of drugs. The skin permeation parameter for this compound indicates that it is moderately good for topical applications. Its bioavailability score

shows that it is sufficiently absorbable and available throughout the body when administered *via* the oral route. The predicted LD50 is also sufficiently high. This, coupled with a good drug-likeness score, makes this compound a very potent lead that can be further explored and developed into an efficient drug against dengue protease. ZINC36596404 also shows similar properties as that of ZINC22973642, but only differs in that it does not inhibit CYP1A2. The fact that these two ZINC compounds showed good binding stability and affinity to Dengue Protease, combined with their positive drug-likeness, show that these compounds can be studied further *in vitro* in order to develop them into effective anti-Dengue drugs.

Conclusion

In this study, a ligand-based QSAR and pharmacophore model of Dengue protease inhibitors was developed using 4-Benzyloxy Phenyl Glycine derivatives. The GATS6e, GATS5i, VE1_DzZ, VE2_DzZ, VE3_DzZ, SpMAD_Dzp, SpMax3_Bhp, ETA_Epsilon_5, IC1, IC2, TIC0, MIC1, WTPT-3 descriptors were seen to have an effect on the anti-dengue protease activity. The validated QSAR model showed significant statistical parameters and can be used to predict the activity of unknown compounds for anti-dengue protease activity. Using this QSAR model and the pharmacophore features presented above, other 4-Benzyloxy Phenyl Glycine derivatives can be modified to enhance their activities. This model can be a helpful tool to reduce the time and expense involved in dengue protease antagonist synthesis and activity determination. Further, the molecular docking and dynamics simulation studies performed using the compounds identified from the ZINC database have indicated that ZINC36596404 and ZINC22973642 show excellent binding with the dengue protease. The complexes also show structural stability. They also have good drug-likeness and compatible ADMET properties. It can be inferred that these two compounds form promising candidates in the development of dengue protease antagonists. Further work that aims to test the *in vitro* and *in vivo* effects of these two compounds is required in order to validate these results. Thus, our findings, coupled with laboratory testing of the identified potential leads can help to develop strong antagonists for dengue protease.

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

AP-formal analysis, writing–original draft. PP-formal analysis, writing–original draft. TM-conceptualization, methodology, data curation, writing–review and editing. MM-conceptualization, methodology, data curation, writing–review and editing. SK-formal analysis, writing–original draft, writing-review and editing. MS-formal analysis, writing–original draft, AR-formal analysis, data curation, writing–review and editing.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2023.1106128/full#supplementary-material

References

Adawara, S. N., Shallangwa, G. A., Mamza, P. A., and Ibrahim, A. (2020). Molecular docking and QSAR theoretical model for prediction of phthalazinone derivatives as new class of potent dengue virus inhibitors. *Beni-Suef Univ. J. Basic Appl. Sci.* 9. doi:10.1186/s43088-020-00073-9

Behnam, M. A. M., Graf, D., Bartenschlager, R., Zlotos, D. P., and Klein, C. D. (2015). Discovery of nanomolar dengue and west nile virus protease inhibitors containing a 4-benzyloxyphenylglycine residue. *J. Med. Chem.* 58 (23), 9354–9370. doi:10.1021/ACS.JMEDCHEM.5B01441

Daina, A., Michielin, O., and Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* 7 (11), 42717–42813. doi:10.1038/srep42717

Drwal, M. N., Banerjee, P., Dunkel, M., Wettig, M. R., and Preissner, R. (2014). ProTox: A web server for the *in silico* prediction of rodent oral toxicity. *Nucleic Acids Res.* 42 (W1), W53–W58. doi:10.1093/NAR/GKU401

Dwivedi, V. D., Arya, A., Yadav, P., Kumar, R., Kumar, V., and Raghava, G. P. S. (2021). DenvInD: Dengue virus inhibitors database for clinical and molecular research. *Briefings Bioinforma.* 22 (3), bbaa098. doi:10.1093/BIB/BBAA098

Elsherif, M. A., Hassan, A. S., Moustafa, G. O., Awad, H. M., and Morsy, N. M. (2020). Antimicrobial evaluation and molecular properties prediction of pyrazolines incorporating benzofuran and pyrazole moieties ARTICLE INFO. *J. Appl. Pharm. Sci.* 10 (02), 37–043. doi:10.7324/JAPS.2020.102006

Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeerschd, T., Zurek, E., and Hutchison, G. R. (2012). Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics* 4 (8), 17. doi:10.1186/1758-2946-4-17

Hollas, B. (2003). An analysis of the autocorrelation descriptor for molecules. *J. Math. Chem.* 3333 (22), 91–101. doi:10.1023/A:1023247831238

Karthikeyan, B. S., Ravichandran, J., Aparna, S. R., and Samal, A. (2021). DEDuCT 2.0: An updated knowledgebase and an exploration of the current regulations and guidelines from the perspective of endocrine disrupting chemicals. *Chemosphere* 267, 128898. doi:10.1016/J.CHEMOSPHERE.2020.128898

Koes, D. R., and Camacho, C. J. (2012). ZINCPharmer: Pharmacophore search of the ZINC database. *Nucleic Acids Res.* 40, W409–W414. Web Server issue). doi:10.1093/NAR/GKS378

Kurniawan, I., Rosalinda, M., and Ikhsan, N. (2020). Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent. *SAR QSAR Environ. Res.* 31 (6), 477–492. doi:10.1080/1062936X.2020.1773534

Li, Z., Wan, H., Shi, Y., and Ouyang, P. (2004). Personal experience with four kinds of chemical structure drawing software: Review on chemdraw, chemwindow, ISIS/draw, and chemsketch. *J. Chem. Inf. Comput. Sci.* 44 (5), 1886–1890. doi:10.1021/ci049794h

Luo, P. H., Zhang, X. R., Huang, L., Yuan, L., Zhou, X. Z., Gao, X., et al. (2017). 3D-QSAR pharmacophore-based virtual screening, molecular docking and molecular dynamics simulation toward identifying lead compounds for NS2B–NS3 protease inhibitors. *J. Recept. Signal Transduct.* 37 (5), 481–492. doi:10.1080/10799893.2017.1358283

Murahari, M., Kharkar, P. S., Lonikar, N., and Mayur, Y. C. (2017). Design, synthesis, biological evaluation, molecular docking and QSAR studies of 2,4-dimethylacridones as anticancer agents. *Eur. J. Med. Chem.* 130, 154–170. doi:10.1016/j.ejmech.2017.02.022

Pawar, S. S., and Rohane, S. H. (2021). Review on Discovery Studio: An important tool for molecular docking. *Asian J. Res. Chem.* 14 (1), 1–3. doi:10.5958/0974-4150.2021.00014.6

Qamar, M. T., Ashfaq, U. A., Tusleem, K., Mumtaz, A., Tariq, Q., Goheer, A., et al. (2017). In-silico identification and evaluation of plant flavonoids as dengue NS2B/NS3 protease inhibitors using molecular docking and simulation approach. *Pak. J. Pharm. Sci.* 30 (6), 2119–2137.

Roy, K., and Ghosh, G. (2003). Introduction of extended topochemical atom (ETA) indices in the valence electron mobile (VEM) environment as tools for QSAR/QSPR studies. In Internet Electronic Journal of Molecular Design.

Saqallah, F. G., Abbas, M. A., and Wahab, H. A. (2022). Recent advances in natural products as potential inhibitors of dengue virus with a special emphasis on NS2b/NS3 protease. *Phytochemistry* 202, 113362. doi:10.1016/j.phytochem.2022.113362

Sarwar, M. W., Riaz, A., Dilshad, S. M. R., Al-Qahtani, A., Nawaz-Ul-Rehman, M. S., and Mubin, M. (2018). Structure activity relationship (SAR) and quantitative structure activity relationship (QSAR) studies showed plant flavonoids as potential inhibitors of dengue NS2B-NS3 protease. *BMC Struct. Biol.* 18 (1), 6. doi:10.1186/s12900-018-0084-5

Schmid, N., Eichenberger, A. P., Choutko, A., Riniker, S., Winger, M., Mark, A. E., et al. (2011). Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophysics J.* 40 (7), 843–856. doi:10.1007/s00249-011-0700-9

Schüttelkopf, A. W., and Van Aalten, D. M. (2004). PRODRG: A tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallogr. D Biol. Crystallogr.* 60 (8), 1355–1363. doi:10.1107/S0907444904011679

Seeliger, D., and De Groot, B. L. (2010). Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J. Computer-Aided Mol. Des.* 24 (5), 417–422. doi:10.1007/s10822-010-9352-6

Singh, V. K., Chaurasia, H., Mishra, R., Srivastava, R., Naaz, F., Kumar, P., et al. (2022). Docking, ADMET prediction, DFT analysis, synthesis, cytotoxicity, antibacterial screening and QSAR analysis of diarylpyrimidine derivatives. *J. Mol. Struct.* 1247, 131400. doi:10.1016/J.MOLSTRUC.2021.131400

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. (2005). GROMACS: Fast, flexible, and free. *J. Comput. Chem.* 26 (16), 1701–1718. doi:10.1002/jcc.20291

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi:10.1093/nar/gky427

Wellekens, K., Betrains, A., De Munter, P., and Peetermans, W. (2022). Dengue: Current state one year before WHO 2010-2020 goals. *Acta Clin. Belg.* 77 (2), 436–444. doi:10.1080/17843286.2020.1837576

WHO (2022). *Dengue and severe dengue*. Geneva, Switzerland: World Health Organization.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46 (D1), D1074–D1082. doi:10.1093/NAR/GKX1037

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32 (7), 1466–1474. doi:10.1002/JCC.21707

Yuan, S., Chan, H. C. S., and Hu, Z. (2017). Using PyMOL as a platform for computational drug design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 7 (2), e1298. doi:10.1002/WCMS.1298

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 04 April 2023
DOI 10.3389/fmed.2023.1154417

OPEN ACCESS

EDITED BY
Surapaneni Krishna Mohan,
Panimalar Medical College Hospital and Research Institute, India

REVIEWED BY
Feifei Pu,
Huazhong University of Science and Technology, China
Gianfranco Pintus,
University of Sharjah, United Arab Emirates
Noor Ahmad Shaik,
King Abdulaziz University, Saudi Arabia

*CORRESPONDENCE
Karthick Vasudevan
✉ karthick.1087@gmail.com
Achraf El Allali
✉ Achraf.ELALLALI@um6p.ma

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

RECEIVED 30 January 2023
ACCEPTED 20 March 2023
PUBLISHED 04 April 2023

CITATION
Dey H, Vasudevan K, Doss C. GP, Kumar SU, El Allali A, Alsamman AM and Zayed H (2023) Integrated gene network analysis sheds light on understanding the progression of Osteosarcoma. *Front. Med.* 10:1154417. doi: 10.3389/fmed.2023.1154417

Integrated gene network analysis sheds light on understanding the progression of Osteosarcoma

Hrituraj Dey[1], Karthick Vasudevan[1]*, George Priya Doss C.[2], S. Udhaya Kumar[2], Achraf El Allali[3]*, Alsamman M. Alsamman[4,5] and Hatem Zayed[6]

[1]Department of Biotechnology, School of Applied Sciences, REVA University, Bangalore, India, [2]Department of Integrative Biology, School of BioSciences and Technology, Vellore Institute of Technology (VIT), Vellore, India, [3]African Genome Center, Mohammed VI Polytechnic University, Ben Guerir, Morocco, [4]Agriculture Genetic Engineering Research Institute (AGERI), Agriculture Research Center (ARC), Giza, Egypt, [5]International Center for Agricultural Research in the Dry Areas (ICARDA), Giza, Egypt, [6]Department of Biomedical Sciences College of Health Sciences, QU Health, Qatar University, Doha, Qatar

Introduction: Osteosarcoma is a rare disorder among cancer, but the most frequently occurring among sarcomas in children and adolescents. It has been reported to possess the relapsing capability as well as accompanying collateral adverse effects which hinder the development process of an effective treatment plan. Using networks of omics data to identify cancer biomarkers could revolutionize the field in understanding the cancer. Cancer biomarkers and the molecular mechanisms behind it can both be understood by studying the biological networks underpinning the etiology of the disease.

Methods: In our study, we aimed to highlight the hub genes involved in gene-gene interaction network to understand their interaction and how they affect the various biological processes and signaling pathways involved in Osteosarcoma. Gene interaction network provides a comprehensive overview of functional gene analysis by providing insight into how genes cooperatively interact to elicit a response. Because gene interaction networks serve as a nexus to many biological problems, their employment of it to identify the hub genes that can serve as potential biomarkers remain widely unexplored. A dynamic framework provides a clear understanding of biological complexity and a pathway from the gene level to interaction networks.

Results: Our study revealed various hub genes viz. TP53, CCND1, CDK4, STAT3, and VEGFA by analyzing various topological parameters of the network, such as highest number of interactions, average shortest path length, high cluster density, etc. Their involvement in key signaling pathways, such as the FOXM1 transcription factor network, FAK-mediated signaling events, and the ATM pathway, makes them significant candidates for studying the disease. The study also highlighted significant enrichment in GO terms (Biological Processes, Molecular Function, and Cellular Processes), such as cell cycle signal transduction, cell communication, kinase binding, transcription factor activity, nucleoplasm, PML body, nuclear body, etc.

Conclusion: To develop better therapeutics, a specific approach toward the disease targeting the hub genes involved in various signaling pathways must have opted to unravel the complexity of the disease. Our study has highlighted the candidate hub genes viz. TP53, CCND1 CDK4, STAT3, VEGFA. Their involvement in the major signaling pathways of Osteosarcoma makes them potential candidates to be targeted for drug development. The highly enriched signaling pathways include FOXM1 transcription pathway, ATM signal-ling pathway, FAK mediated

signaling events, Arf6 signaling events, mTOR signaling pathway, and Integrin family cell surface interactions. Targeting the hub genes and their associated functional partners which we have reported in our studies may be efficacious in developing novel therapeutic targets.

KEYWORDS

Osteosarcoma, gene interaction network, hub genes, TP53, FOXM1 transcription factor

1. Introduction

The prominence of Osteosarcoma dates back to the early nineteenth century when the French surgeon Alexis Boyer first coined the term and William Enneking described the disease. A recent study by the American Cancer Society found that 186.6 per 100,000 children and adolescents were diagnosed with cancer each year from birth to age 19 (1). Osteosarcoma is the most common type of bone cancer, originating in the mesenchyme tissue. The tumor usually develops around the pelvis or long bone and then metastasizes to neighboring tissue (2). The most prevalent locations in femur (42%, of which 75% is in the distal femur), the tibia (19%, of which 80% is in the proximal tibia), and the humerus (10%). The jaw or skull (8%) and the pelvis (8%) are additional potential sites. In the ribs, Osteosarcomas only comprise 1.25 percent of cases (3), (4).

Although it is seen in both young and adults, it has been observed that the tumor spreads rapidly when the bone undergoes the stages of its growth. It has a bimodal age distribution with an adolescence and elderly peak in incidence. The incidence often peaks between the ages of 10 and 14 years, after which it starts to subside. Adults over 65 see the second peak in Osteosarcoma incidence, more likely to be a second malignancy commonly linked to Paget disease (5). The genomic landscape of Osteosarcoma based on various sequencing methods revealed that alterations in the sequence are due to somatic point mutations such as single base substitutions, insertions, and deletions. Other structural variants such as rearrangements and somatic copy number alterations leading to copy number decrease may downregulate a tumor suppressor gene driver and copy number increase may trigger an oncogene driver (6, 7). Numerous familial syndromes are associated with Osteosarcoma. Li-Fraumeni syndrome is one such condition with a high prevalence of Osteosarcoma. This condition is characterized by various malignancies, including leukemia, breast, sarcoma, adrenocortical, and brain tumors (8). It is an autosomal dominant disorder where the p53 tumor suppressor gene is rendered inactive, which helps advance the cell cycle in the presence of DNA damage.

Additionally, it has been demonstrated that alteration in additional p53 pathway genes, such as *MDM2, p14ART, and CDK4*, may increase a person's risk of acquiring Osteosarcoma (9). DNA helicase anomalies have also been reported in Osteosarcoma. In the autosomal recessive disorder, Rothman-Thomas syndrome, which is associated with skin changes, alopecia, and Osteosarcoma, gene *RECQL4* coding for DNA helicase is found to be defective. Similar DNA helicase aberrations are found in Werner syndrome where the *WRN or RECQL4* gene is defective causing melanoma, Osteosarcoma, etc. (10).

During the mid-1970s, chemotherapy was shown to be adequate for treating Osteosarcoma. Osteosarcoma is typically treated with neoadjuvant chemotherapy that includes cisplatin, doxorubicin, ifosfamide, and high-dose methotrexate given along with leucovorin. This is followed by surgical resection and adjuvant chemotherapy (11). Although the current treatment regime has proven to be partially effective, it is associated with short- and long-term concomitant side effects such as accumulating toxic compounds in other organs such as the liver, kidney, heart, etc., leading to other detrimental effects on the body. For instance, higher dosage rates were linked to an increased risk of nephrotoxicity and gonadal dysfunction brought on by cisplatin. The dosage intensity and the total dose of doxorubicin were associated with an increased risk of cardiac toxicity (12). Thus, the hub genes involved in the various enriched biological processes and signaling pathways must be identified to develop better treatment strategies. These hub genes are essential because they play a role in regulating the molecular mechanism. Our study aimed to highlight the hub genes involved in the gene-gene interaction network to understand their interaction and how they affect the various biological processes and signaling pathways involved in Osteosarcoma. Using networks of omics data to identify cancer biomarkers could revolutionize the field. Cancer biomarkers and the molecular mechanisms behind it can both be understood by studying the biological networks underpinning the disease (13). Several network-based analysis tools were used for biomarker identification in recent years. For instance, a gene co-expression network (GCN) was developed to effectively identify biomarkers in glioma. It was also utilized to assess a gene module relevant to lung cancer, predictive biomarkers for estrogen receptor-positive breast cancer treated with tamoxifen, and biomarkers for anticipating the chemotherapy response in breast cancer (14–16). In our earlier studies, we have used advanced computational tools to decipher and predict the pathogenicity of the various diseases (17, 18). Gene interaction network provides a broad view of functional gene analysis by giving an insight into how genes cooperatively interact to elicit a response. A dynamic framework offers a clear understanding of biological complexity and a pathway from the gene level to interaction networks. The term "interaction" refers to the relationship between genes that can affect other genes' operations. Because gene interaction networks serve as a nexus to many biological problems, their employment of it to identify the hub genes that can serve as potential biomarkers remain widely unexplored. Gene interaction network assists in identifying novel candidate genes, based on the idea that the neighboring genes located near the disease-causing gene in a network are more likely to cause a similar disease (19).

2. Materials and methods

2.1. Cancer genetics web server

The Cancer genetics web server is an online resource portal, which provides information on various cancers, particularly for researchers and health professionals exploring this field. Sever is available at www.cancer-genetics.org/. Using PubMed, the data were obtained by utilizing information from numerous data sources and literary reviews. It offers comprehensive links to credible information about genes, their associated proteins, and genetic alterations linked to cancer and related disorders. The site includes a directory of genes identified as the oncogenes and the tumor suppressor genes. Every gene page includes accessible links to major genetic databases and abstracts, references, external searches, and summary information wherever possible.

2.2. STRING database

STRING (Search tool for retrieval of the interacting gene) (https://string-db.org/) is an online, publicly accessible database harboring information on protein-protein interaction. The interactions include direct as well as indirect connections. It provides a versatile way for analyzing and visualizing the data, such as setting confidence scores that reflect the level of interaction, no of interactors, network type, display mode, etc. In order to categorize the interactions, String uses the confidence scores: highest (above 0.90), high (0.7–0.89), medium (0.4–0.69), and low (0.15–0.39). The STRING database accepts the input in various forms, such as protein by name, protein by sequence, multiple proteins, protein families (COG), etc. The outcome of the network can be saved in a variety of formats such as bitmap image, vector graphic, TSV format, tab-delimited file, etc. (20).

2.3. Cytoscape

Gene interaction networks can be visualized and analyzed using Cytoscape (https://cytoscape.org/). It provides a user-friendly interface that allows the user to seamlessly operate the software. It supports various plugins, which serve various purposes such as clustering of genes, enrichment analysis, annotation, determining topological properties of a network, etc. Output from STRING was used as an input for the Cytoscape.

2.3.1. Network analyzer

It is a plugin in Cytoscape that calculates topological parameters in a network. Numerous parameters can be computed, such as degree, number of nodes, edges, average no. of neighbors, clustering coefficient, average shortest path length, closeness centrality, and betweenness centrality. The degree and average shortest length are the essential parameters while analyzing the network since the degree represents the direct interactors of the desired gene, whereas the average shortest path is the distance between two nodes. Closeness centrality measures how fast information travels from one node to another node in a network, whereas betweenness centrality represents the degree of influence a node exerts upon other interactions of a node (21). The results generated can be exported as a CSV file or directly analyzed in the software.

2.3.2. MCODE

MCODE is a plugin used to identify clusters in a network. Clusters are highly interrelated regions that are grouped in a network. The MCODE method is based on analyzing densely interconnected regions where nodes have more interconnected nodes, detecting potential clusters, and evaluating the number of interconnected nodes (node scoring). Genes are clustered by MCODE based on their connectivity, in which the same cluster contains more interconnected genes with the optimal neighborhood density. Genes that are associated with MCODE scores are clustered together. (22).

2.4. FunRich

FunRich (http://www.funrich.org/) is a tool used for functional enrichment and network analysis. It can be utilized to conduct functional enrichment analysis on background databases incorporating diverse genomic and proteomic resources. The outcomes of the enrichment studies may be depicted using a wide range of graphical layouts, such as column graphs, bar graphs, pie charts, Venn diagrams, heat maps, and doughnut charts. Users can download information from the UniProt and standard human-specific FunRich databases. Additionally, users can create their custom datasets and carry out enrichment analyses regardless of the organism (23).

3. Results

3.1. Data collection

The genes for Osteosarcoma responsible for its growth and development were curated from Cancer genetics web database. The sites host information on genes for 76 different cancers and associated conditions. The information on genes related to Osteosarcoma was searched based on the keywords. We were able to gather 58 genes and their related information. This data was used for a STRING interaction network. The interaction network was maximized, with a medium confidence score (0.4) which gave an interaction for 71 genes and their functional partners. Gene networks were constructed and further analyzed based on STRING interaction data (Figure 1).

3.2. Network analysis

The network analysis of 71 genes was carried out using NetworkAnalyzer. To study the gene interaction network, it analyzed different topological parameters such as degree, no of nodes and edges, characteristic path length, clustering coefficient, closeness centrality, and betweenness centrality. The top genes with the highest degree values are *TP53*, *CCND1*, *CDK4*, and *STAT3* with no interactors 45, 33, 28, and 27, respectively. Table 1 lists the 20 genes along with their various analyzed parameters. The network

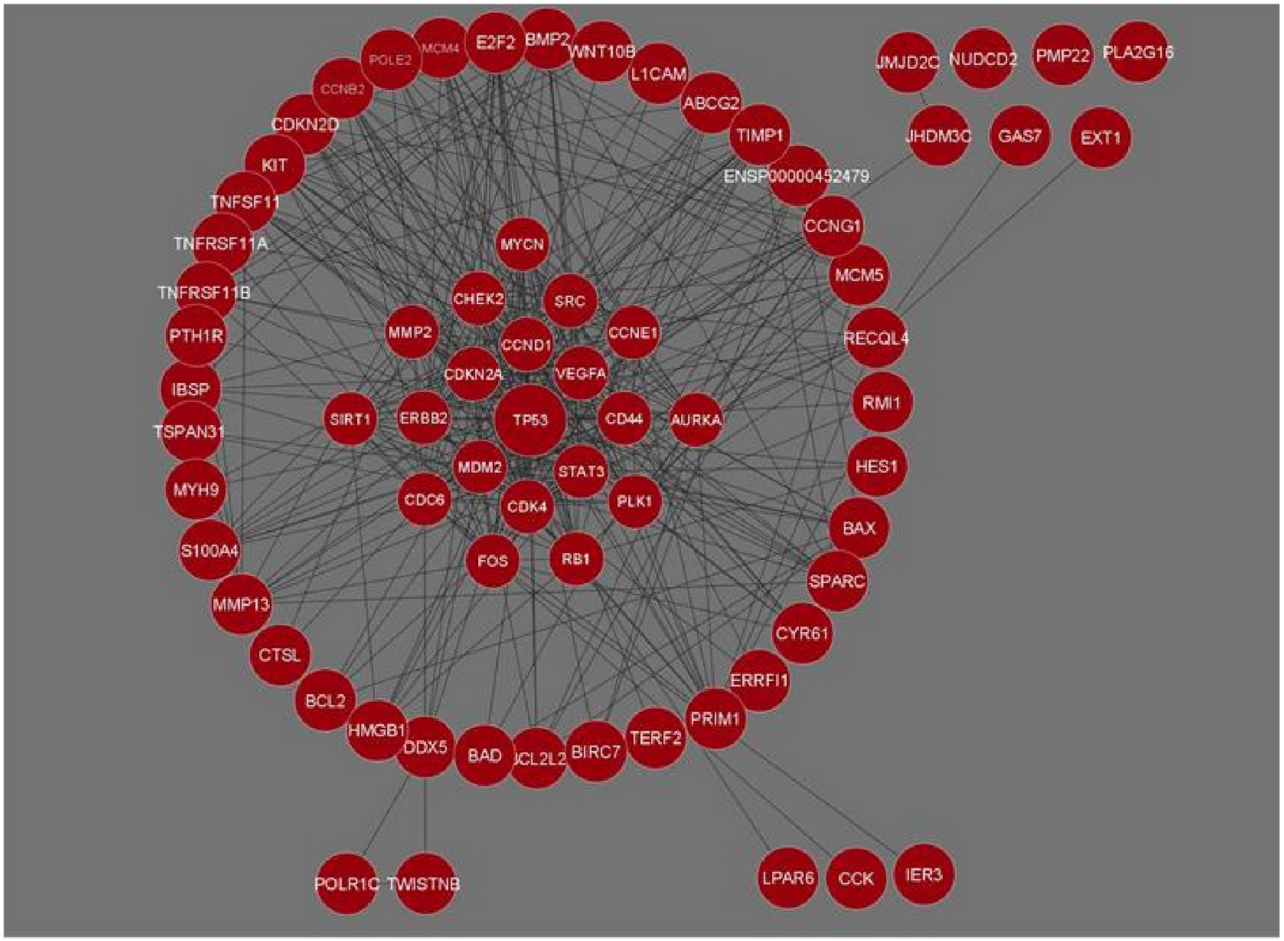


FIGURE 1
Gene interaction network of Osteosarcoma comprising 71 genes and 426 interactions built in Cytoscape. Genes with the maximum number of network interactions are positioned in the center (ENSP00000452479 is the Ensembl protein ID for sequence BCL2L2-PABPN1).

analysis revealed the no. of nodes to be 71 and no. of edges to be 426, while the clustering coefficient of the entire network was 0.583.

3.3. Clustering analysis

Clustering analysis of the gene interaction network was done using MCODE, resulting in the genes in a cluster of 3. *viz.* C1, C2, and C3 (Figure 2). The clustering of genes allowed us to understand the highly interconnected regions. Clustering of MCODE is influenced by both directed interactions and interactions between the associated interactors. Out of 71 genes in the network, 36 are identified as part of the cluster. Among the three, cluster C1 had the most inter-connected regions constituting 24 nodes and 137 edges with an MCODE score of 11.913, followed by C2 with five nodes and ten edges with a score of 5.0, and C3 with 7 nodes and 14 edges with a score of 4.667 (Table 2).

3.4. Functional enrichment analysis

Following clustering analysis, functional enrichment analysis was performed using the STRING database and FunRich tool, clarifying genes' contribution to various processes and pathways. The Bonferroni correction method obtained Gene ontology terms with a p-value $\leq$ 0.05. Using the Bonferroni correction, multiple comparisons are compensated by dividing the significance level by the number of comparisons. A significance level indicates the likelihood that a given test will detect an incorrect difference in the sample that does not exist in the population (false positive). Commonly, significance levels of 0.05 are considered significant. The genes observed in Osteosarcoma revealed various contributions in Gene Ontology terms such as Biological Processes (BP), Molecular Function (MF), and Cellular Compartment (CC). The significantly enriched terms in BP included regulation of cell cycle signal transduction, cell communication, regulation of nucleobase, nucleoside, nucleotide, and nucleic acid metabolism (Supplementary File 1), MF included kinase binding, kinase regulator activity, transcription factor activity (Supplementary File 2) and CC included nucleoplasm, PML body, nuclear body, nucleus, and cytosol (Supplementary File 3; Figure 3). The enriched signaling pathways is of utmost importance while studying the progression of cancer. Cancer involves various signal transmission pathways, which promote its progression. The signaling pathways involved in tumor progression of Osteosarcoma are the FOXM1 transcription pathway, ATM signaling pathway, FAK mediated signaling events, Arf6 signaling events, Class 1 PI3K signaling events, mTOR

TABLE 1 The list of the top 20 genes in Osteosarcoma analyzed by NetworkAnalyzer.

Genes	Degree	Avg. shortest path length	Closeness centrality	Betweenness centrality
TP53	45	1.343283582	0.744444	0.278506532
CCND1	33	1.582089552	0.632075	0.076460459
CDK4	28	1.626865672	0.6146788	0.047913393
STAT3	27	1.71641791	0.587798	0.033714855
VEGFA	27	1.701492537	0.582606	0.033879065
CDKN2A	27	1.641791045	0.609009	0.037095448
MDM2	26	1.701492537	0.587718	0.041774175
SRC	25	1.74626865	0.572643	0.053286042
CHEK2	24	1.746268657	0.572643	0.023052737
ERBB2	23	1.776119403	0.567713	0.036390641
RB1	23	1.76119403	0.563020	0.080214837
FOS	23	1.76119403	0.567794	0.051156614
CD44	22	1.805970149	0.553719	0.014110304
CCNE1	21	1.835820896	0.544715	0.007678749
PLK1	20	1.880597015	0.544715	0.025338494
MMP2	20	1.835820896	0.531746	0.009431072
CDC6	20	1.835820896	0.544715	0.075482361
MYCN	19	1.925373134	0.51937	0.012188444
AURKA	19	1.835820896	0.544715	0.004759471
SIRT1	18	1.776119403	0.563028	0.020802171

signaling pathway, and Integrin family cell surface interactions (Supplementary File 4; Figure 4). The genes involved in various signaling pathways of Osteosarcoma are mentioned in Table 3.

4. Discussion

A cancer cell will essentially have six hallmark capabilities to be recognized as a cancer cell. The six core hallmarks outlined by Hananah and Weinberg include self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis, along with the emerging hallmarks of cancer which includes deregulating cellular energetics and avoiding immune destruction (24, 25). Attaining each capability will likely involve inactivating or eluding a specific control mechanism. We have utilized a gene interaction network in our study to understand the development and progression of the tumor cells in Osteosarcoma. This helped us decipher a group of highly interactive genes responsible for the pathogenesis and spread of the disease.

During analysis, MF observed were kinase binding, kinase regulator activity, and transcription factor activity. Prior studies on Osteosarcoma have highlighted that protein tyrosine kinases are essential signaling molecules involved in the signaling pathways that regulate cellular differentiation and proliferation (26). The enriched BPs of Osteosarcoma included signal transduction, cell communication, regulation of cell cycle, regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism, apoptosis, protein metabolism, energy pathways, metabolism along with cell cycle checkpoint signaling, DNA damage checkpoint signaling, and response to hypoxia. Earlier studies have shown that impairment in signal transduction, cell communication, and cell cycle checkpoint signaling has significantly promoted Osteosarcoma (27). Signal transduction is a sequential event where an extracellular signal is transduced by the cell to create a response, which is necessary for the normal growth and development of the cell. Since genetic alterations drive cancer, these alterations create a wide range of aberrant signaling networks that drives the expansion of the tumor. These signaling pathways control tumor growth, development, and fate (28). The signal transduction pathway involved 14 genes namely *CCND1, CDK4, VEGFA, CDKN2A, SRC, CHEK2, ERBB2, CD44, CCNE1, PLK1, CDC6, AURKA, CCNB2,* and *TNFSF11*. It has been reported that patients suffering from Osteosarcoma cells develop resistance toward the kinase inhibitor drug, Sorafenib due to the mTOR signaling pathway. The mammalian target of rapamycin (mTOR) facilitates all cell proliferation, apoptosis, and autophagy. There is evidence showing that the mTOR signaling pathway plays a significant role in a number of diseases, including osteosarcoma. (29). The mTOR is structurally made up of a dimer complex called the mammalian target of rapamycin complex 1 (mTORC1) and the mammalian target of rapamycin complex 2

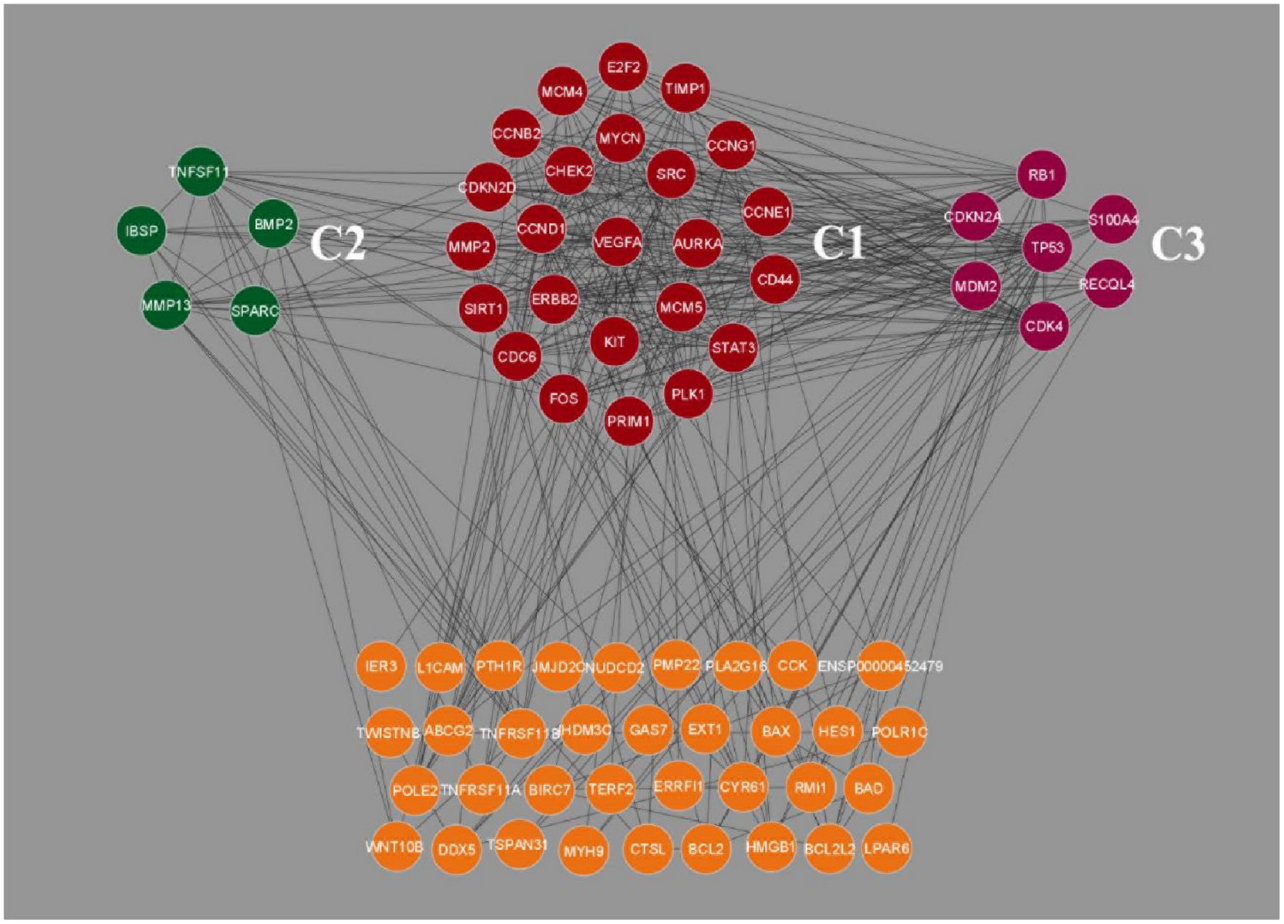


FIGURE 2
Clustering analysis of Osteosarcoma gene interaction network using MCODE. The genes were grouped into three clusters, *viz.* C1, C2 and C3. Cluster C1 showed the highest level of clustering, followed by C2 and C3. The unclustered genes are located beneath, highlighted in orange color.

(mTORC2) (30). mTORC1 has been mostly seen in controlling cell growth and metabolism, while mTORC2 primarily governs cell proliferation and survival (31). Numerous signaling pathways in the body, such as phosphoinositide-3-kinase (PI3K)/AKT, tuberous sclerosis complex subunit 1 (TSC1)/tuberous sclerosis complex subunit 2 (TSC2)/Rheb, LKBL/adenosine 5′monophosphate-activated protein kinase (AMPK), VAM6/Rag GTPases, and others, are regulated by mTOR (32). Under normal circumstances, mTOR plays a significant role in regulating cell growth and division. However, it is hyper-activated in tumor cells sending aberrant signals that help tumor cells grow and proliferate, thus promoting malignancy (33). mTOR pathway incessantly activates the AKT signaling pathway among the other pathways (34). Our study revealed 19 genes involved in the mTOR signaling pathway of Osteosarcoma viz. *TP53, CCND1, CDK4, STAT3, VEGFA, CDKN2A, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG1,* and *TNFSF11*. The involvement of mutated genes *TP53* and *VEGFA* is closely associated with all types of cancer. *TP53* controls cell growth and proliferation by acting as a tumor suppressor gene. The alteration in the sequence of TP53 leads to tumor development. *VEGFA* promotes the mTOR signaling in Osteosarcoma by promoting angiogenesis in the tumor (35). It has been studied that there is significant upregulation in mTORC1 during tumor growth and development, and mTORC1 is comparatively more sensitive to rapamycin than mTORC2. Thus, rapamycin acts as an inhibitor of mTOR (36). Different approaches can be sought, such as down-regulating the mTOR complexes to control cell proliferation. Because of its close linkage with Osteosarcoma, mTOR pathways, and the associated genes can serve as a therapeutic target for the disease. Cell communication was also seen to be significantly enriched in the Biological Processes. Communication between the neighboring cells is crucial for normal cellular activities. Numerous studies have demonstrated that a complex intercellular communication system, whether through direct cell-to-cell contact or traditional paracrine/endocrine signaling, plays a crucial role in the growth and expansion of tumors (37). The most basic signal transmission to the proximal or the distant cells is the release of soluble substances into the extracellular space, such as cytokines, chemokines, and growth factors. Along with it, another cell interaction involves adhesion molecules and gap junction (38). Recent studies have also demonstrated that healthy mitochondria and other organelles may be donated by non-cancer cells through tunnel nanotubes to keep cancer cells alive, but it has also been

TABLE 2 List of Osteosarcoma related genes and their associated signaling pathways.

Signaling pathways	Genes
FOXM1 transcription factor network	*CCND1, CDK4, CDKN2A, CHEK2, RB1, FOS, CCNE1, PLK1, MMP2,* and *CCNB2*
FAK-mediated signaling events	*TP53, CCND1, CDK4, STAT3, VEGFA, CDKN2A, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG,* and *TNFSF11*
ATM pathway	*TP53, CDKN2A, MDM2, CHEK2, RB1, CCNE1, PLK1, MMP2, CDC6, SIRT1, E2F2,* and *CCNG1*
Arf6 signaling events	*TP53, CCND1, CDK4, STAT3, VEGFA, CDKN2A, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG1,* and *TNFSF11*
Class I PI3K signaling events	TP53, CCND1, CDK4, STAT3, VEGFA, CDKN2A, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG1, and TNFSF11
mTOR signaling pathway	*TP53, CCND1, CDK4, STAT3, VEGFA, CDKN2A, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG1,* and *TNFSF11*
EGF receptor(ErbB1) signaling pathway	*TP53, CCND1, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CDK4, STAT3, VEGFA, CDKN2A, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG1,* and *TNFSF11*
Integrin family cell surface interactions	*TP53, CCND1, CDK4, STAT3, VEGFA, CDKN2A, MDM2, SRC, CHEK2, ERBB2, RB1, FOS, CCNE1, PLK1, MMP2, SIRT1, E2F2, CCNG1,* and *TNFSF11*

revealed that horizontal mitochondrial transfer from cancer cells to neighboring cells is equally possible (39).

The Integrin family of proteins binds extracellular matrix ligands and cell-surface ligands to act as cell adhesion receptors during cell communication. Our study of Osteosarcoma significantly enriches the integrin family of cell surface interactions. Integrins connect with the extracellular matrix (ECM) via the extracellular domain, supplying anchoring (40). Integrins are also responsible for transmitting chemical signals into the cells, where the signals develop in the ECM after ligation and involve receptor clustering and binding of a particular ligand (41). As a response to this clustering and the presence of GTPase Rho A, cytoskeletal proteins like focal adhesion kinase (FAK) are formed. The Ras protein, which plays a crucial role in cell signaling and gene expression, is phosphorylated by FAK to activate the mitogen-activated protein (MAP) kinase pathway (42). FAK plays a role in co-localizing with integrin receptors in adherent cell types at cell-substratum contact points known as focal adhesions (43). FAK stimulates cell motility, survival, and proliferation through kinase-dependent and -independent processes during the development of various malignancies (44). Studies have reported that FAK signaling is located at the junction of other signaling pathways promoting metastasis (45). According to various reports, FAK signaling is linked to the maintenance of cancer stem cells (46). It has been highlighted that the tumor cells of Osteosarcoma interact with their microenvironment, where β4 integrin plays a significant role in metastasis and the invasive nature of cancer (47). Growth factors and integrin ligands work synergistically to regulate the differentiation of osteogenic cells from stem cells (48). Growth factors called Bone Morphogenic Proteins (BMPs) substantially impact the development and remodeling of postnatal skeletal tissue, among other things (49). There are 14 known BMPs, collectively constituting a subfamily with Growth Differentiation Factors (GDFs). Among the 14 known BMPs, BMP-2, BMP-4, BMP-6, BMP-7, and BMP-9 are especially important as they have been found to induce complete bone morphogenesis (50). It has been reported that the inhibition of β4 integrin has gradually mitigated and inhibited metastasis in patients with Osteosarcoma (51). Thus, analyzing the network targeting genes involved in the integrin family of cell surface interactions can help develop therapeutic targets for the disease.

Our study has also revealed various highly enriched pathways, such as the FOXM1 transcription factor network, ATM pathway, signaling event mediated by FAK, Arf6 signaling events, and Class 1 PI3K signaling events. FOXM1, a Forkhead Box Transcription Factor, is known for maintaining the homoeostatic environment and other cellular functions, such as cell proliferation, cell cycle progression, DNA damage repair, angiogenesis, etc. Being associated with a large number of cellular processes, it has also manifested its role in several diseases as well as cancer. It has been studied that FOXM1 plays a role in tumor growth and progression (52). Forkhead box (Fox) proteins belong to a superfamily of evolutionarily conserved transcriptional factors characterized by a common DNA binding domain known as the forkhead box or winged helix domain (53). FOXM1 preferentially binds promoter regions with the consensus "TAAACA" recognition sequence (54). Cell cycle regulation regulates its expression at mRNA and protein levels. It increases during the S-phase, peaks G2 and M, and degrades during mitotic exit (55). Genetic alteration and gene copy amplification of FOXM1 has been seen at loci 12p13.33, exhibiting oncogenic properties (56). Various studies have highlighted that the alterations arise in FOXM1 during post-transcriptional and post-translational modifications, which leads to its deregulation and overexpression in cancer cells (55, 57). The role of FOXM1 in tumor cells is its participation in the self–renewal and proliferation of cancer stem cells through Wnt signaling, the MAPK-ERK pathway, and the PI3K-mTOR pathway (58). Studies conducted on patients suffering from Osteosarcoma have revealed that the upregulation of miR-370 suppressed the expression of FOXMI. On the contrary, it was also evident that miR-370 was reduced in Osteosarcoma cells where FOXM1 expression was elevated. The miR-370 is a class of micro-RNA involved in various cellular processes such as proliferation, differentiation, apoptosis, and tumor suppression (59). Thus, micro-RNA can serve as a potential drug target in controlling the spread of Osteosarcoma by FOXM1 factor since the contribution of this transcription factor in promoting the disease is exemplary. The Ataxia-Telangiectasia Mutated (ATM) kinase is an essential sensor and signal transducer in the DNA damage response. It is noteworthy that ATM is often considered a major tumor suppressor because of its ability to induce cell cycle arrest. However, certain tumor cells in the advanced stages exhibit enhanced ATM signaling, which benefits cancer cell survival, resistance to radiation and chemotherapy, biosynthesis, proliferation, and metastasis (60). ATM is an active serine/threonine kinase and is an important member of the P13K-related protein kinase family (PIKK). The two main types of ATM signaling are the

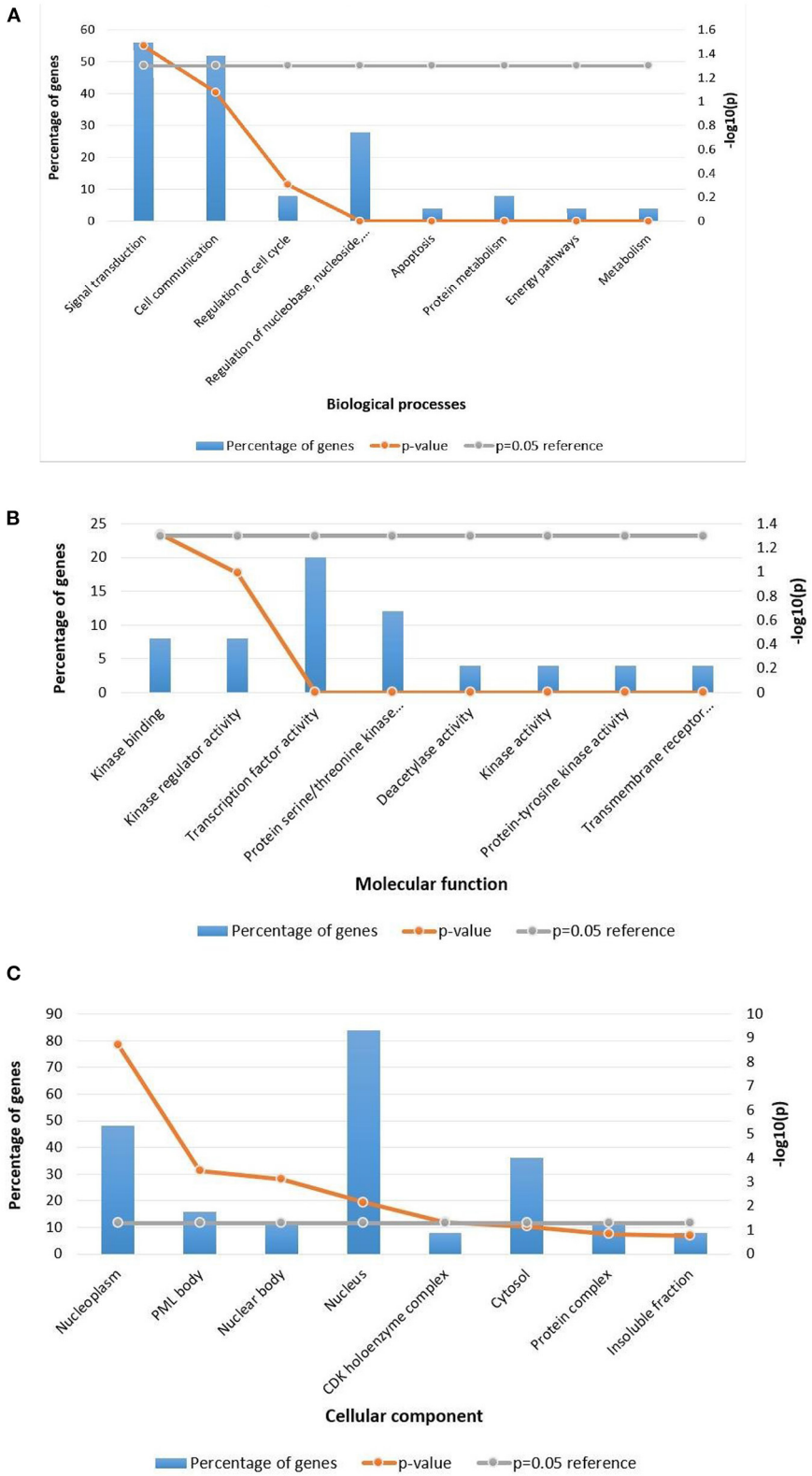


FIGURE 3
Functional enrichment analysis of Osteosarcoma gene. Significantly enriched **(A)** Biological processes, **(B)** Molecular function, **(C)** Cellular component. The *p*-value is taken as 0.05 for reference.

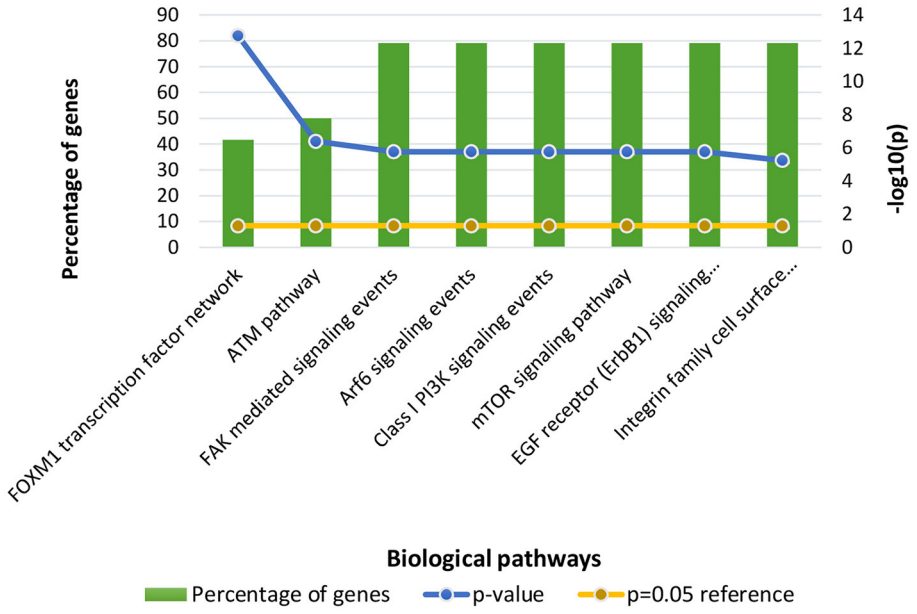


FIGURE 4
Functional enrichment of Osteosarcoma genes involved in significant biological pathways. The *p*-value is considered 0.05 for reference.

TABLE 3 Clustering analysis of Osteosarcoma gene interaction network.

Cluster	MCODE score	No. of nodes	No. of edges	Node IDs
C1	11.913	24	137	*VEGFA, MYCN, SRC, AURKA, MCM5, KIT, ERBB2, CCND1, CHEK2, E2F2, TIMP1, CCNG1, CCNE1, CD44, STAT3, PLK1, PRIM1, FOS, CDC6, SIRT1, MMP2, CDKN2D, CCNB2,* and *MCM4*
C2	5.00	5	10	*TNFSF11, BMP2, SPARC, MMP13,* and *IBSP*
C3	4.677	7	14	*TP53, RB1, S100A4, RECQL4, CDK4, MDM2,* and *CDKN2A*

canonical route, which is activated by DNA damage and signals with the Mre11-Rad50-NBS1 (MRN) complex, and many non-canonical modes of activation triggered by other types of cellular stress. Both types of signaling are likely to play a part in ATM's ability to limit tumor growth (61). PI3K family members such as ATM are routinely auto-inhibited when they are in their resting state (dimers or polymers), and are only activated when they attach to their partners. The ATM canonical pathway is activated upon DNA double-strand breaks (DSBs), where ATM dimers are dissociated to monomers, activation is triggered, and ATM monomers are recruited to the DNA damage sites (62, 63). Since ATMs induce cell cycle arrest and apoptosis whenever genetic alteration occurs, cancer cells use various mechanisms to downregulate ATMs. For instance, ATM expression can be decreased in some cancers due to miRNA-18a (64). Arf6 is a member of the adenosine diphosphate (ADP)-ribosylation factor (ARF) family of small GTPases. By regulating the transit of proteins and lipids in eukaryotic cells, ARFs influence cellular behavior and function (65). Arf6 controls cytoskeletal remodeling, cell shape alterations, extracellular matrix proteolysis, and cell adhesion mechanisms involved in tumor cell migration (66). Degradation of the ECM by matrix metalloproteinases (MMPs) is required for tumor cell invasion. MMPs are released into the extracellular environment by both invadopodia and tumor cell-expelled microvesicles, aiding the breakdown of the ECM and invasion (67). Initiation of Arf6 leads to the activation of Rho and Rac1 pathways, which promotes both microvesicle shedding and formation of invadopodia, whereas expression of a dominant negative Arf6 prevents the development of invadopodia and microvesicle shedding (68, 69). The phosphoinositide 3-kinase (PI3K) family is crucial to almost every aspect of cell and tissue biology and hyperactivation of PI3K is one of the central events in cancer (70). Studies carried out in the early 2000s were the first to show that class I PI3K catalytic isoforms had the ability to alter themselves. Since the discovery of its mutated form, PI3KCA, PI3K has been placed on the frontline as a big player in understanding cancer. The enrichment analysis of Osteosarcoma genes has also revealed clinical phenotypes and sites of gene expression where the former consisted of neoplasia, somatic mutation, osteogenic sarcoma, and painful tender mass at long bone metaphysis (Figure 5) while the latter comprised of the esophagus, oral mucosa, malignant glioma, endometrium, uterine cervix, and vulva (Figure 6).

To identify possible drug targets for Osteosarcoma, which plays an essential role in various biological pathways, we used NetworkAnalyzer, which is a built plugin in Cytoscape.

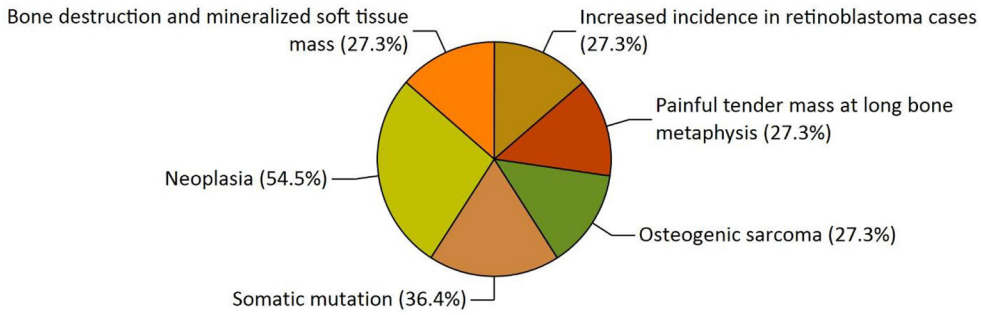


FIGURE 5
The functional enrichment analysis of osteosarcoma genes performed by the FunRich tool revealed clinical phenotypes.

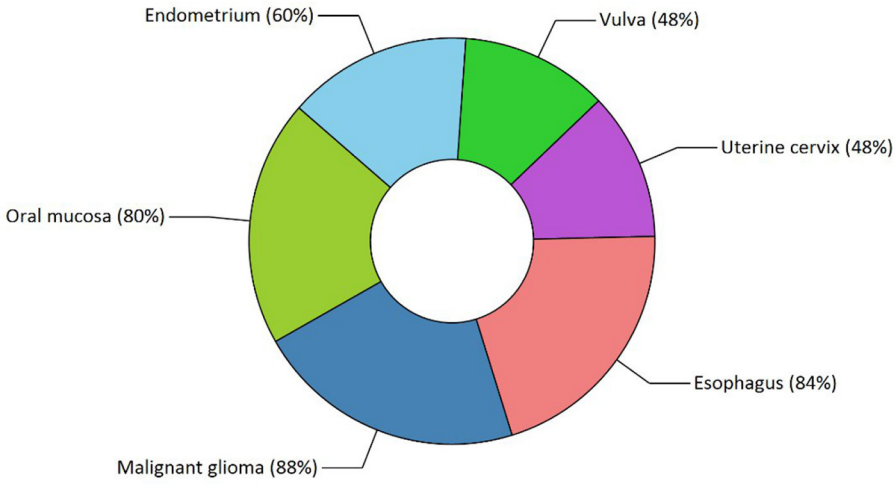


FIGURE 6
Functional enrichment analysis revealed sites of expression of osteosarcoma genes. The various highlighted portions show different sites of expression.

NetworkAnalyzer uses various parameters such as degree, average shortest path length, closeness centrality, and betweenness centrality. Degree refers to the no. of direct interactors, and more no. of degrees will indicate more no. of gene interactors which will help us to understand the progression of a pathway. The significance of the gene in gene-to-gene communication increases with decreasing average shortest path length and increasing closeness centrality. Based on the parameters mentioned above, our study has revealed the top five genes viz. *TP53*, *CCND1*, *CDK4*, *STAT3*, and *VEGFA*, can be considered potential biomarkers because they are involved in the major biological pathways of Osteosarcoma.

TP53 gene is a potential biomarker with the most no. of direct interactors of 45 with the shortest average path length of 1.343 and the highest closeness centrality of 0.744. *TP53* is seen to be involved in various biological pathways. In our study, such as the FOXM1 transcription factor network, ATM pathway, Class 1 PI3K signaling events, and mTOR signaling pathway. In normal conditions, *TP53* is a tumor suppressor gene that initiates numerous stress-induced pathways, including DNA damage, senescence, cellular death, and reprogramming. It stimulates numerous genes encoding proteins responsible for apoptosis (71). In a cancerous state, *TP53* is mutated, which loses its ability to suppress the tumor, thereby promoting uncontrolled cell proliferation. Over 50% of human neoplasms have somatic mutations in the *TP53* gene. About 10% of the changes are nonsense mutations, resulting in shortened p53 proteins, while most variants are missense mutations. Sixty percent of neoplasms with missense *TP53* mutations have their second *TP53* allele deleted (72). Earlier studies demonstrated that FOXM1 expression is increased when p53 is partially deleted or inactivated by negatively regulating the expression of FOXM1. Similar studies on *TP53* have revealed that reverse regulation of *TP53* through the mTOR pathway also modifies the synchronization of growth signals and stressors (73). The *TP53* gene is considered a hallmark in cancer studies and serves maximum potential for developing therapeutic targets for treating Osteosarcoma.

CCND1 gene can serve as a drug target with 33 direct interactors having a path length of 1.582 and closeness centrality of

0.632. *CCND1* or *Cyclin D1* gene synthesizes a protein that governs cyclin-dependent kinases in the cell cycle. It is well recognized as a regulator of cell cycle progression in the nucleus, modifying the transition from the G1 to the S phase. Although Cyclin D1 is well recognized for its function in the nucleus, current clinical investigations link it to tumor invasion and metastasis when it is present in the cytoplasmic membrane (74). It is altered in 4.10% of all cancers, typically by post-transcriptional regulation, translocation, or amplification (75).

Additionally, emerging evidence reveals that *CCND1* gene mutations that cause nuclear retention and constitutive CDK4/6 kinase activation are oncogenic (76). *CCND1* is also seen to be involved in biological pathways such as the FOMX1 transcription pathway, mTOR signaling pathway, etc. *CCND1* is seen to be involved in response to leptin, which is a peptide hormone produced by adipocytes. Leptin helps in the maintenance of normal cellular homeostasis. Downregulation of the apoptotic reaction and upregulation of the cell cycle is due to the pro-carcinogenic impact of leptin (77). Therefore, targeting the *CCND1* gene may aid in halting Osteosarcoma development.

CDK4 gene plays a significant role in the completion of the cell cycle and are often hyperactive in cancer. CDKs are serine/threonine kinases that are activated in association with a cyclin partner. It has a no. of direct interactors of 28 with an average shortest path length of 1.626 and closeness centrality of 0.614. During the G1-S transition, retinoblastoma protein acts as a negative cell cycle regulator by binding to the transcription factor E2F and suppressing transcriptional activity during the early G1 phase. D-type cyclins express themselves more often in response to mitogenic stimuli, and they join forces with CDK4/6 to phosphorylate RB. The E2F transcription factor family's inhibitory control on RB is partially relieved by hypo phosphorylated RB, which encourages the expression of E2F target genes like cyclin E and speeds up the G1 phase transition (78). Studies have also shown that *CDK4* is involved in the regulation of the mTOR pathway activated, thus making it a potential drug target (79).

The *VEGFA* gene is considered a hallmark in cancer-related studies because of its role in angiogenesis, accomplished periodically from pre-existing vascular networks (80). The tumor angiogenesis is achieved in four steps. First is disruption of the basement membrane leading to hypoxia. Second is the dispersion of endothelial cells activated by VEGFA, followed by the proliferation and stabilization of endothelial cells. At last, the angiogenesis regulating factors regulates the repeated process of angiogenesis (81). Studies have also demonstrated that *the* FOXM1 transcription factor regulates *VEGFA* to promote tumor angiogenesis (82). *VEGFA* gene had a degree value of 27 and an average shortest path length of 1.701.

The signal transducer and activator of transcription, *STAT3*, plays a vital role in DNA replication. Being an essential STAT protein family member, it plays a crucial part in various vital cellular functions, including proliferation, differentiation, survival, immunosuppression, angiogenesis, and cancer (83). *STAT3*-activated genes enhance angiogenesis and metastasis, prevent apoptosis, promote cell proliferation and survival, and suppress antitumor immune responses (84). In addition to its established role as a transcription factor in cancer, *STAT3* regulates mitochondrion functions (85). *STAT* gene has been revealed to have direct interactors of 27 with an average shortest path length of 1.716 and closeness centrality of 0.587.

5. Conclusion

Osteosarcoma is one of the most frequently occurring sarcomas with a high potency of tumorigenesis. Although chemotherapy and radiotherapy are available as treatment options that have improved patients' lives, there is still some gray area regarding the etiology of the disease. To develop better therapeutics, a specific approach toward the disease targeting the hub genes involved in various signaling pathways must be opted to unravel the complexity of the disease. Our study has mentioned hub genes viz. *TP53, CCND1 CDK4, STAT3,* and *VEGFA* have the highest no. of interactions and showed a high clustering density. Their involvement in the major signaling pathways of Osteosarcoma makes them potential candidates to be targeted for drug development. The highly enriched signaling pathways include the FOXM1 transcription pathway, ATM signaling pathway, FAK mediated signaling events, Arf6 signaling events, Class 1 PI3K signaling events, mTOR signaling pathway, and Integrin family cell surface interactions. Targeting the hub genes and their associated functional partners, which we have reported in our studies, may be efficacious in developing novel therapeutic targets.

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.

Author contributions

Conceptualization: HD, KV, GD, and AE. Methodology: HD and KV. Formal analysis: HD, SK, and AA. Investigation: HD, KV, GD, and HZ. Data curation: HD and SK. Writing—original draft preparation: HD, KV, and AE. Writing—review and editing: KV, AE, GD, and HZ. Supervision: KV and AE. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

The authors express deep gratitude to the management of REVA University, Bengaluru, Karnataka, India, and Vellore Institute of Technology, Vellore, Tamil Nadu, India, for all the support, assistance, and constant encouragement to carry out this work.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1154417/full#supplementary-material

References

1. Ward E, DeSantis C, Robbins A, Kohler B, Jemal A. Childhood and adolescent cancer statistics. *CA Cancer J Clin.* (2014) 64:83–103. doi: 10.3322/caac.21219

2. Ismiarto YD, Sitanggang GL, Kamal AFK, Prabowo et al. Orthopedic oncology completed. *Indian J Orthop.* (2018) 10:1–7. doi: 10.1007/978-3-319-07323-1

3. Mckenna RJ, Schwinn CP, Soong KY, Higinbotham NL. Sarcomata of the osteogenic series (osteosarcoma, fibrosarcoma, chondrosarcoma, parosteal osteogenic sarcoma, and sarcomata arising in abnormal bone): an analysis of 552 cases. *JBJS.* (1966) 48:1–26. doi: 10.2106/00004623-196648010-00001

4. Dahlin DC, Coventry MB. Osteogenic sarcoma. A study of six hundred cases. *J Bone Joint Surg Am.* (1967) 49:101–10. doi: 10.2106/00004623-196749010-00008

5. Deyrup AT, Montag AG, Inwards CY, Xu Z, Swee RG, Krishnan Unni K, et al. Sarcomas arising in Paget disease of bone: a clinicopathologic analysis of 70 cases. *Arch Pathol Lab Med.* (2007) 131:942–6. doi: 10.5858/2007-131-942-SAIPDO

6. Reimann E, Kõks S, Ho XD, Maasalu K, Märtson A. Whole exome sequencing of a single osteosarcoma case–integrative analysis with whole transcriptome RNA-seq data. *Hum Genomics.* (2014) 8:20. doi: 10.1186/PREACCEPT-1873296159134645

7. Chen X, Bahrami A, Pappo A, Easton J, Dalton J, Hedlund E, et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. *Cell Rep.* (2014) 7:104–12. doi: 10.1016/j.celrep.2014.03.003

8. Bougeard G, Renaux-Petel M, Flaman JM, Charbonnier C, Fermey P, Belotti M, et al. Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers. *J Clin Oncol Off J Am Soc Clin Oncol.* (2015) 33:2345–52. doi: 10.1200/JCO.2014.59.5728

9. Kansara M, Thomas DM. Molecular pathogenesis of osteosarcoma. *DNA Cell Biol.* (2007) 26:1–18. doi: 10.1089/dna.2006.0505

10. Wang LL. Biology of osteogenic sarcoma. *Cancer J.* (2005) 11:294–305. doi: 10.1097/00130404-200507000-00005

11. Isakoff MS, Bielack SS, Meltzer P, Gorlick R. Osteosarcoma: current treatment and a collaborative pathway to success. *J Clin Oncol.* (2015) 33:3029–35. doi: 10.1200/JCO.2014.59.4895

12. Janeway KA, Grier HE. Sequelae of osteosarcoma medical therapy: a review of rare acute toxicities and late effects. *Lancet Oncol.* (2010) 11:670–8. doi: 10.1016/S1470-2045(10)70062-0

13. Yan W, Xue W, Chen J, Hu G. Biological networks for cancer candidate biomarkers discovery. *Cancer Inform.* (2016) 15:1–7. doi: 10.4137/CIN.S39458

14. Liu R, Cheng Y, Yu J, Lv Q-L, Zhou H-H. Identification and validation of gene module associated with lung cancer through coexpression network analysis. *Gene.* (2015) 563:56–62. doi: 10.1016/j.gene.2015.03.008

15. Liu R, Guo C-X, Zhou H-H. Network-based approach to identify prognostic biomarkers for estrogen receptor-positive breast cancer treatment with tamoxifen. *Cancer Biol Ther.* (2015) 16:317–24. doi: 10.1080/15384047.2014.1002360

16. Liu R, Lv QL, Yu J, Hu L, Zhang LH, Cheng Y, et al. Correlating transcriptional networks with pathological complete response following neoadjuvant chemotherapy for breast cancer. *Breast Cancer Res Treat.* (2015) 151:607–18. doi: 10.1007/s10549-015-3428-x

17. Udhaya Kumar S, Saleem A, Thirumal Kumar D, Anu Preethi V, Younes S, Zayed H, et al. A systemic approach to explore the mechanisms of drug resistance and altered signaling cascades in extensively drug-resistant tuberculosis. *Adv Protein Chem Struct Biol.* (2021) 127:343–64. doi: 10.1016/bs.apcsb.2021.02.002

18. Mishra S, Shah MI, Udhaya Kumar S, Thirumal Kumar D, Gopalakrishnan C, Al-Subaie AM, et al. Network analysis of transcriptomics data for the prediction and prioritization of membrane-associated biomarkers for idiopathic pulmonary fibrosis (IPF) by bioinformatics approach. *Adv Protein Chem Struct Biol.* (2021) 123:241–73. doi: 10.1016/bs.apcsb.2020.10.003

19. Miryala SK, Anbarasu A, Ramaiah S. Discerning molecular interactions: a comprehensive review on biomolecular interaction databases and network analysis tools. *Gene.* (2018) 642:84–94. doi: 10.1016/j.gene.2017.11.028

20. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* (2015) 43:D447–52. doi: 10.1093/nar/gku1003

21. Assenov Y, Ramírez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological parameters of biological networks. *Bioinformatics.* (2008). 24:282–4. doi: 10.1093/bioinformatics/btm554

22. Saito R, Smoot ME, Ono K, Ruscheinski J, Wang PL, Lotia S, et al. A travel guide to Cytoscape plugins. *Nat Methods.* (2012) 9:1069–76. doi: 10.1038/nmeth.2212

23. Pathan M, Keerthikumar S, Ang CS, Gangoda L, Quek CY, Williamson NA, et al. FunRich: an open access standalone functional enrichment and interaction network analysis tool. *Proteomics.* (2015) 15:2597–601. doi: 10.1002/pmic.201400515

24. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell.* (2011) 144:646–74. doi: 10.1016/j.cell.2011.02.013

25. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell.* (2000) 100:57–70. doi: 10.1016/S0092-8674(00)81683-9

26. Tian Z, Niu X, Yao W. Receptor tyrosine kinases in osteosarcoma treatment: which is the key target? *Front Oncol.* (2020) 10:1642. doi: 10.3389/fonc.2020.01642

27. de Azevedo JW, Fernandes TA, Fernandes JV, de Azevedo JC, Lanza DC, Bezerra CM, et al. Biology and pathogenesis of human osteosarcoma. *Oncol Lett.* (2020) 19:1099–116. doi: 10.3892/ol.2019.11229

28. Sever R, Brugge JS. Genetic and epigenetic mechanisms of cancer progression. *Cold Spring Harb Perspect Med.* (2015) 5a:00609821.

29. Adamopoulos C, Gargalionis AN, Basdra EK, Papavassiliou AG. Deciphering signaling networks in osteosarcoma pathobiology. *Exp Biol Med.* (2016) 241:1296–305. doi: 10.1177/1535370216648806

30. Zou Z, Tao T, Li H, Zhu X. MTOR signaling pathway and mTOR inhibitors in cancer: Progress and challenges. *Cell Biosci.* (2020) 10:1–11. doi: 10.1186/s13578-020-00396-1

31. Unni N, Arteaga CL. Is Dual mTORC1 and mTORC2 therapeutic blockade clinically feasible in cancer? *JAMA Oncol.* (2019) 5:1564–5. doi: 10.1001/jamaoncol.2019.2525

32. Dowling RJO, Topisirovic I, Fonseca BD, Sonenberg N. Dissecting the role of mTOR: lessons from mTOR inhibitors. *Biochim Biophys Acta.* (2010) 1804:433–9. doi: 10.1016/j.bbapap.2009.12.001

33. Lener MS. 乳鼠心肌提取HHS public access. *Physiol Behav.* (2016) 176:139–48. doi: 10.1016/j.cell.2017.02.004

34. Mossmann D, Park S, Hall MN. mTOR signaling and cellular metabolism are mutual determinants in cancer. *Nat Rev Cancer.* (2018) 18:744–57. doi: 10.1038/s41568-018-0074-8

35. Ding L, Congwei L, Bei Q, Tao Y, Ruiguo W, Heze Y, et al. (2016). mTOR: an attractive therapeutic target for osteosarcoma? *Oncotarget* 7:50805–13. doi: 10.18632/oncotarget.9305

36. Porta C, Paglino C, Mosca A. Targeting PI3K/Akt/mTOR signaling in cancer. *Front Oncol.* (2014) 4:64. doi: 10.3389/fonc.2014.00064

37. Bergfeld SA, DeClerck YA. Bone marrow-derived mesenchymal stem cells and the tumor microenvironment. *Cancer Metastasis Rev.* (2010) 29:249–61. doi: 10.1007/s10555-010-9222-7

38. Walker C, Mojares E, Del Río Hernández A. Role of extracellular matrix in development and cancer progression. *Int J Mol Sci.* (2018) 19. doi: 10.3390/ijms19103028

39. Dominiak A, Chełstowska B, Olejarz W, Nowicka G. Communication in the cancer microenvironment as a target for therapeutic interventions. *Cancers.* (2020) 12. doi: 10.3390/cancers12051232

40. Hynes RO. Integrins: bidirectional, allosteric signaling machines. *Cell.* (2002) 110:673–87. doi: 10.1016/S0092-8674(02)00971-6

41. Jones JL, Walker RA. Integrins: a role as cell signaling molecules. *Mol Pathol.* (1999) 52:208–13. doi: 10.1136/mp.52.4.208

42. Harburger DS, Calderwood DA. Calderwood. Integrin signaling at a glance. *J Cell Sci.* (2009) 122(Pt 2):159–63. doi: 10.1242/jcs.018093

43. Schlaepfer DD, Hauck CR, Sieg DJ. Signaling through focal adhesion kinase. *Prog Biophys Mol Biol.* (1999) 71:435–78. doi: 10.1016/S0079-6107(98)00052-2

44. Zhao J, Guan J-L. Signal transduction by focal adhesion kinase in cancer. *Cancer Metastasis Rev.* (2009) 28:35–49. doi: 10.1007/s10555-008-9165-4

45. Sulzmaier FJ, Jean C, Schlaepfer DD. FAK in cancer: mechanistic findings and clinical applications. *Nat Rev Cancer.* (2014) 14:598–610. doi: 10.1038/nrc3792

46. Barbero S, Mielgo A, Torres V, Teitz T, Shields DJ, Mikolon D, et al. Caspase-8 association with the focal adhesion complex promotes tumor cell migration and metastasis. *Cancer Res.* (2009) 69:3755–63. doi: 10.1158/0008-5472.CAN-08-3937

47. Tawil B. The importance of cell signaling—integrins and growth factors—in bone tissue engineering: applications for the treatment of osteosarcoma. *Adv Tissue Eng Regen Med Open Access.* (2017) 2:e21. doi: 10.15406/atroa.2017.02.00021

48. Wei Q, Pohl TLM, Seckinger A, Spatz JP, Cavalcanti-Adam EA. Regulation of integrin and growth factor signaling in biomaterials for osteodifferentiation. *Beilstein J Org Chem.* (2015) 11:773–83. doi: 10.3762/bjoc.11.87

49. Chen D, Zhao M, Mundy GR. Bone morphogenetic proteins. *Growth Factors.* (2004) 22:233–41. doi: 10.1080/08977190412331279890

50. Miyazono K, Kamiya Y, Morikawa M. Bone morphogenetic protein receptors and signal transduction. *J Biochem.* (2010) 147:35–51. doi: 10.1093/jb/mvp148

51. Wan X, Kim SY, Guenther LM, Mendoza A, Briggs J, Yeung C, et al. Beta4 integrin promotes osteosarcoma metastasis and interacts with ezrin. *Oncogene.* (2009) 28:3401–11. doi: 10.1038/onc.2009.206

52. Kalathil D, John S, Nair AS. FOXM1 and cancer: faulty cellular signaling derails homeostasis. *Front Oncol.* (2021) 10:626836. doi: 10.3389/fonc.2020.626836

53. Alvarez-Fernández M, Medema RH. Novel functions of FoxM1: from molecular mechanisms to cancer therapy. *Front Oncol.* (2013) 3:1–5. doi: 10.3389/fonc.2013.00030

54. Littler DR, Alvarez-Fernández M, Stein A, Hibbert RG, Heidebrecht T, Aloy P, et al. Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. *Nucleic Acids Res.* (2010) 38:4527–38. doi: 10.1093/nar/gkq194

55. Park HJ, Wang Z, Costa RH, Tyner A, Lau LF, Raychaudhuri P, et al. An N-terminal inhibitory domain modulates activity of FoxM1 during cell cycle. *Oncogene.* (2008) 27:1696–704. doi: 10.1038/sj.onc.1210814

56. Barger CJ, Branick C, Chee L, Karpf AR. Pan-cancer analyses reveal genomic features of FOXM1 overexpression in cancer. *Cancers (Basel).* (2019) 11:251. doi: 10.3390/cancers11020251

57. Li Y, Guo H, Jin C, Qiu C, Gao M, Zhang L, et al. Spliceosome-associated factor CTNNBL1 promotes proliferation and invasion in ovarian cancer. *Exp Cell Res.* (2017) 357:124–34. doi: 10.1016/j.yexcr.2017.05.008

58. Sher G, Masoodi T, Patil K, Akhtar S, Kuttikrishnan S, Ahmad A, et al. (2022). Dysregulated FOXM1 signaling in the regulation of cancer stem cells. *Semin. Cancer Biol.* 86:107–21. doi: 10.1016/j.semcancer.2022.07.009

59. Duan N, Hu X, Yang X, Cheng H, Zhang W. MicroRNA-370 directly targets FOXM1 to inhibit cell growth and metastasis in osteosarcoma cells. *Int J Clin Exp Pathol.* (2015) 8:10250–60.

60. Phan LM, Rezaeian A-H. ATM: main features, signaling pathways, and its diverse roles in dna damage response, tumor suppression, and cancer development. *Genes.* (2021) 12:845. doi: 10.3390/genes12060845

61. Cremona CA, Behrens A. ATM signaling and cancer. *Br Dent J.* (2014) 217:3351–60. doi: 10.1038/onc.2013.275

62. Uziel T, Lerenthal Y, Moyal L, Andegeko Y, Mittelman L, Shiloh Y, et al. Requirement of the MRN complex for ATM activation by DNA damage. *EMBO J.* (2003) 22:5612–21. doi: 10.1093/emboj/cdg541

63. Lee J-H, Paull TT. ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science.* (2005) 308:551–4. doi: 10.1126/science.1108297

64. Song L, Lin C, Wu Z, Gong H, Zeng Y, Wu J, et al. miR-18a impairs DNA damage response through downregulation of ataxia telangiectasia mutated (ATM) kinase. *PLoS ONE.* (2011) 6:e25454–e25454. doi: 10.1371/journal.pone.0025454

65. Kahn RA, Cherfils J, Elias M, Lovering RC, Munro S, Schurmann A, et al. Nomenclature for the human Arf family of GTP-binding proteins: ARF, ARL, and SAR proteins. *J Cell Biol.* (2006) 172:645–50. doi: 10.1083/jcb.200512057

66. Schweitzer JK, Sedgwick AE, D'Souza-Schorey C. ARF6-mediated endocytic recycling impacts cell movement, cell division and lipid homeostasis. *Semin Cell Dev Biol.* (2011) 22:39–47. doi: 10.1016/j.semcdb.2010.09.002

67. Sedgwick AE, Clancy JW, Olivia Balmert M, D'Souza-Schorey C. Extracellular microvesicles and invadopodia mediate non-overlapping modes of tumor cell invasion. *Sci Rep.* (2015) 5:14748. doi: 10.1038/srep14748

68. Muralidharan-Chari V, Clancy J, Plou C, Romao M, Chavrier P, Raposo G, et al. ARF6-regulated shedding of tumor cell-derived plasma membrane microvesicles. *Curr Biol.* (2009) 19:1875–85. doi: 10.1016/j.cub.2009.09.059

69. Muralidharan-Chari V, Hoover H, Clancy J, Schweitzer J, Suckow MA, Schroeder V, et al. ADP-ribosylation factor 6 regulates tumorigenic and invasive properties in vivo. *Cancer Res.* (2009) 69:2201–9. doi: 10.1158/0008-5472.CAN-08-1301

70. Thorpe LM, Yuzugullu H, Zhao JJ. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat Rev Cancer.* (2015) 15:7–24. doi: 10.1038/nrc3860

71. Synoradzki KJ, Bartnik E, Czarnecka AM, Fiedorowicz et al. Tp53 in biology and treatment of osteosarcoma. *Cancers.* (2021) 13:1–23. doi: 10.3390/cancers13174284

72. Willis A, Jung EJ, Wakefield T, Chen X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene.* (2004) 23:2330–8. doi: 10.1038/sj.onc.1207396

73. Cui D, Qu R, Liu D, Xiong X, Liang T, Zhao Y, et al. The cross talk between p53 and mTOR pathways in response to physiological and genotoxic stresses. *Front Cell Dev Biol.* (2021) 9:775507. doi: 10.3389/fcell.2021.775507

74. Montalto FI, De Amicis F. Cyclin D1 in cancer: a molecular connection for cell cycle control, adhesion and invasion in tumor and stroma. *Cells.* (2020) 9:2648. doi: 10.3390/cells9122648

75. Moreno-Bueno G, Rodríguez-Perales S, Sánchez-Estévez C, Hardisson D, Sarrió D, Prat J, et al. Cyclin D1 gene (CCND1) mutations in endometrial cancer. *Oncogene.* (2003) 22:6115–8. doi: 10.1038/sj.onc.1206868

76. Alt JR, Cleveland JL, Hannink M, Diehl JA. Phosphorylation-dependent regulation of cyclin D1 nuclear export and cyclin D1-dependent cellular transformation. *Genes Dev.* (2000) 14:3102–14. doi: 10.1101/gad.854900

77. Samad N. Role of leptin in cancer: a systematic review. *Biomed J Sci Tech Res.* (2019) 18:13226–35. doi: 10.26717/BJSTR.2019.18.003091

78. Yang Y, Luo J, Chen X, Yang Z, Mei X, Ma J, et al. CDK4/6 inhibitors: a novel strategy for tumor radiosensitization. *J Exp Clin Cancer Res.* (2020) 39:188. doi: 10.1186/s13046-020-01693-w

79. Romero-Pozuelo J, Figlia G, Kaya O, Martin-Villalba A, Teleman AA. Cdk4 and Cdk6 couple the cell-cycle machinery to cell growth via mTORC1. *Cell Rep.* (2020) 31:107504. doi: 10.1016/j.celrep.2020.03.068

80. Sa-nguanraksa D, O-charoenrat P. The role of vascular endothelial growth factor A polymorphisms in breast cancer. *Int J Mol Sci.* (2012) 13:14845–64. doi: 10.3390/ijms131114845

81. Pagès G, Pouysségur J. Transcriptional regulation of the vascular endothelial growth factor gene—a concert of activating factors. *Cardiovasc Res.* (2005) 65:564–73. doi: 10.1016/j.cardiores.2004.09.032

82. Wang R-T, Miao R-C, Zhang X, Yang G-H, Mu Y-P, Zhang Z-Y, et al. Fork head box M1 regulates vascular endothelial growth factor-A expression to promote the angiogenesis and tumor cell growth of gallbladder cancer. *World J Gastroenterol.* (2021) 27:692–707. doi: 10.3748/wjg.v27.i8.692

83. Yu H, Lee H, Herrmann A, Buettner R, Jove R. Revisiting STAT3 signaling in cancer: new and unexpected biological functions. *Nat Rev Cancer.* (2014) 14:736–46. doi: 10.1038/nrc3818

84. Frank DA. STAT3 as a central mediator of neoplastic cellular transformation. *Cancer Lett.* (2007) 251:199–210. doi: 10.1016/j.canlet.2006.10.017

85. Wegrzyn J, Potla R, Chwae Y-J, Sepuri NBV, Zhang Q, Koeck T, et al. Function of mitochondrial Stat3 in cellular respiration. *Science.* (2009) 323:793–7. doi: 10.1126/science.1164551

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 17 April 2023
DOI 10.3389/fmed.2023.1150933

OPEN ACCESS

EDITED BY
Balu Kamaraj,
Imam Abdulrahman Bin Faisal University,
Saudi Arabia

REVIEWED BY
Udhaya Kumar S.,
Baylor College of Medicine,
United States
Karthick Vasudevan,
Reva University,
India

*CORRESPONDENCE
Rohini Karunakaran
✉ rohini@aimst.edu.my

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

RECEIVED 25 January 2023
ACCEPTED 09 March 2023
PUBLISHED 17 April 2023

CITATION
Subramani S, Varshney N, Anand MV, Soudagar MM, Al-keridis LA, Upadhyay TK, Alshammari N, Saeed M, Subramanian K, Anbarasu K and Rohini K (2023) Cardiovascular diseases prediction by machine learning incorporation with deep learning.
Front. Med. 10:1150933.
doi: 10.3389/fmed.2023.1150933

Cardiovascular diseases prediction by machine learning incorporation with deep learning

Sivakannan Subramani[1], Neeraj Varshney[2], M. Vijay Anand[3], Manzoore Elahi M. Soudagar[4], Lamya Ahmed Al-keridis[5], Tarun Kumar Upadhyay[6], Nawaf Alshammari[7], Mohd Saeed[7], Kumaran Subramanian[8], Krishnan Anbarasu[9] and Karunakaran Rohini[10,11,12]*

[1]Department of Advanced Computing, St. Joseph's University, Bengaluru, Karnataka, India, [2]Department of Computer Engineering and Applications, GLA University, Mathura, Uttar Pradesh, India, [3]Department of Mechanical Engineering, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India, [4]Department of VLSI Microelectronics, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India, [5]Faculty of Science, Princess Norah Bint Abdulrahman University, Riyadh, Saudi Arabia, [6]Department of Biotechnology, Parul Institute of Applied Sciences and Centre of Research for Development, Parul University, Vadodara, India, [7]Department of Biology, College of Science, University of Hail, Hail, Saudi Arabia, [8]Centre for Drug Discovery and Development, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India, [9]Department of Bioinformatics, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India, [10]Unit of Biochemistry, Centre of Excellence for Biomaterials Engeneering, Faculty of Medicine, AIMST University, Semeleing, Bedong, Malaysia, [11]Centre for Excellence for Biomaterials Science AIMST University, Semeling, Bedong, Malaysia, [12]Department of Computational Biology, Saveetha School of Engineering, SIMATS, Chennai, Tamil Nadu, India

It is yet unknown what causes cardiovascular disease (CVD), but we do know that it is associated with a high risk of death, as well as severe morbidity and disability. There is an urgent need for AI-based technologies that are able to promptly and reliably predict the future outcomes of individuals who have cardiovascular disease. The Internet of Things (IoT) is serving as a driving force behind the development of CVD prediction. In order to analyse and make predictions based on the data that IoT devices receive, machine learning (ML) is used. Traditional machine learning algorithms are unable to take differences in the data into account and have a low level of accuracy in their model predictions. This research presents a collection of machine learning models that can be used to address this problem. These models take into account the data observation mechanisms and training procedures of a number of different algorithms. In order to verify the efficacy of our strategy, we combined the Heart Dataset with other classification models. The proposed method provides nearly 96 percent of accuracy result than other existing methods and the complete analysis over several metrics has been analysed and provided. Research in the field of deep learning will benefit from additional data from a large number of medical institutions, which may be used for the development of artificial neural network structures.

KEYWORDS

cardiovascular disease, AI-based technologies, internet of things, machine learning, computational method

1. Introduction

Cardiovascular disease (CVD), which is the leading cause of death globally, has become a significant problem in public health all over the world. As a result, patients, their families, and the governments of these countries have incurred substantial socioeconomic expenses. Patients at high risk for CVD can be identified by prediction models that use risk stratification. After that, measures that are tailored to this group, such as dietary changes and the use of statins, can help reduce that risk and contribute to the primary prevention of CVD (1).

Several guidelines for the evaluation and management of CVD have suggested using predictive models as a means of identifying patients at high risk and assisting with clinical decision-making. The Pooled Cohort Equations and the Framingham CV risk equation6, for example, have both been subjected to independent evaluations in a variety of populations; however, the findings indicated that both of these equations were only weakly discriminating and had a poor level of calibration (2).

As a direct consequence of this, the predictive power of the majority of the models that are now in use is restricted, and there is room for advancement. For instance, the assumption of linearity is necessary for logistic regression, while the assumption of predictor independence is necessary for the Cox proportional hazard model (3).

In the area of study pertaining to the cardiovascular system, machine learning (ML) algorithms have been demonstrated to be extremely helpful predictors. They are more adept than standard statistical models at capturing the complex interactions and nonlinear linkages that exist between the variables and the results (4). Several different investigations (5–15) came to the conclusion that random forests (RF) and support vector machines (SVM) performed better than traditional models.

Cardiovascular diseases such as coronary artery disease (CAD), atrial fibrillation (AF), and other cardiac or vascular ailments continue to be the leading cause of death in the world (10). As people living standards improve and their stress levels continue to rise, the number of people who suffer from CVD is growing at an alarming rate.

According to the most recent estimations (16, 17), CVD will be responsible for the deaths of about 23 million people by the year 2030. Infarction of the myocardium, atrial fibrillation, and heart failure are all instances of different types of CVD (18, 19). The incidence of cardiovascular disease can be influenced by a number of factors, including racial or ethnic background, age, gender, body mass index (BMI), height, and length of torso, as well as the outcomes of blood tests that evaluate factors such as renal function, liver function, and cholesterol levels (20, 21) which is shown in Figure 1.

The development of a wide variety of health problems can be influenced by the complex interactions that take place between these factors. Standard statistical approaches are incapable of accounting for all of the intricate causal links that exist between risk factors because there are so many of them (22, 23). In this day and age of big data, the Internet of Things (IoT) has been shown to be of critical importance. It has made it possible for patients to use smart drugs and smart bracelets to monitor and collect accurate data during a pandemic (24).

Researchers are employing artificial intelligence (AI) in an effort to mine new medical information that can be used by clinicians to better understand the symptoms of various diseases and, as a result, make more informed decisions for patients (25). This comes as the prevalence of data from the internet of things (IoT) grows within healthcare systems. In order to investigate previously unknown risk factors, current efforts to standardise medical data, and efforts to organise national health screening data (26–28), we will first standardise medical data. These risk variables may have a correlation with the occurrence of the disease, which means that they could offer insights into the mechanisms underlying the disease. Furthermore, accurate disease incidence prediction models necessitate the analysis of large amounts of data (29, 30). The use of artificial intelligence (AI) and massive amounts of data in the prediction of CVD models is becoming increasingly common.

The main contribution and novelty of this research is mentioned below:

- To extract a total of 11 distinct characteristics from the dataset.
- After that, we started by normalising the data and then proceeded to divide the Heart dataset into training and testing sets using an 8:2 split.
- Afterwards, the incorporated GBDT is utilised in the SHAP method for the purpose of selecting features.
- It helps to construct a stacking model consisting of a base learner layer in addition to a meta learner layer.
- Finally, we will achieve the results over several performance metrics analyses and method in the stacking model.

2. Background

Weng et al. (31) tested four different models using clinical data from over 300,000 homes in the United Kingdom. According to the findings, NN was the method that produced the most accurate CVD prediction results for the larger amount of data that were analysed.

The three traditional machine learning models that were tested and evaluated by Dimopoulos et al. (32) based on ATTICA data with 2020 samples for the little CVD dataset were K-Nearest Neighbour (KNN), Random Forest (RF), and Decision Tree. When compared, RF was shown to have produced the best results by using the HellenicSCORE tool, which is a calibration of the ESC Score.

In view of the growing popularity of machine learning techniques in IoT applications, Mohan et al. (15) have proposed a hybrid HRFLM strategy as a means of further improving the accuracy of the model predictions in light of the aforementioned popularity of machine learning methods.

An IoT-ML method was investigated by Akash et al. (33) with the goal of predicting the condition of the cardiovascular system in the human body. The algorithm model uses machine learning (ML) techniques to compute and predict the patient cardiovascular health after it has obtained essential data from the human body. This data include the patient heart rate, ECG signal, and cholesterol.

Within the framework of Yang et al. (34) examination of local locations with separate prediction models, LR was utilised to evaluate 30 cardiovascular disease-related characteristics utilising more than 200,000 high-risk participants in eastern China. The results of the experiments led to the development of an RF model that is more suited to eastern China.

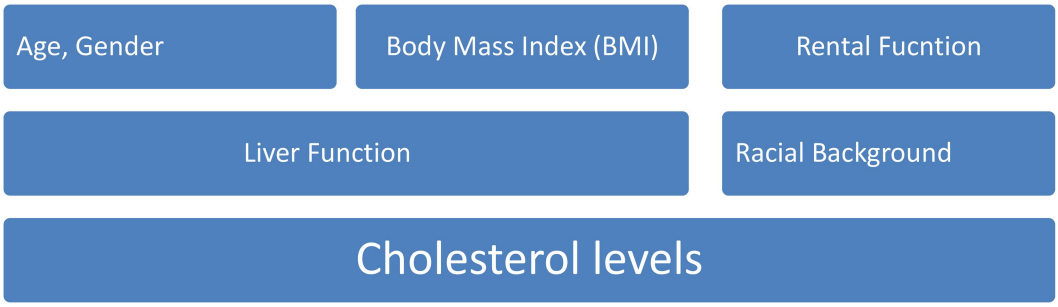


FIGURE 1
Several factor influencing incidence in cardiovascular disease.

For the first time in the study of CVDs, the idea of a stacking model was presented for the very first time by Yang et al. (35). The data on air pollution and weather were considered in order to have a better understanding of how the stacking model influences the daily hospitalisation rate for CVDs. In order to assist in the construction of the stacking model, a grassroots level of five basic learners was first constructed.

During this period, digital, fully automated ecosystems as well as cyber-physical systems are fast growing and finding applications all over the world. The creation of smart healthcare, which offers tools and processes for early diagnosis of life-threatening disorders, is one example of the innovative concepts and technical compositions that are being implemented in nearly every business. As the fourth industrial revolution moves towards a society that is more technologically advanced, there is an urgent requirement for additional research into CVD Zheng et al. (36).

3. Proposed method

The first thing that needs to be done is to combine the data from the Heart Dataset, which already contains information from Cleveland, Hungarian, and Swizerlang, as well as data from Long Beach VA and Stalog (Heart). From the five sources, we extracted a total of 11 distinct characteristics. After that, we started by normalising the data and then proceeded to divide the Heart dataset into training and testing sets using an 8:2 split. Afterwards, the incorporated GBDT is utilised in the SHAP method for the purpose of selecting features.

In the following stage, we will construct a stacking model consisting of a base learner layer in addition to a meta learner layer. The study uses RF, LR, MLP, ET, and CatBoost classifiers to serve as our base learners. LR is utilised in the role of the meta learner. Finally, the suggested stacking model is assessed with regard to its accuracy, precision, recall, F1 score, and area under the curve (AUC). In order to evaluate the model adaptability to new contexts, we made use of a publicly available dataset known as the Heart Attack Dataset.

The Cleveland, Hungarian, Swizerlang, Long Beach VA, and Stalog (Heart) datasets, together with others from the machine learning repository at the University of California, Irvine (UCI), were combined to form the Heart Dataset. We began with a total of 1,190 samples, and after deleting 272 duplicates, we were left with 918 unique sample datasets. We started with 1,190 samples. The whole Heart dataset is displayed in Table 1, and it consists of 11 features that were taken from five different datasets that contained significant relevant features.

3.1. Feature select and analysis

It is feasible to increase model performance and save a considerable amount of runtime by selecting the ideal subset of features that have a significant impact on the prediction outcomes. This process is referred to as feature selection, and it is possible to accomplish both of these goals.

The three most common methods for picking characteristics are called filters, wrappers, and embedding. The research we conducted utilised the embedded approach known as GBDT as a means of selecting feature variables. This was due to the fact that embedded techniques offer superior prediction performance compared to filter methods and are noticeably quicker than wrapper methods.

GBDT makes use of an additive model and a forward stepwise algorithm in order to achieve learning. These two components work together to accomplish this. For non-leaf nodes, the significance of the features increases proportionately with the magnitude of the reduction in weighted impurity that occurs during splitting.

Because of this, it is not possible to provide a detailed explanation of the role that each attribute plays in determining the overall accuracy of the predictions made by the integrated GBDT. In order to find a solution to this issue, we make use of a technique known as feature imputation, in which the explanatory model is a linear function of the values produced by feature imputation.

$$l\left(z' = \varnothing 0 + \sum \mathrm{Ni} = 1 \varnothing \mathrm{i} Z' i\right) \quad (1)$$

where N—features; $\varnothing_i$—feature attribute value, and Z′i—feature is valid or not.

The Φ_i value of Equation (1) can be determined by employing a tree-valued estimate methodology (also known as the SHAP method), which is founded on the concepts of game theory and used as the feature attribute values. Below is a formulation for a model f and a set S of non-zero Z' indices, as well as the conventional spherically valued attribute $\varnothing_i$ for each feature.

$$\varnothing i = \sum S \in M\{i\}|S|!\left(N - |S| - 1\right)!N!\left[f\left(S \cup \{i\} - f\left(S\right)\right)\right] \quad (2)$$

TABLE 1 Heart dataset features.

Feature	Detailed Information
Age	Age of the patient
Sex	Sex of the patient (Male: 0 or female: 1)
Chest pain type	Four chest pain types
	• ATA: atypical angina
	• TA: typical angina
	• ASY: asymptomatic
	• NAP: non-angina
Resting BP	Value of blood pressure during fasting (Unit mm hg)
Cholesterol	Concentration of serum cholesterol (Unit mm/dL)
Fasting BS	Value of blood glucose during fasting (1: blood glucose >120 mg/dL, 0: other)
Resting ECG	Resting electrocardiogram
Max HR	maximum heart rate
Exercise angina	Presence of exercise angina
Old peak	ST value decision
ST_Slope	Slope of ST section at the movement peak (up, flat, and down)

where M—input feature set.

It is essential to keep in mind that the SHAP strategy is adapted to the specific context and tailored to individual needs. In contrast to the tree model gain, this method can produce consistent results for global feature attributes. This is an advantage over the tree model gain. In the course of our study, we make use of the SHAP methodology in order to isolate and assess several individual characteristics.

In addition to this, we investigate the ways in which various characteristics interact with one another in order to improve our ability to predict outcomes. We differentiate between the contributions of individual features and the contributions of feature interactions by referring to the former as individual feature contribution and the latter as joint feature contribution $\Phi_{i,j}$. In the same way as the value, the Shapeley interaction index is calculated using a formula, and the joint feature contribution i and j can be found by doing the calculation as follows.

$$\varnothing i, j = \sum S \in M\{i\}|S|!(N-|S|-Z)!Z(N-1)!\nabla i, j(S) \quad (3)$$

When $i \neq j$:

$$\nabla i, j(S) = f(S \cup \{i,j\}) - f(S \cup \{i\}) - f(S \cup \{j\}) + f(S) \quad (4)$$

where Z represents the indices. i,j represent the feature contributions. S represents the Shapeley interaction Index.

Equations (3) and (4) in the GBDT model quantify the twinning relationships between joint features. So, when judging the model, you can get a good idea of how the different factors that interact with each other contribute together.

3.2. Model building

To the extent that the model predictions are accurate, each individual in the base population has a stronger capacity for learning, and the degree of correlation between them decreases. When the individual learners are already more accurate, a fusion of models will be more successful if the individual learners come from a diverse range of backgrounds. This is the foundation upon which the concept of error-ambiguity decomposition is built.

This suggests that when picking the foundation learners for our organisation, we should take into account the performance of individual learners while also taking into account the distinctiveness of each individual learner. Theoretically, it is conceivable to build layers of the stacking model indefinitely as long as their fundamental classifier is operational. This, of course, results in an increase in the level of complexity represented by the model.

To ensure accuracy while reducing the level of complexity exhibited by the model, we solely employ the stacking model, which is comprised of a two-tiered structure consisting of base learners and meta-learners. As a direct consequence of this, SVM, KNN, LR, and ET were decided upon as the possible models for base learners to utilise in the prediction of CVDs. XGBoost, LightGBM, CatBoost, and MLP were some of the other options that were thought about. Following the selection of the most reliable models as the foundation for our education, we restricted the pool of potential candidates to five people who exemplified a comprehensive representation of the population as a whole. The optuna framework was used in order to get the optimal values for the model parameters.

After running a 5-fold CV, this model may generate a large number of features. 5-fold CV is a technique that is frequently utilised in the first layer of a stacking framework to collect input features for the second layer. This paper employs linear regression (LR) as the

classifier for the fusion model predictions since generalised linear regression, also known as GLR, has historically been employed in the second layer of the stacking architecture. Because adjusting the complexity of the output layer of a neural network does not require the employment of more complex functions, this example makes use of functions that are simpler in nature.

The primary learners are the LR, RF, DT, MLP, and CatBoost protocols. At the beginning, we give the training sets eight times as many points as the testing sets. Within the training package that we provide for each of the five foundational learners, we utilise a 5-fold CV. A single base learner is capable of producing five predictions, and each of these five predictions is arranged in a vertical column within a one-dimensional matrix. It possible that the second stage of training will be based on a five-dimensional matrix that been developed using the data of five different learners as its foundation.

When applied to the testing set, the 5-fold CV model is utilised once more to make predictions about our initial testing set, which results in the production of five predictions once more. The base learners can be concatenated into a matrix for the stage second iteration. We use LR on the meta-learner so that it does not become too good at its job. In the second step of the process, we use these predictions to build the final results.

4. Results and discussion

The outcomes of the experiments will be discussed here in order to illustrate the benefits of the stacking paradigm that was recommended by us. Python version 3.9.7 was used throughout each and every test. In this investigation, the sklearn 1.0.2 toolbox is used for model prediction. The SHAP 40.0 toolbox is used for feature selection, and the Optuna 2.10.0 framework is used to determine the optimum values for the model parameters which is shown in Table 2. We executed 10 splits of the data set using various random seeds in order to account for the small sample size of this study and the aforementioned randomisation. After doing so, we averaged the results of all 10 experiments.

Before we started the feature selection process, our dataset contained a total of 11 features. Using the Tree SHAP approach, you are able to determine the contribution value that corresponds to each feature that is contained inside the sample dataset. The ranking of the feature contributions is determined by using the average SHAP value for all of the samples. In accordance with the GBDT model, the contributions of the global features are depicted. The ST Slope and Chest Pain Type have a significant influence on the patient condition (CVD) in patients with cardiovascular disease. In order to cut the model operating time even more, some features that aren't necessary will have to be eliminated. We chose to adopt a cutoff of 0.02, which led to the elimination of the Resting ECG characteristic while permitting the retention of the other 10 features. We used the AUC to evaluate the performance both before and after the feature selection process. Even though the AUC values of GBDT went down, the drop wasnot substantial at all, and there was not any difference that could be considered statistically significant by performing metrics such as AUC, Threshold, Sensitivity, Specificity which is shown in Figures 2–5.

TABLE 2 Software specifications.

Language	Python Version 3.9.7
Operating system	Windows 11
Tool box for model predicition	Sklearn 1.0.2
Feature selection	SHAP 40.0
Optimum values	Optuna 2.10.0 framework

The incidence of CVD was quite low in this experiment, resulting in poor PPV and NPV values for each of the seven different ML models. Because of this, their therapeutic value may suffer as a result of an increase in the number of false-positive results. The probabilities that were predicted by each machine learning model were unique, and the risk distribution for LR was comparable to that of SVM. Patients who had a CVD episode had estimated risks that were greater, across all ML models, than those patients who had not had a CVD episode. The plots also demonstrated that all ML models overestimated the risks of those individuals who had not experienced any CVD events. This finding suggests that this variable may also affect how well a model predicts what will happen.

It is necessary to have a risk model in order to determine whether individuals have a high probability of developing CVD. We intended to test seven machine learning (ML)-based models to evaluate how correctly they could predict the risk of CVD. The findings demonstrated that each one of them had good to excellent discrimination and that they were all accurately calibrated. When it came to forecasting the risk of CVD, penalised LR performed better than other machine learning models, just like SVM did. The specificity of the SVM was higher than that of the LR, while the LR had a lower level of sensitivity. It is possible that a higher level of specialisation was favoured in this Kazakh Chinese group because the majority of the population was nomadic and there were few medical services available. In addition to this, when taking calibration and DCA into consideration, SVM fared marginally better than LR. Because of this, SVM and LR can be used to find people in this group who are at a higher risk of CVD and to find out if putting risk-mitigation interventions in place for these people will improve their CVD outcomes during the clinical decision-making process.

Linear regression has been widely used in the clinic to construct predictive models due to the ease with which it may be interpreted and its general straightforwardness. In a study aimed at predicting myocardial ischemia, both LR and SVM were shown to have the same level of predictive ability, which was consistent with our findings. A recent and exhaustive study concluded that there is no performance benefit to be gained from using ML in clinical prediction models over using LR. It was determined that when machine learning algorithms were applied to small datasets with a limited number of predictors, LR models might perform better than ML models in terms of overall performance. It is possible that the small sample size and the L1 penalised technique used in this work are to blame for the superior performance of LR in comparison to other machine learning models, with the exception of SVM.

The Support Vector Machine (SVM) is a well-known supervised machine learning approach that has found applications in a wide variety of business sectors. The fundamental idea behind support vector machines (SVM) is to locate the hyperplane that has the

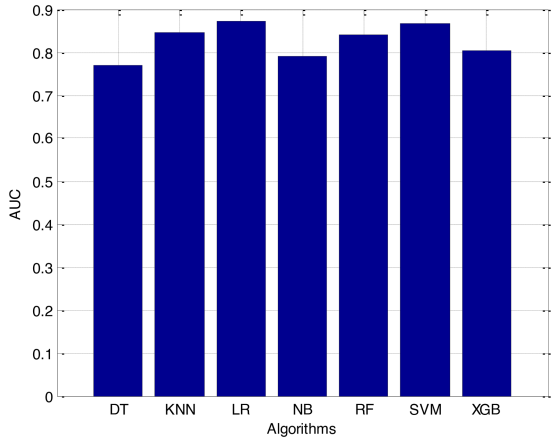


FIGURE 2
Area under the curve (AUC).

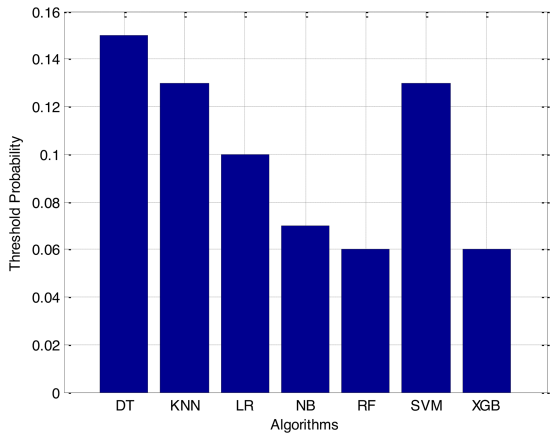


FIGURE 3
Threshold probability.

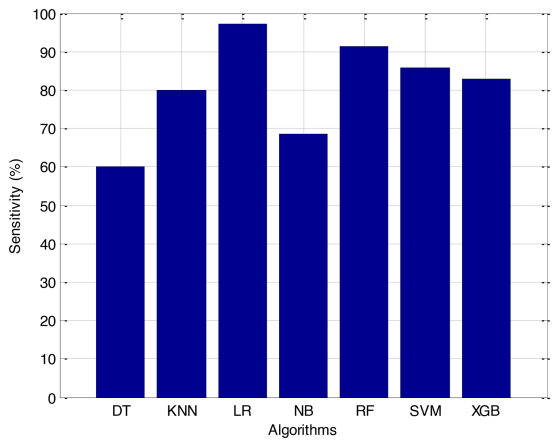


FIGURE 4
Sensitivity (%).

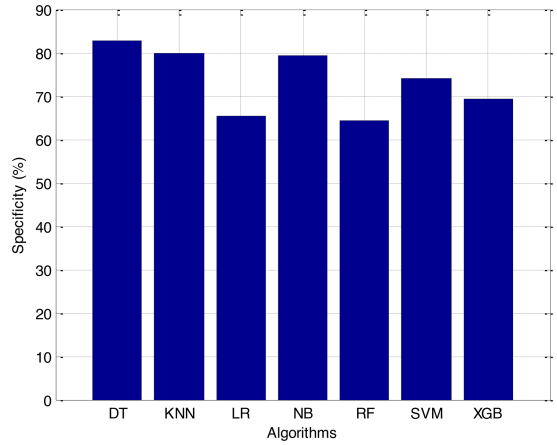


FIGURE 5
Specificity (%).

capacity to effectively classify the data while also providing the biggest geometric margin. In addition to this, it possesses significant kernel capabilities, which simplify the process of dealing with nonlinear classification issues. The outstanding performance of SVM demonstrates that it is a great tool for tackling classification challenges on small, non-linear, and high-dimensional datasets. This demonstrates that SVM is an excellent tool. In our experiment, we observed that the SVM performed significantly better than other machine learning models.

When it comes to classification, RF is among the most successful ensemble learning strategies that may be used. Its predictions were not nearly as accurate as those generated by the LR and SVM algorithms, which were the other two options. It is likely that the limited number of participants in this study will prevent RF from achieving its full potential as a prediction tool. The concept of variable importance was utilised in order to locate potential indicators of CVD. Some studies suggest that RF may be capable of revealing crucial but undisclosed predictions.

According to the results of feature selection that was based on RF, the age of the patient was the most significant predictor in the classification of CVD. In this study, it was discovered that certain risk factors, such as smoking and alcohol intake, were not as predictive as previously believed. Previous studies have shown that the synthetic indices BAI and LHR are both very good indicators of cardiovascular disease. Inflammation plays a significant part in the formation of atherosclerotic plaques as well as1 the progression of cardiovascular disease is shown in Figures 6–11.

There is evidence that inflammatory cytokines, such as high-sensitivity CRP and interleukin-6, are associated with an elevated risk of cardiovascular disease. The Hs-CRP inflammatory marker was included in the Reynolds Risk Score in order to account for its role as a potential contributor to cardiovascular disease. hs-CRP has been shown in a number of other epidemiological studies to be an important predictor of CVD. These studies have also shown that hs-CRP acts as a mediator in the pathogenesis of vascular disease and is a marker of endothelial dysfunction. These findings are consistent with the findings of the aforementioned studies. It was discovered that Hs-CRP increases the risk of atherosclerotic plaque rupture in

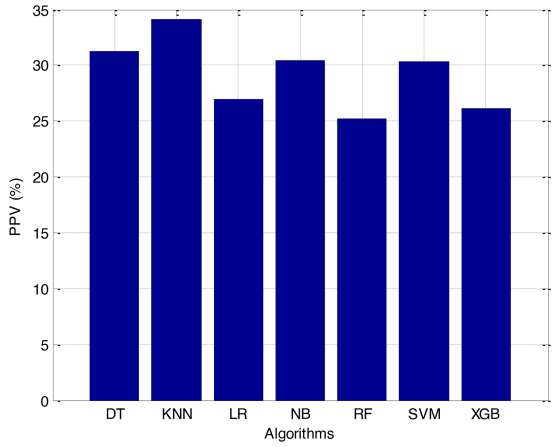


FIGURE 6
PPV (%).

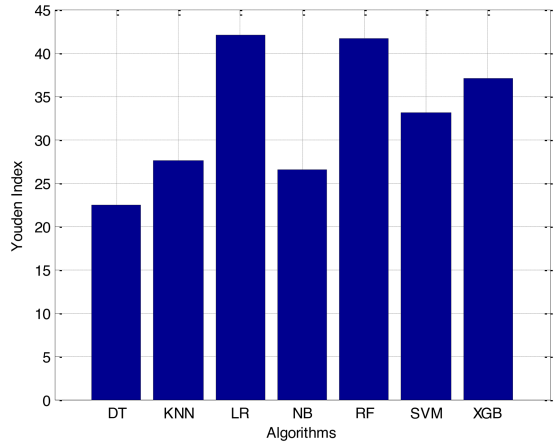


FIGURE 9
High-risk patients (%).

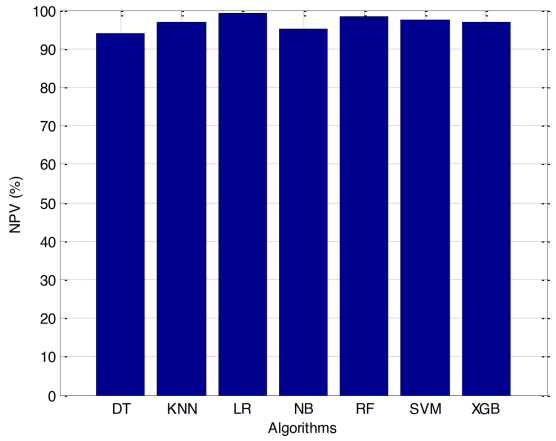


FIGURE 7
NPV (%).

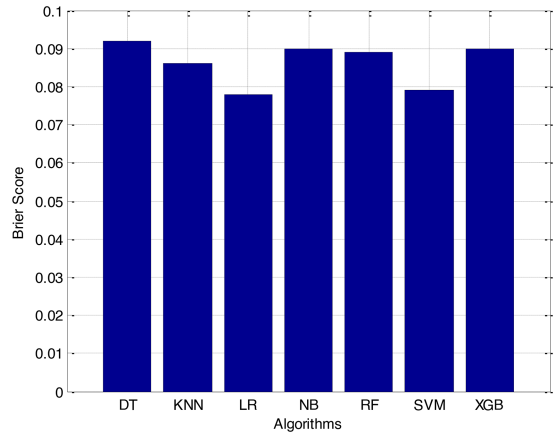


FIGURE 10
Brier score.

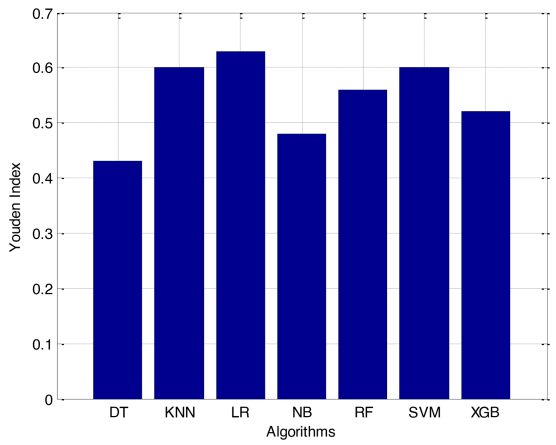


FIGURE 8
Youden index.

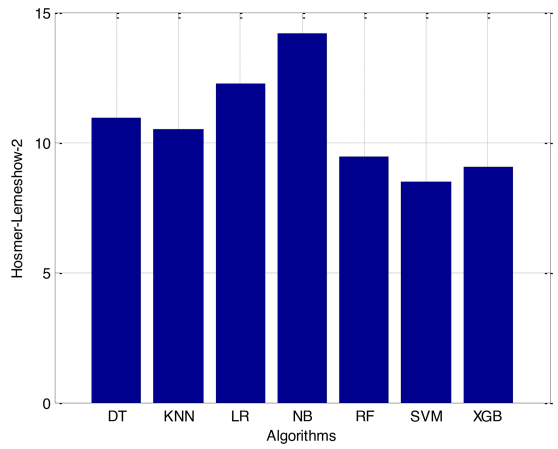


FIGURE 11
Hosmer-Lemeshow-2.

addition to destabilising atherosclerotic plaques *via* NO, IL-6, and prostacyclin.

In addition, hs-CRP has been demonstrated to enhance thrombosis and cardiomyocyte apoptosis in response to hypoxia, which provides more support for its position as a risk factor for cardiovascular disease. It has been demonstrated that IL-6 is a factor in the course of atherosclerosis and that it promotes the creation of atherosclerotic plaques. This factor may have contributed to the increase in the number of cases of CVD. Taking statins, which can reduce a person chance of acquiring CVD, is beneficial for a great number of people and can help them avoid developing the condition. In clinical practise, Hs-CRP and IL-6 can be used as biomarkers for the early diagnosis of patients who have an increased likelihood of developing cardiovascular disease.

According to the findings of our study, a decrease in ADP was associated with an increased risk of developing cardiovascular disease. The adipose hormone ADP, which is secreted by adipocytes, has anti-inflammatory effects. These effects manifest themselves as a reduction in the levels of CRP and lymphocyte recruitment in atherosclerotic lesions, a reduction in the expression of TNF-, and an increase in the production of cytokines that are protective against inflammation. On the other hand, there is evidence from a small number of studies that suggests an increase in ADP may assist in avoiding an ischemic stroke. Increased NEFA concentrations have been associated with an increased risk of cardiovascular disease in previous research, and our study came to the same conclusion. The possible effects of NEFA on cardiovascular disease include the potential to exacerbate or worsen a number of illnesses, including type 2 diabetes, hypertension, the metabolic syndrome, and endothelial deterioration, to name a few. Patients can have a lower chance of developing cardiovascular disease if they are treated to have a lower ADP (CVD).

The risk prediction models that are currently being used in CVD domains were built using traditional statistical methodologies, as many studies have revealed. Nevertheless, these models have been proven to be erroneous in external populations. In the field of cardiology, machine learning algorithms have proven to be superior methods for deriving predictions from big datasets that are notoriously difficult to understand. No prior assumptions are made by machine learning algorithms, which means that any data can be used to develop accurate and resilient models. Because of this, ML is able to model more complex relationships between outcomes and predictors, which are typically more challenging to express using standard statistical methods. Discovering interactions between marginal predictors can help improve risk-management strategies when using ML.

Machine learning has the potential to identify new genotypes and phenotypes for a variety of CVDs, as well as novel risk factors for CVDs. All aspects of medical picture recognition, diagnosis, outcome prediction, and prognosis evaluation can be improved with the application of more sophisticated machine learning techniques such as deep learning and artificial neural networks (ANN). It possible that in the future, cardiologists will make better clinical decisions if they use machine learning models rather than the CVD risk stratifications that are currently used. On the other hand, most ML models may be hard for medical professionals to understand and use, which may limit how often they can be used in clinical settings.

5. Conclusion

According to the findings of this research, a stacking fusion model-based classifier performs better than individual models on all assessment criteria. This finding suggests that stacking models can combine the benefits of a variety of model types to achieve superior prediction performance. The recommended stacking approach offers improved prediction performance, increased resilience, and increased utility for individuals who are at high risk of developing cardiovascular disease. Hospitals can utilise this information to identify patients who are at a high risk of developing cardiovascular disease and provide them with early clinical intervention in order to reduce that risk. Research in the field of deep learning will benefit from additional data from a large number of medical institutions, which may be used for the development of artificial neural network structures or for the usage of deep learning frameworks in the future. In future work, the other deep learning techniques algorithms can be incorporated into Internet of Things (IoT) environments which helps to achieve more accuracy in terms of result and it can be useful to the hospitals and saving several human life.

Data availability statement

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

Author contributions

MA and KS: conceptualization. SS, NV, and MA: methodology and investigation. TU and LA-k: software. MSo, TU, and NA: validation. KA and RK: formal analysis. KA and MA: data curation. SS and NV: writing—original draft preparation. MA, MSo, LA-k, and RK: writing—review and editing. LA-k, NA, and RK: supervision. All authors contributed to the article and approved the submitted version.

Acknowledgments

The authors are thankful to the princess Nourah Bint Abdulrahman University researcher, supporting program number (PNURSP2023R82) Princess Nourah bint Abdulrahman University, Riyadh Saudia arabia.

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.

References

1. Al Aref SJ, Anchouche K, Singh G, Slomka PJ, Kolli KK, Kumar A, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. (2019) 40:1975–86. doi: 10.1093/eurheartj/ehy404

2. Ghosh P, Azam S, Jonkman M, Karim A, Shamrat FJM, Ignatious E, et al. Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access*. (2021) 9:19304–26. doi: 10.1109/ACCESS.2021.3053759

3. Krittanawong C, Virk HUH, Bangalore S, Wang Z, Johnson KW, Pinotti R, et al. Machine learning prediction in cardiovascular diseases: a meta-analysis. *Sci Rep*. (2020) 10:1–11. doi: 10.1038/s41598-020-72685-1

4. Alaa AM, Bolton T, Di Angelantonio E, Rudd JH, Van der Schaar M. Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK biobank participants. *PLoS One*. (2019) 14:e0213653. doi: 10.1371/journal.pone.0213653

5. Mezzatesta S, Torino C, De Meo P, Fiumara G, Vilasi A. A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. *Comput Methods Prog Biomed*. (2019) 177:9–15. doi: 10.1016/j.cmpb.2019.05.005

6. Brites I. S., Silva L. M., Barbosa J. L., Rigo S. J., Correia S. D., Leithardt V. R. (2022) "Machine learning and iot applied to cardiovascular diseases identification through heart sounds: a literature review." in *International Conference on Information Technology & Systems*. Springer, Cham, 356–388.

7. Dinesh K. G., Arumugaraj K., Santhosh K. D., Mareeswari V. (2018) "Prediction of cardiovascular disease using machine learning algorithms." in *2018 International Conference on Current Trends towards Converging Technologies (ICCTCT)*. IEEE, 1–7.

8. Li Y, Sperrin M, Ashcroft DM, Van Staa TP. Consistency of variety of machine learning and statistical models in predicting clinical risks of individual patients: longitudinal cohort study using cardiovascular disease as exemplar. *BMJ*. (2020) 371:1–8. doi: 10.1136/bmj.m3919

9. Padmanabhan M, Yuan P, Chada G, Nguyen HV. Physician-friendly machine learning: a case study with cardiovascular disease risk prediction. *J Clin Med*. (2019) 8:1050. doi: 10.3390/jcm8071050

10. Aryal S, Alimadadi A, Manandhar I, Joe B, Cheng X. Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. *Hypertension*. (2020) 76:1555–62. doi: 10.1161/HYPERTENSIONAHA.120.15885

11. Dinh A, Miertschin S, Young A, Mohanty SD. A data-driven approach to predicting diabetes and cardiovascular disease with machine learning. *BMC Med Inform Decis Mak*. (2019) 19:1–15. doi: 10.1186/s12911-019-0918-5

12. Ponnusamy M, Bedi P, Suresh T, Alagarsamy A, Manikandan R, Yuvaraj N. Design and analysis of text document clustering using salp swarm algorithm. *J Supercomput*. (2022) 78:16197–213. doi: 10.1007/s11227-022-04525-0

13. Allan S, Olaiya R, Burhan R. Reviewing the use and quality of machine learning in developing clinical prediction models for cardiovascular disease. *Postgrad Med J*. (2022) 98:551–8. doi: 10.1136/postgradmedj-2020-139352

14. Yadav A, Singh A, Dutta MK, Travieso CM. Machine learning-based classification of cardiac diseases from PCG recorded heart sounds. *Neural Comput Applic*. (2020) 32:17843–56. doi: 10.1007/s00521-019-04547-5

15. Mohan S, Thirumalai C, Srivastava G. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*. (2019) 7:81542–54. doi: 10.1109/ACCESS.2019.2923707

16. Juhola M, Joutsijoki H, Penttinen K, Aalto-Setälä K. Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods. *Sci Rep*. (2018) 8:1–10. doi: 10.1038/s41598-018-27695-5

17. Maheshwari V, Mahmood MR, Sravanthi S, Arivazhagan N, ParimalaGandhi A, Srihari K, et al. Nanotechnology-based sensitive biosensors for COVID-19 prediction using fuzzy logic control. *J Nanomater*. (2021) 2021:1–8. doi: 10.1155/2021/3383146

18. Maini E., Venkateswarlu B., Gupta A. (2018). "Applying machine learning algorithms to develop a universal cardiovascular disease prediction system" in *International Conference on Intelligent Data Communication Technologies and Internet of Things*. Springer, Cham. 627–632.

19. Li Q, Campan A, Ren A, Eid WE. Automating and improving cardiovascular disease prediction using machine learning and EMR data features from a regional healthcare system. *Int J Med Inform*. (2022) 163:104786. doi: 10.1016/j.ijmedinf.2022.104786

20. Maiga J., Hungilo G. G. (2019). "Comparison of machine learning models in prediction of cardiovascular disease using health record data." in *2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)* IEEE, 45–48.

21. Sivasankari S. S., Surendiran J., Yuvaraj N., Ramkumar M., Ravi C. N., Vidhya R. G. (2022). "Classification of diabetes using multilayer perceptron." in *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE, 1–5.

22. Rahim A, Rasheed Y, Azam F, Anwar MW, Rahim MA, Muzaffar AW. An integrated machine learning framework for effective prediction of cardiovascular diseases. *IEEE Access*. (2021) 9:106575–88. doi: 10.1109/ACCESS.2021.3098688

23. Maurovich-Horvat P. Current trends in the use of machine learning for diagnostics and/or risk stratification in cardiovascular disease. *Cardiovasc Res*. (2021) 117:e67–9. doi: 10.1093/cvr/cvab059

24. Ahn I, Gwon H, Kang H, Kim Y, Seo H, Choi H, et al. Machine learning–based hospital discharge prediction for patients with cardiovascular diseases: development and usability study. *JMIR Med Inform*. (2021) 9:e32662. doi: 10.2196/32662

25. Arunachalam SK, Rekha R. A novel approach for cardiovascular disease prediction using machine learning algorithms. *Concurr Comput Pract Exp*. (2022) 34:e7027. doi: 10.1002/cpe.7027

26. Kannan S, Yuvaraj N, Idrees BA, Arulprakash P, Ranganathan V, Udayakumar E, et al. Analysis of convolutional recurrent neural network classifier for COVID-19 symptoms over computerised tomography images. *Int J Comput Appl Technol*. (2021) 66:427–32. doi: 10.1504/IJCAT.2021.120453

27. Smita , Kumar E. Probabilistic decision support system using machine learning techniques: a case study of cardiovascular diseases. *J Discret Math Sci Cryptogr*. (2021) 24:1487–96. doi: 10.1080/09720529.2021.1947452

28. Di Castelnuovo A, Bonaccio M, Costanzo S, Gialluisi A, Antinori A, Berselli N, et al. Common cardiovascular risk factors and in-hospital mortality in 3,894 patients with COVID-19: survival analysis and machine learning-based findings from the multicentre Italian CORIST study. *Nutr Metab Cardiovasc Dis*. (2020) 30:1899–913. doi: 10.1016/j.numecd.2020.07.031

29. Shu S, Ren J, Song J. Clinical application of machine learning-based artificial intelligence in the diagnosis, prediction, and classification of cardiovascular diseases. *Circ J*. (2021) 85:1416–25. doi: 10.1253/circj.CJ-20-1121

30. Nashif S, Raihan MR, Islam MR, Imam MH. Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. *World J Eng Technol*. (2018) 06:854–73. doi: 10.4236/wjet.2018.64057

31. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One*. (2017) 12:e0174944. doi: 10.1371/journal.pone.0174944

32. Dimopoulos AC, Nikolaidou M, Caballero FF, Engchuan W, Sanchez-Niubo A, Arndt H, et al. Machine learning methodologies versus cardiovascular risk scores, in predicting disease risk. *BMC Med Res Methodol*. (2018) 18:1–11. doi: 10.1186/s12874-018-0644-1

33. Zaman M. I. U., Tabassum S., Ullah M. S., Rahaman A., Nahar S., Islam A. M. (2019). "Towards IoT and ML driven cardiac status prediction system." in *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, 1–6. IEEE

34. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep*. (2020) 10:1–8. doi: 10.1038/s41598-020-62133-5

35. Hu Z, Qiu H, Su Z, Shen M, Chen Z. A stacking ensemble model to predict daily number of hospital admissions for cardiovascular diseases. *IEEE Access*. (2020) 8:138719–29. doi: 10.1109/ACCESS.2020.3012143

36. Zheng H, Sherazi SWA, Lee JY. A stacking ensemble prediction model for the occurrences of major adverse cardiovascular events in patients with acute coronary syndrome on imbalanced data. *IEEE Access*. (2021) 9:113692–704. doi: 10.1109/ACCESS.2021.3099795

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 05 May 2023
DOI 10.3389/fmed.2023.1164305

OPEN ACCESS

EDITED BY
Balu Kamaraj,
Imam Abdulrahman Bin Faisal University,
Saudi Arabia

REVIEWED BY
Khurshid Ahmad,
Yeungnam University, Republic of Korea
Neeraja Reddy Matavalam,
Sri Venkateswara University, India

*CORRESPONDENCE
Mahmoud Hisham Mosli
✉ hmosli@hotmail.com
Ramu Elango
✉ relango@kau.edu.sa

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

RECEIVED 12 February 2023
ACCEPTED 14 March 2023
PUBLISHED 05 May 2023

CITATION
Jan RM, Al-Numan HH, Al-Twaty NH, Alrayes N, Alsufyani HA, Alaifan MA, Alhussaini BH, Shaik NA, Awan Z, Qari Y, Saadah OI, Banaganapalli B, Mosli MH and Elango R (2023) Rare variant burden analysis from exomes of three consanguineous families reveals *LILRB1* and *PRSS3* as potential key proteins in inflammatory bowel disease pathogenesis. *Front. Med.* 10:1164305.
doi: 10.3389/fmed.2023.1164305

Rare variant burden analysis from exomes of three consanguineous families reveals *LILRB1* and *PRSS3* as potential key proteins in inflammatory bowel disease pathogenesis

Rana Mohammed Jan[1,2], Huda Husain Al-Numan[1,2], Nada Hassan Al-Twaty[1], Nuha Alrayes[2,3], Hadeel A. Alsufyani[4], Meshari A. Alaifan[5], Bakr H. Alhussaini[5], Noor Ahmad Shaik[2,6], Zuhier Awan[7], Yousef Qari[8], Omar I. Saadah[5,9], Babajan Banaganapalli[2,6], Mahmoud Hisham Mosli[7,9]* and Ramu Elango[2,6]*

[1]Department of Biological Sciences, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia, [2]Princess Al-Jawhara Al-Brahim Center of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia, [3]Department of Medical Laboratory Sciences, Faculty of Applied Medical Sciences, King Abdulaziz University, Jeddah, Saudi Arabia, [4]Department of Medical Physiology, Faculty of Medicine, King Abdulaziz University Hospital, Jeddah, Saudi Arabia, [5]Department of Pediatrics, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, [6]Department of Genetic Medicine, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, [7]Department of Clinical Biochemistry, Faculty of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, [8]Department of Internal Medicine, King Abdulaziz University, Jeddah, Saudi Arabia, [9]Inflammatory Bowel Disease Research Group, King Abdulaziz University, Jeddah, Saudi Arabia

Background: Inflammatory bowel disease (IBD) is a chronic autoimmune disorder characterized by severe inflammation and mucosal destruction of the intestine. The specific, complex molecular processes underlying IBD pathogenesis are not well understood. Therefore, this study is aimed at identifying and uncovering the role of key genetic factors in IBD.

Method: The whole exome sequences (WESs) of three consanguineous Saudi families having many siblings with IBD were analyzed to discover the causal genetic defect. Then, we used a combination of artificial intelligence approaches, such as functional enrichment analysis using immune pathways and a set of computational functional validation tools for gene expression, immune cell expression analyses, phenotype aggregation, and the system biology of innate immunity, to highlight potential IBD genes that play an important role in its pathobiology.

Results: Our findings have shown a causal group of extremely rare variants in the *LILRB1* (Q53L, Y99N, W351G, D365A, and Q376H) and *PRSS3* (F4L and V25I) genes in IBD-affected siblings. Findings from amino acids in conserved domains, tertiary-level structural deviations, and stability analysis have confirmed that these variants have a negative impact on structural features in the corresponding proteins. Intensive computational structural analysis shows that both genes have very high expression in the gastrointestinal tract and immune organs and are involved in a variety of innate immune system pathways. Since the innate

immune system detects microbial infections, any defect in this system could lead to immune functional impairment contributing to IBD.

Conclusion: The present study proposes a novel strategy for unraveling the complex genetic architecture of IBD by integrating WES data of familial cases, with computational analysis.

KEYWORDS

inflammatory bowel disease, missense mutation, Crohn's disease, gastrointestinal tract, protein modeling

1. Introduction

Inflammatory bowel disease (IBD) is a chronic immune disorder characterized by severe inflammation and mucosal destruction in the colon and small intestine (1, 2). Crohn's disease (CD) and ulcerative colitis (UC) are the two major forms of IBD, which share identical pathological and clinical symptoms (2, 3). However, each condition shows a variable clinical presentation, response to treatment, and genetic risk factors (3). Recent decades have seen a sharp increase in the prevalence of IBD, which could be attributed to industrialization and lifestyle changes. The high prevalence of consanguinity in the Arab population results in the perpetuation of numerous harmful genetic variants in society. This aggregation of damaging variants in key genes may cause rare monogenic diseases and increase the genetic contribution to complex diseases such as IBD. Although the primary cause of IBD is unknown, interactions between environmental and immunoregulatory variables have been identified as a probable cause in genetically predisposed individuals (4, 5).

There is a clear evidence that genetic factors play an important role, with relatives of UC and CD patients having 8- to 10-fold increased risk of developing IBD (6). The strongest evidence for a genetic predisposition to IBD came from twin studies. While genetic defects in the IL-10 signaling pathway have been identified as an underlying molecular cause for very-early-onset IBD (VEO-IBD), no single causal genetic factor has been identified for late-onset IBD. This is because, late-onset IBD has a polygenic etiology, and environmental factors determine the susceptibility and age of onset of the disease (7). However, genome-wide association studies (GWAS) have uncovered more than 200 common risk loci in IBD pathogenesis (8–12). Some of these risk alleles are missense variants that have been mapped to genes such as Interleukin 23 Receptor (*IL23R*), Nucleotide Binding Oligomerization Domain containing 2 (*NOD2*), and Autophagy-related 16 like 1 (*ATG16L1*) (13). Majority of these risk markers are intronic variants (14).

Thus, to better understand the pathogenesis of complex diseases, application of next-generation sequencing technologies is having a greater impact, especially in consanguineous societies (15–17). They will provide an excellent opportunity to identify rare variants with intermediate to high effect ranges more efficiently. These rare variants are believed to have high odds ratios (ORs) and high penetrance and are suitable for functional experimental validation. In genetics, OR is often used to quantify the risk of developing a particular disease in individuals who carry a specific genetic variant or mutation. In a recent study, one rare coding variant in the *BTNL2* gene within the Major histocompatibility complex (MHC) region was associated with higher IBD risk (OR-2.3), giving an insight into T cell activation mechanisms and IBD sub-phenotype developments (18). It provides strong support for our planned approach to identify potential causal variants and genes for IBD through familial studies. Since published information on the genetics of Arab IBD familial patients is limited, the goal of this study is to find out the causal genetic variants involved in IBD pathogenesis.

2. Materials and methods

2.1. Recruitment of families with IBD

The Biomedical Ethics Research Committee of King Abdulaziz University Hospital in Jeddah (KAUH) approved the proposed research project. At the Internal Medicine specialty gastroenterology clinics at King Abdulaziz University Hospital, Jeddah (KAUH), three unrelated Saudi consanguineous families with many affected siblings, who fulfilled the inclusion criteria of the study, reporting abdominal pain along with weight loss and persistent diarrhea, were recruited. An informed consent to join the research as participants was signed by all family members before we collected clinical data and blood samples. Family A has two siblings with IBD, and families B and C each have three siblings with IBD. All these patients were examined by a consultant gastroenterologists, and the diagnosis was arrived at as per the standard diagnostic criteria set out by the European Crohn's and Colitis Organization (ECCO) 2019 (19). After collecting the full family history, a three-generation pedigree for each family was constructed. Hospital electronic health records were accessed to collect clinical history on all affected siblings. For genetic analysis, approximately 3–4 mL of peripheral blood was collected in EDTA tubes from all participants and stored at −80°C until used.

2.2. DNA purification

Genomic DNA was purified according to the manufacturer's instructions using the QIAamp DNA Blood Kit (Qiagen, United States). A Nanodrop (ND-1000 UV–VIS) spectrophotometer was used to measure DNA concentration and purity. The DNA integrity for high molecular weight DNA was evaluated using 1% agarose gel electrophoresis, and the gel image was captured in a UV transilluminator attached camera. All the samples were stored at −20°C until they were used for genetic analysis.

2.3. Whole exome sequence analysis

Whole exome sequencing was performed using the Illumina HiSeq2000 next-generation sequencer (Illumina Inc., San Diego, CA, United States). The whole exome-enriched library was constructed using genomic DNA at an average concentration of 60 ng/μL, including DNA tagmentation (fragmentation and adapter ligation at both ends), target capturing, and amplification using the ligated adapters. The Agilent SureSelect exome capture kit V7.0 (Agilent Technologies, United States) was used to shear all exonic sections of protein-coding genes that were registered in the CCDS and RefSeq databases, resulting in ideal size-range fragments. Ultra-long 120-mer biotinylated cRNA library baits were used to hybridize the fragmented DNA. Capillary electrophoresis was used to determine the concentration and size of the library. During enrichment, various adapters were incorporated, allowing the samples to be amplified for subsequent sequencing. For variant calling and annotation, the sequencing reads (in the FASTQ format) were matched to the human genome reference sequence build 38 (GRCH38.p12) using BLAST (version 0.6.4d). Variants were filtered based on the following criteria: depth (30), maximum quality read (60), alternative to total depth ratio (>80% for homozygous variants and 40–70% for heterozygous variants), minor allele frequency (<0.01) based on the 1,000 genomes, gnomAD database, and location (coding regions or regulatory sites). The rare variants were further filtered based on the segregation pattern of the variants under different genetic inheritance models such as autosomal recessive (AR), compound heterozygous (CH), and *de novo* to identify the disease-causing variants.

2.3.1. Identifying the rare variant burden genes

Since IBD is a complex disease with polygenic involvement, we tried to identify the genes with a rare variant burden. From the exome sequencing data of individual families, we attempted to identify genes harboring rare variants to see which genes are potentially involved in the disease causation.

2.4. Functional enrichment analysis using immune pathways

The rare variant harboring genes shared between the three families were initially identified by the Venny 2.1.0 web tool.[1] The ClueGo, a Cytoscape plug-in was then used to perform functional enrichment analysis on these rare variant genes. For pathway enrichment of query genes, the GO annotations was chosen in the ClueGo settings (6). In this enrichment test, default stringent statistical options, such as Bonferroni multiple testing correction and enrichment/depletion (Two-sided hypergeometric test), were applied. The common pathways (enriched GO terms) among all three families were identified by the VENNY tool. The pathways corresponding to the mapped genes with rare variants that were shared by all three families were then further filtered to exclude contributing genes that were not included in the initial query list of shared rare variant genes.

1 https://bioinfogp.cnb.csic.es/tools/venny/

2.5. Computational functional validation of selected potential IBD genes

The shared genes with rare variants from the pathway analysis were further filtered to validate their potential contribution to disease development. To this end, several databases were used to explore their gene expression levels in different organs and to prioritize the potential therapeutic drug targets and disease phenotype annotations.

2.5.1. Gene expression analysis and exome validation

We examined the changes in the expression status of our query genes in IBD tissues by downloading 24 IBD-related transcript expression datasets hosted in Expression Atlas.[2] This database is maintained by the European Bioinformatics Institute and provides information on gene expression patterns from RNA-seq, microarray studies, and protein expression from proteomics studies. The keywords searched in the database were IBD and inflammation. Different experimental samples were used, such as colonic, mucosal biopsies and peripheral blood monocytes, for different diseases such as UC, IBD, CD, irritable bowel syndrome, colorectal cancer, and colon adenomas. From the resultant datasets, we identified differentially expressed genes (DEGs) using a logFC cutoff fold change of >1 at $p < 0.05$. Furthermore, the EBI gene expression atlas (GXA) interface in Ensembl was used to search for transcript expression data of the query genes in different organs and tissues. The input is the gene name, and the output is the baseline expression in transcripts per million (TPM). Only the expression data of query genes (>0.5 TPM cutoff value) in the gastrointestinal tract, immunological organs, and blood were chosen from the output.

2.5.2. Immune cell expression analysis

The Database of Immune Cell Expression (DICE)[3], expression quantitative trait loci (eQTLs), and epigenomics were used to reveal the effect of IBD risk-associated genetic polymorphisms on specific immune cell types which might influence disease pathogenesis. This database delivers comprehensive information on immune cell expression generated by 15 immune cell types (subsets of T cells, B cells, monocytes, and NK cells). The input is the query gene ID, and the output is the expression level of genes in transcripts per million (TPM) on the *x*-axis, and cell types are sorted based on the *y*-axis of box plot graphs.

2.5.3. Open target phenotype identification

The query hub genes were further analyzed using the Open Targets Platform.[4] This website accesses several databases to help in clarifying the causal relationships between enzymatic reactions, physical binary interactions, or functional relationships between disease phenotypes and therapeutic targets (6). The input is the query gene list, and the output is the evidence score for a given target-disease pair. The significant value was set at a 0.5 cutoff score to detect the druggable molecular targets.

2 https://www.ebi.ac.uk/gxa/home
3 https://dice-database.org/
4 https://platform.opentargets.org/

2.5.4. System biology of innate immunity

The innate immunity interactions for the query genes were further explored by using the InnateDB website.[5] This publicly available database with an integrated platform facilitates the systems-level analysis of innate immunity networks, pathways, and genes (20). The input is the gene name, and the output is the interactions and signaling responses involved in innate immunity processes.

2.6. Interaction gene networks and function prediction

The GeneMANIA plugin from Cytoscape was used to identify gene interaction networks from query genes and predict the gene's putative function and annotation. The plugin uses a large database of functional interaction networks from *Homo sapiens*, and each related gene is traceable to the source network used to make the prediction. The input is the query gene list and the organism type. The output is a network of interconnected genes (21, 22).

2.7. Amino acid conserved domains

The functional relevance of rare genetic variants on candidate proteins was predicted by comparing the nucleotide and amino acid sequences to the functional domains of the concerned protein as listed in the Conserved Domain Database (CDD). CDD program uses RPS-BLAST, which efficiently scans the query protein for pre-computed position-specific score matrices (PSSMs), to estimate the sequence conservation characteristics of the functional domains of the candidate protein. Protein domains annotated with query input sequence and imagining options are shown in the output file.

2.8. Protein structure analysis

2.8.1. Protein modeling and stability analysis

The Artificial Intelligence (AI) program developed by Alphabet/Google DeepMind, AlphaFold, generated protein structure at the molecular level[6], which was extensively used to study the structural effect of the variants on the candidate proteins. The input is the protein, gene name, or UniProt accession, and organism name. The output is a predicted 3D protein model from its amino acid sequence with high accuracy (including side chains), a per residue confidence metric (PLDDT) that is used to color the residues of the prediction, and a predicted aligned error that is necessary to assess confidence in the domain packing and large-scale topology of the protein. The I-TASSER web tool was also used along with AlphaFold for the generation of protein structures that were not available in AlphaFold. I-TASSER predicts the 3D structure and biological activity of protein molecules based on their amino acid sequences using high-quality model predictions. The input is the amino acid sequence, and the output is several full-length atomic models along with their estimated accuracy (including a confidence score for all models, predicted TM-score, and RMSD for the first model), GIF images of the predicted models, and predicted secondary structure and solvent accessibility. To generate mutant protein models, SWISS-MODEL, a fully automated protein structure homology-modeling tool, was used. The input is the mutated amino acid sequence along with the wild-type template file in PDB format. Outputs include the 3D structure of models, their target–template sequence alignment, and model coordinates. The protein model PDB file is viewed by a molecular visualization system, PyMOL 2.5.[7] PyMOL represents the protein in a three-dimensional (3D) model and is capable of editing molecules.

2.8.2. Structural deviation and stability findings

The structural deviation between optimized native and variant protein models was determined using YASARA, a molecular graphics, modeling, and simulation tool. Two protein atomic coordinates were superimposed on top of each other, and the corresponding RMSD values were calculated to quantify structural similarity at both the global and local residue levels. The cut-off RMSD values for variant-induced structure deviations at the polypeptide chain and residue levels were > 0.2 and > 2, respectively. The effect of a candidate variant on protein structure stability was determined using the MAESTRO webserver. MAESTRO provides a confidence estimation C_{pred} for its total predicted change in stability (kcal/mol) ΔΔG predictions. ΔΔGpred <0.0 indicates a stabilizing mutation and C_{pred} is given as a value between 0.0 (not reliable) and 1.0 (highly reliable).

3. Results

3.1. Clinical and family history

In family A (Figure 1A), the proband (III.2), aged 27 years, is the offspring of a consanguineous marriage between first cousins and has no family history of inflammatory bowel disease. He was first diagnosed with Crohn's disease at the age of 22 years, suffering from several symptoms such as nausea, anorexia, and night sweats. His endoscopy test findings confirmed the diagnosis of Crohn's disease with an eroded, punctate, white-spotted mucosa in the esophagus and a hemorrhagic gastropathy (Figure 2). He is currently being treated with Pentaza (mesalamine), which is a 5-aminosalicylic acid derivative, and Imuran (azathioprine AZA), an immunosuppressive medication. His younger brother (III.3), now 20 years old, was first diagnosed with Crohn's when he was 17 years old. He had comparably severe symptoms including lethargy, dizziness, and anorexia. At first, he was diagnosed with tuberculosis and was treated for 9 months. After that, gastrointestinal inflammation recurrence was noticed when a confirmatory endoscopy test was performed. The endoscopic findings were a tight, inflamed terminal ileum and an enterocolonic fistula. Then, after 2 years, a second endoscopic test was done that found severe inflammation at the ascending colon and cecum with anatomical distortion characterized by altered vascularity, congestion (edema), erythema, and pseudopolyps. The findings were worse when

5 https://www.innatedb.com/
6 https://alphafold.ebi.ac.uk/
7 https://pymol.org/2/

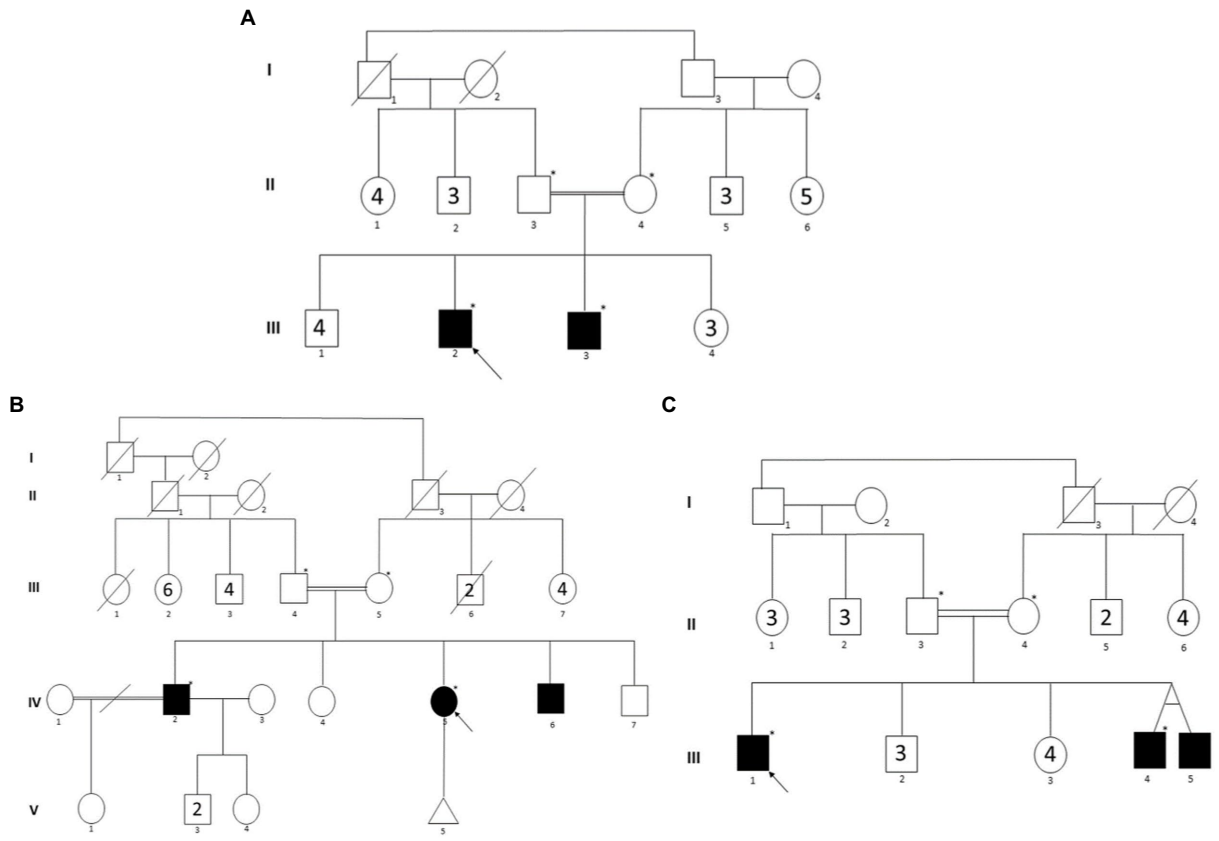


FIGURE 1
Three-generation pedigrees of families with IBD. **(A–C)** Represent the pedigrees of families A, B, and C, respectively. The black color circles or boxes represent patients with IBD. The arrow represents the proband in each family. The star represents individuals selected for WES.

compared to previous examinations (Figure 2). He is currently being administered Remicade (Infliximab), which is a chimeric monoclonal antibody used to treat several autoimmune diseases, including IBD.

In family B (Figure 1B), the parents are healthy distant relatives from the same Arabian tribe. In this family, Crohn's disease was diagnosed in one female and two male siblings. The proband (IV.2) was diagnosed when he was 25 years old and is currently taking Humira (monoclonal antibody). His sister (IV.5) was diagnosed in her late 20s, and she had a colectomy and an ostomy bag. His younger brother (IV.6) was diagnosed when he was 25 years old and was kept on Infliximab (Remicade) for 2 years.

In family C (Figure 1C), the parents are first cousins and healthy, except that the mother has some intestinal inflammation. Interestingly, of the three affected male siblings, two were monozygotic twins. The proband (III.1) was diagnosed 3 years ago, and he is 32 years old now. He has been on Remicade monoclonal antibody treatment every 2 months since the diagnosis. Both twins (III.4 and III.5), now aged 29 years, were diagnosed 6 years ago, and both underwent colectomy at the ages of 26 and 24 years, respectively.

3.2. Whole exome sequence analysis

Whole exome sequencing of many family members provided an average of 97,242, 98,011, and 96,297 variants in families A, B, and C, respectively. These massive numbers of variants were further filtered out by excluding 3′ and 5′ UTR variants, conservative and disruptive inframe deletion or insertion, synonymous, intergenic, and intronic variants, coding variants with high allele frequency (>0.01), and poor quality variants with a Phred score of <30. The inclusion of rare coding variants has resulted in 3,498 variants (in 1,455 genes) for family A, 3,721 variants (in 1,571 genes) for family B, and 3,679 variants (in 1,668 genes) for family C. Most of the coding variants in all three families were of the missense type (Table 1). The segregation analysis of the variants in the respective families with IBD did not detect any single rare variant following a classical AR, CH (compound heterozygotes), or *de novo* inheritance pattern. Therefore, we searched for the aggregation of rare variants that would increase the burden status of genes in these families.

3.3. Functional enrichment analysis using immune pathways

The functional enrichment analysis of rare variant genes from individual families revealed a total of 180, 114, and 116 immune-related pathways for families A, B, and C, respectively. In families A, B, and C, 23, 21, and 29 immune pathways respectively were significantly enriched ($p = <0.05$).

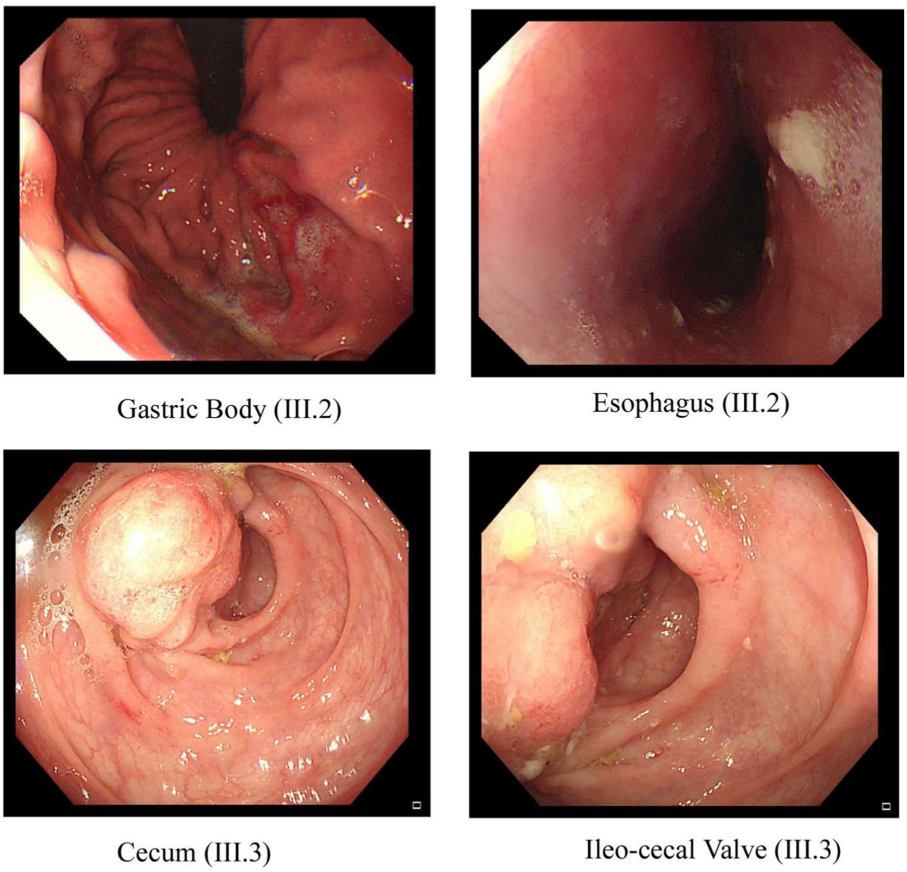


FIGURE 2
Endoscopic images of the GI tract of the proband III.2 and affected sibling III.3 in family A. Two pictures of the top row from proband III.2 show inflammation and ulceration lesions in the gastric body and eroded, punctate white spotted mucosa in the esophagus with a hemorrhagic gastropathy. Bottom row images from III.3 show inflammation and ulceration lesions in the cecum and ileocecal valve.

TABLE 1 The exome variants yield from siblings of three IBD families.

Case		Total variants	Coding*	Rare**	Number of genes	Homozygous variants	Heterozygous variants
Family A	III.2	98,823	13,300	1721	734	195	1,526
	III.3	95,660	13,157	1777	721	173	1,604
Family B	IV.2	97,925	13,287	1809	785	122	1,687
	IV.5	98,097	13,435	1912	786	150	1762
Family C	III.1	97,737	12,933	1862	833	148	1714
	III.4	94,857	12,796	1817	835	149	1,668

Coding* includes Frameshift, missense, splice acceptor, splice donor, start lost, and stop retained variants.
Rare** (Minor allele frequency < 0.01 in gnomAD, 1,000 genomes, ExAC dbs).

Table 2 presents the top five significant immune pathways for each family.

A total of 95 (61.3%) GO terms were shared by the three families and analyzed with the VENNY tool. These GO terms were associated with 163 genes after excluding the human leukocyte antigen (*HLA*) complex gene family owing to their known involvement in multiple autoimmune diseases. When we analyzed all 163 genes, only eight genes with rare variants were found to be common among all three families (Figures 3A,B).

3.4. Transcript expression analysis of candidate genes in IBD and healthy tissue samples

Out of the eight rare variant genes, seven genes were differentially expressed in colonic and mucosal tissues. Of them, two (*ZDHHC11* and *PRSS3*) were downregulated (FC: <−1.1) and five (*LILRB3, LILRA2, LILRB1, PRSS2,* and *LILRA1*) were upregulated. The expression of the upregulated genes (FC: >1.1) is presented in

10.3389/fmed.2023.1164305

TABLE 2 Top five immune system-related pathways enriched in genes with rare coding variants in three families with IBD.

ID	Term	*P*-value*	% Associated genes	Associated genes found
Family A				
GO:0002483	Antigen processing and presentation of endogenous peptide antigen	0.00	38.10	*LEF1, LIG4, PRKDC*
GO:0002697	Regulation of immune effector process	0.01	3.02	*LEF1, LIG4, PRKDC*
GO:0038093	Fc receptor signaling pathway	0.01	2.30	*LILRB1, TMEM176A, TMEM176B*
GO:0045088	Regulation of innate immune response	0.01	8.98	*LIG4, PRKDC, SOS1, SOS2*
GO:0002220	Innate immune response activating cell surface receptor signaling pathway	0.02	11.20	*ERAP1, ERAP2, IDE, SEC14L3*
Family B				
GO:0002250	Adaptive immune response	0.00	2.88	*AHR, BTNL9, CARD9, CD79A, CEACAM1, HLA-B, HLA-C, HLA-DQB1, HLA-DQB2, HLA-DRB1, HLA-DRB5, IL17RA, IRF7, LILRA1, LILRB1, LILRB3, ORAI1, OTUD7B, PDCD1LG2, PPL, RAPGEF3, RASGRP1, RIF1*
GO:0045088	Regulation of innate immune response	0.00	9.58	*A2M, CARD9, CEACAM1, DHX58, HLA-B, IKBKB, IL18RAP, IRF7, KIR2DL4, LILRA2, LILRB1, MUC12, MUC16, MUC17, MUC19, MUC2, MUC20, MUC3A, MUC4, MUC5AC, MUC6, NCR1, NLRC5, OTOP1, PIK3R6, PRKDC, PSMB11, PSME3, PSPC1, RASGRP1, SOCS1, TRIM5*
GO:0045088	Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.00	2.40	*AHR, CARD9, CEACAM1, HLA-B, HLA-DQB1, HLA-DQB2, HLA-DRB1, IRF7, LILRB1, RIF1*
GO:0042269	Adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.00	18.60	*CEACAM1, HLA-B, IL18RAP, KIR2DL4, LILRB1, NCR1, PIK3R6, RASGRP1*
GO:0002218	Activation of innate immune response	0.00	11.39	*CARD9, IKBKB, LILRA2, MUC12, MUC16, MUC17, MUC19, MUC2, MUC20, MUC3A, MUC4, MUC5AC, MUC6, PRKDC, PSMB11, PSME3, PSPC1, TRIM5*
Family C				
GO:0002220	Innate immune response activating cell surface receptor signaling pathway	0.00	14.40	*CARD9, ICAM3, KLRC2, LILRA2, MUC1, MUC12, MUC16, MUC17, MUC19, MUC2, MUC20, MUC21, MUC3A, MUC4, MUC5AC, MUC5B, MUC6, PSMA8*
GO:0002758	Innate immune response-activating signal transduction	0.00	14.29	*CARD9, ICAM3, KLRC2, LILRA2, MUC1, MUC12, MUC16, MUC17, MUC19, MUC2, MUC20, MUC21, MUC3A, MUC4, MUC5AC, MUC5B, MUC6, PSMA8*
GO:0002223	Stimulatory C-type lectin receptor signaling pathway	0.00	14.05	*CARD9, ICAM3, KLRC2, MUC1, MUC12, MUC16, MUC17, MUC19, MUC2, MUC20, MUC21, MUC3A, MUC4, MUC5AC, MUC5B, MUC6, PSMA8*
GO:0002218	Activation of innate immune response	0.00	12.66	*CARD9, CGAS, ICAM3, KLRC2, LILRA2, MUC1, MUC12, MUC16, MUC17, MUC19, MUC2, MUC20, MUC21, MUC3A, MUC4, MUC5AC, MUC5B, MUC6, PSMA8, PSPC1*
GO:0042113	B cell activation	0.00	2.72	*ATM, HLA-DQB1, HLA-DQB2, IGLC1, IGLL5, LFNG, MSH2, SAMSN1, SLC25A5, YY1AP1*

P-value* significant (<0.05).

Figure 4B. In the control tissue (gastrointestinal, blood, and immune organs) samples, seven (of the eight) shared genes showed differential expression. *PRSS3* has high expression in the small intestine and colon (Figure 3C).

3.5. Immune cell gene expression

Based on RNA sequencing data, we investigated the immune cell type representations of the eight prioritized genes, and only

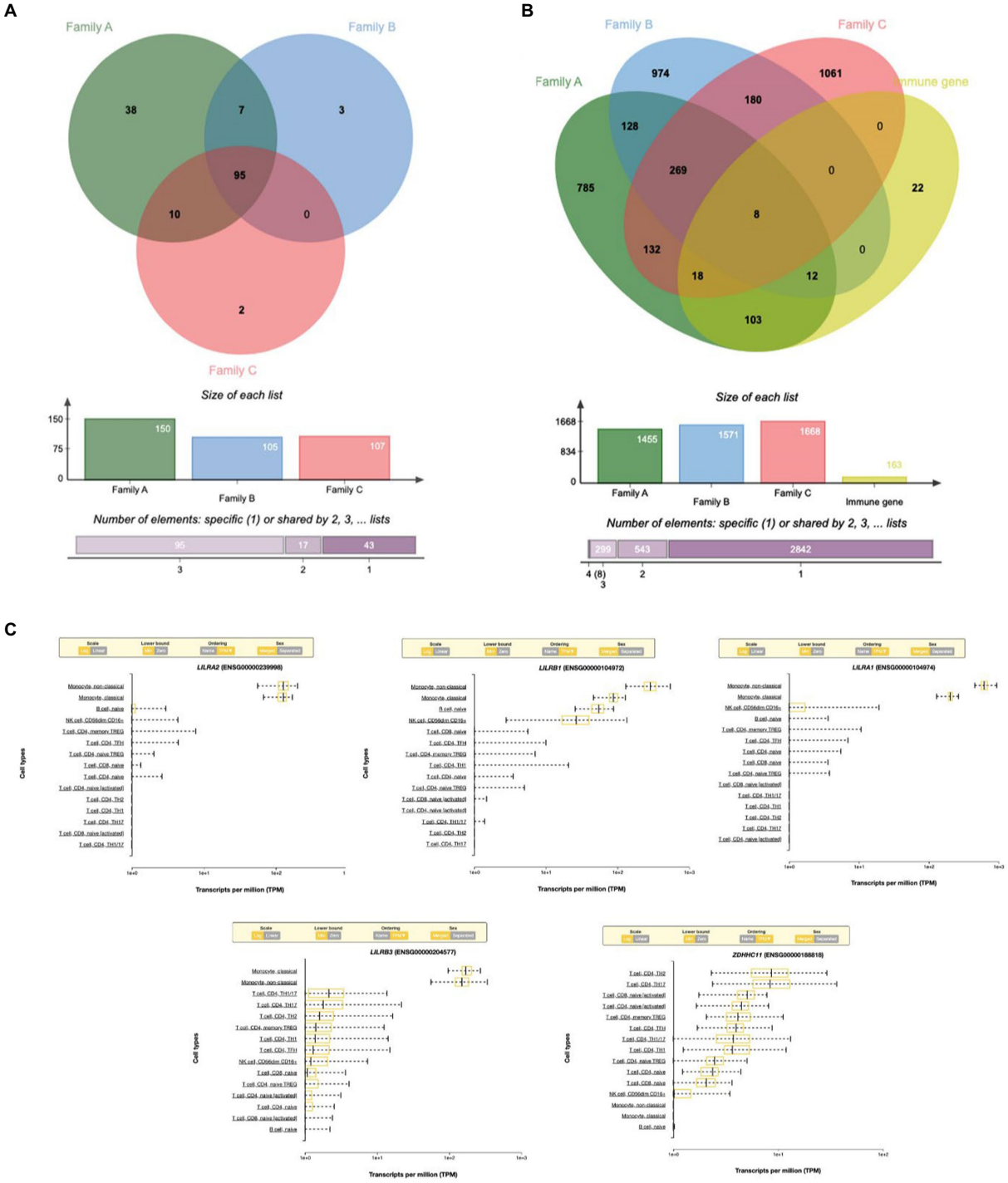


FIGURE 3
(A) Venn diagram showing the shared immune Go-terms between three IBD families. **(B)** Venn diagram showing the shared gene with rare variants between three families. **(C)** Top immune cell gene expression patterns of the genes with rare variants.

seven genes had significant expression in various immune cells with the log FC of 0.4 (Figure 4A). Leukocyte immunoglobulin-like receptor genes (*LILRB1, LILRB3, LILRA1, and LILRA2*) are highly expressed in immune cells such as monocytes and natural killer (NK) cells. The *LILRB3* gene is highly expressed in classical and non-classical monocytes with an average of 175.84 TPM and 152.3 TPM, respectively. However, this gene is barely expressed in the T cells, with a mean TPM of <2.73. Furthermore, *LILRB1, LILRA1*, and *LILRA2* are highly enriched in non-classical monocytes with means of 290.46, 632.40, and 128.35, respectively, compared to the classical monocytes with an average of 87.91, 204.29, and 123.98, respectively. *KMT2C* is also highly expressed in non-classical monocytes, with a mean of more than 90 TPM (Table 3).

FIGURE 4
(A) Eight genes that are highly expressed in the gastrointestinal tract and several immune tissues. **(B)** Differentially expressed genes in IBD datasets from NCBI-GEO database.

3.6. Open target phenotype identification

From the eight genes identified from the WES rare variant burden analysis, only four genes have shown an association score of >0.1 with gastrointestinal or immune system disease phenotypes. The *KMT2C* gene shared phenotypes with UC with an overall association score of >0.37 (Table 4).

3.7. Concordance analysis

We used the Venny tool to find genes that were present in both IBD and normal healthy tissue expression analyses, immune cell restricted expression analysis, and open target platform analysis. Of the eight genes, seven (90%) were expressed in IBD tissues, normal healthy tissues (GI, immune organs), and different immune cell types

10.3389/fmed.2023.1164305

such as monocytes and NK cells. In addition, four genes (50%) showed a strong association (>0.1 score) with gastrointestinal and immune system disease phenotypes. However, all eight genes, *LILRB1, LILRB3, LILRA2, LILRA1, KMT2C, ZDHHC11, PRSS2,* and *PRSS3*, were found to be significant in at least two tools (Table 3).

3.8. System biology analysis of innate immunity

Only *LILRB1* and *PRSS3* have physical interactions or associations with the innate immune response in humans, out of the eight genes obtained in the preceding step. *LILRB1* is mapped to chromosome 19, and it has 12 experimentally validated interactions with other genes. Most of the interacting partners are from the *HLA* gene family, such as A, C, G, and F. The gene *PRSS3* interacts with six other genes (Table 4).

3.9. Shared genes with rare variants to pathway analysis

These three families shared 10 rare variants for *LILRB1* (ENST00000324602.12) including six missense and four novel frameshift variants. However, 11 unique missense variants were shared only between families B and C. Furthermore, two unique missense variants each were observed in families B and C. Families A and C shared a missense variant in *PRSS3* (ENST00000379405.4) (c.244G > A; rs76740888) and family B had one additional missense variant in *PRSS3* (c.10 T > C; rs772714741) (Table 5).

These two genes, *LILRB1* and *PRSS3*, were studied independently to map the biochemical pathways associated with them. Our findings showed that *LILRB1* is connected to three pathways, namely, the adaptive immune system, the immune system, and immunoregulatory interactions between a lymphoid and a non-lymphoid cell. The *PRSS3* gene is involved in 10 different pathways, namely, neutrophil degranulation, the innate immune system, antimicrobial peptides, the metabolism of vitamins and cofactors, the metabolism of water-soluble vitamins and cofactors, alpha-defensins, defensins, the immune system, and cobalamin (Cbl, vitamin B12) transport and metabolism.

3.10. Gene–gene networking analysis

Many of the physically interacting partners of *PRSS3* (such as *TCN1, DEFA4, DEFA1, DEFA5, DEFA3, DEFA6, PRSS2, SPINK1,* and *CBLIF*) are co-regulated and co-expressed with interacting partners of *LILRB1* (*HLA-B, LILRA1,* and *LILRA3*). Indirect dysregulated interactions between many of these proteins might trigger inflammation in IBD (Table 6).

3.11. Amino acid conserved domains

A crucial step in determining the relationship between the nucleotide sequence, protein structure, and function of disease-causing proteins is by mapping the conserved amino acid domains. According to the CDD analysis, the *LILRB1* protein contains an immunoglobulin (Ig) superfamily domain located between 28 and 419 amino acid positions (four domains). *PRSS3* protein consists of a Trypsin-like serine protease domain between 38 and 256 amino acids. We excluded variants that were located outside the conserved domains area (Table 7).

TABLE 3 Summary of the four different computational predictions for potential genes for IBD pathology: normal expression, IBD specific expression, and immune and open target platform.

Gene name	Normal expression (Colon) (average TPM)	IBD specific expression (FC)	Immune (Mean TPM)	Open target platform overall association score
LILRB1	2.32	1.25	290.46	0.145
LILRB3	1.52	1.825	175.84	<0.1
KMT2C	14.6	NA	90.96	0.264
ZDHHC11	0.1	−1.1	10.26	<0.1
LILRA2	0.3	1.85	128.35	<0.1
PRSS2	0	2.23	NA	0.156
PRSS3	45.2	−1.5	0.82	0.109
LILRA1	0.34	1.575	632.40	<0.1

TABLE 4 Number of experimentally validated interactions and predicted interactions for *LILRB1* and *PRSS3* genes from the innate immunity database.

Ensembl gene ID	Organism	Chromosome	Gene symbol	Gene name	Experimentally validated interactions	Interactions predicted by orthology
ENSG00000104972	Homo sapiens	19	*LILRB1*	Leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 1	12	0
ENSG00000010438	Homo sapiens	9	*PRSS3*	Protease, serine, 3	6	0

10.3389/fmed.2023.1164305

TABLE 5 Rare variants of *LILRB1* and *PRSS3* genes in three families.

Gene name	Chr. No.	Position	Rs ID	cDNA position	Amino acid position	Effect	MAF	
							1,000 Gp	GenomAD
Family A								
LILRB1	19	54,631,583	rs554096090	c.154G > A	p.Gly52Ser	Missense variant	0.001	0.000
LILRB1	19	54,631,587	rs199588814	c.158A > T	p.Gln53Leu	Missense variant	0.001	0.000
LILRB1	19	54,631,686	rs200880414	c.257C > T	p.Pro86Leu	Missense variant	0.000	0.000
LILRB1	19	54,631,724	rs570016342	c.295 T > A	p.Tyr99Asn	Missense variant	0.000	0.000
LILRB1	19	54,631,725	rs535742370	c.296A > T	p.Tyr99Phe	Missense variant	0.000	0.000
LILRB1	19	54,631,749	rs142396802	c.320G > T	p.Arg107Leu	Missense variant	0.000	0.000
LILRB1	19	54,633,154	–	c.1098_1099delAT	p.Trp367fs	Frameshift variant	–	–
LILRB1	19	54,633,157	–	c.1100_1101insCT	p.Trp367fs	Frameshift variant	–	–
LILRB1	19	54,633,171	–	c.1114_1115insAG	p.Thr372fs	Frameshift variant	–	–
LILRB1	19	54,633,173	–	c.1117_1118delTA	p.Tyr373fs	Frameshift variant	–	–
PRSS3	9	33,796,675	rs76740888	c.244G > A	p.Val82Ile	Missense variant	–	–
Family B								
LILRB1	19	54,631,583	rs554096090	c.154G > A	p.Gly52Ser	Missense variant	0.001	0.000
LILRB1	19	54,631,587	rs199588814	c.158A > T	p.Gln53Leu	Missense variant	0.001	0.000
LILRB1	19	54,631,605	rs774715846	c.176G > A	p.Arg59His	Missense variant	–	0.000
LILRB1	19	54,631,686	rs200880414	c.257C > T	p.Pro86Leu	Missense variant	0.000	0.000
LILRB1	19	54,631,724	rs570016342	c.295 T > A	p.Tyr99Asn	Missense variant	0.000	0.000
LILRB1	19	54,631,725	rs535742370	c.296A > T	p.Tyr99Phe	Missense variant	0.000	0.000
LILRB1	19	54,631,749	rs142396802	c.320G > T	p.Arg107Leu	Missense variant	0.000	0.000
LILRB1	19	54,631,944	rs370374304	c.368 T > G	p.Ile123Ser	Missense variant	0.000	0.000
LILRB1	19	54,633,033	rs1185911260	c.976G > C	p.Val326Leu	Missense variant	–	–
LILRB1	19	54,633,034	rs1486166961	c.977 T > C	p.Val326Ala	Missense variant	–	–
LILRB1	19	54,633,037	rs974205214	c.980C > T	p.Ser327Phe	Missense variant	–	0.000
LILRB1	19	54,633,049	rs1334566399	c.992A > G	p.Gln331Arg	Missense variant	–	–
LILRB1	19	54,633,108	rs765206177	c.1051 T > G	p.Trp351Gly	Missense variant	–	0.000
LILRB1	19	54,633,116	rs764221410	c.1059A > C	p.Gln353His	Missense variant	–	0.000
LILRB1	19	54,633,150	rs1260040283	c.1093G > T	p.Asp365Tyr	Missense variant	–	–
LILRB1	19	54,633,151	rs12985933	c.1094A > C	p.Asp365Ala	Missense variant	–	–
LILRB1	19	54,633,154	–	c.1098_1099delAT	p.Trp367fs	Frameshift variant	–	–
LILRB1	19	54,633,157	–	c.1100_1101insCT	p.Trp367fs	Frameshift variant	–	–
LILRB1	19	54,633,166	rs1401913528	c.1109G > A	p.Arg370Lys	Missense variant	–	–
LILRB1	19	54,633,171	–	c.1114_1115insAG	p.Thr372fs	Frameshift variant	–	–
LILRB1	19	54,633,173	–	c.1117_1118delTA	p.Tyr373fs	Frameshift variant	–	–
LILRB1	19	54,633,185	rs1240220003	c.1128A > T	p.Gln376His	Missense variant	–	–
LILRB1	19	54,633,210	rs372567136	c.1153G > A	p.Gly385Ser	Missense variant	–	0.000
PRSS3	9	33,795,583	rs772714741	c.10 T > C	p.Phe4Leu	Missense variant	–	–
Family C								
LILRB1	19	54,631,583	rs554096090	c.154G > A	p.Gly52Ser	Missense variant	0.001	0.000
LILRB1	19	54,631,587	rs199588814	c.158A > T	p.Gln53Leu	Missense variant	0.001	0.000
LILRB1	19	54,631,605	rs774715846	c.176G > A	p.Arg59His	Missense variant	–	0.000
LILRB1	19	54,631,686	rs200880414	c.257C > T	p.Pro86Leu	Missense variant	0.000	0.000
LILRB1	19	54,631,724	rs570016342	c.295 T > A	p.Tyr99Asn	Missense variant	0.000	0.000

(Continued)

10.3389/fmed.2023.1164305

TABLE 5 (Continued)

Gene name	Chr. No.	Position	Rs ID	cDNA position	Amino acid position	Effect	MAF	
							1,000 Gp	GenomAD
LILRB1	19	54,631,725	rs535742370	c.296A > T	p.Tyr99Phe	Missense variant	0.000	0.000
LILRB1	19	54,631,749	rs142396802	c.320G > T	p.Arg107Leu	Missense variant	0.000	0.000
LILRB1	19	54,631,944	rs370374304	c.368 T > G	p.Ile123Ser	Missense variant	0.000	0.000
LILRB1	19	54,631,965	rs767704704	c.389A > T	p.Gln130Leu	Missense variant	–	0.000
LILRB1	19	54,633,037	rs974205214	c.980C > T	p.Ser327Phe	Missense variant	–	0.000
LILRB1	19	54,633,049	rs1334566399	c.992A > G	p.Gln331Arg	Missense variant	–	–
LILRB1	19	54,633,108	rs765206177	c.1051 T > G	p.Trp351Gly	Missense variant	–	0.000
LILRB1	19	54,633,116	rs764221410	c.1059A > C	p.Gln353His	Missense variant	–	0.000
LILRB1	19	54,633,150	rs1260040283	c.1093G > T	p.Asp365Tyr	Missense variant	–	–
LILRB1	19	54,633,151	rs12985933	c.1094A > C	p.Asp365Ala	Missense variant	–	–
LILRB1	19	54,633,154	–	c.1098_1099delAT	p.Trp367fs	Frameshift variant	–	–
LILRB1	19	54,633,157	–	c.1100_1101insCT	p.Trp367fs	Frameshift variant	–	–
LILRB1	19	54,633,166	rs1401913528	c.1109G > A	p.Arg370Lys	Missense variant	–	–
LILRB1	19	54,633,171	–	c.1114_1115insAG	p.Thr372fs	Frameshift variant	–	–
LILRB1	19	54,633,173	–	c.1117_1118delTA	p.Tyr373fs	Frameshift variant	–	–
LILRB1	19	54,633,185	rs1240220003	c.1128A > T	p.Gln376His	Missense variant	–	–
LILRB1	19	54,633,210	rs372567136	c.1153G > A	p.Gly385Ser	Missense variant	–	0.000
LILRB1	19	54,636,536	rs41308744	c.1696G > A	p.Glu566Lys	Missense variant	0.004	0.000
PRSS3	9	33,796,675	rs76740888	c.244G > A	p.Val82Ile	Missense variant	–	–

TABLE 6 Protein–protein interactions of *LILRB1* and *PRSS3* genes.

Interactors	Species	Type	Source database ID(s)	Interactor types	Tissue
LILRB1 with HLA-B	Homo sapiens	Physical interaction	BIOGRID-256234	Protein – protein	–
LILRB1 with HLA-A	Homo sapiens	Association	IDB-120686	Protein – protein	Kidney cell line
CSK with LILRB1	Homo sapiens	Association	MINT-8027327; EBI-7351403	Protein – protein	–
HLA-F with LILRB1	Homo sapiens	Physical interaction	BIOGRID-276645	Protein – protein	–
LILRB1 with HLA-A	Homo sapiens	Association	BIOGRID-255783	Protein – protein	–
PTPN6 with LILRB1	Homo sapiens	Physical association	IDB-190120; BIOGRID-318101	Protein – protein	–
CSK with LILRB1	Homo sapiens	Physical interaction	IDB-117837; IDB-117834	Protein – protein	T-lymphocyte cell line
LILRB1 with HLA-C	Homo sapiens	Physical interaction	BIOGRID-256235	Protein – protein	–
LILRB1 with HLA-G	Homo sapiens	Physical interaction	BIOGRID-256236; MINT-7144982; EBI-7087620	Protein – protein	–
PTPN6 with LILRB1	Homo sapiens	Physical interaction	IDB-117838; IDB-117836	Protein – protein	T-lymphocyte cell line
B2M with LILRB1	Homo sapiens	Association	BIND-121495	Protein – protein	–
CSK with LILRB1	Homo sapiens	Association	EBI-7351451; MINT-8027342	Protein – protein	–
PRSS3 with SERPINA1	Homo sapiens	Association	BIND-117882; BIND-90568	Protein – protein	–
Complex of 10 interactors	Homo sapiens	Association	EBI-8770525	Protein – protein	–
PRSS3 with ALB	Homo sapiens	Association	BIOGRID-825632	Protein – protein	–
PRSS3 with HDGF	Homo sapiens	Association	BIOGRID-635705	Protein – protein	–
TFPI with PRSS3	Homo sapiens	Association	BIOGRID-317015	Protein – protein	–
PRSS3 with UBC	Homo sapiens	Association	BIOGRID-627754; BIOGRID-618329	Protein – protein	–

3.12. *LILRB1* and *PRSS3* 3D model construction

The predicted 3D protein structure was collected from Alphafold, the state-of-the-art AI system developed by DeepMind, and I-TASSER. The total length (650 aa) of the structures of human *LILRB1* protein chain A, with model confidence (pLDDT >70), was downloaded as a PDB file. The full-length (247 aa) structure model of human *PRSS3* protein chain A was downloaded as a PDB file, with a model confidence score (C-score of −0.54), an estimated TM score of 0.64 ± 0.13 Å, and an estimated RMSD = 6.9 ± 4.1 Å. The *LILRB1* (p.Gln53Leu, p.Tyr99Asn, p.Trp351Gly, p.Asp365Ala, and p.Gln376His) and *PRSS3* (p.Phe4Leu and p.Val25Ile) were then modeled using homology modeling by the SWISS-MODEL using energy-minimized native protein structures.

3.13. Protein stability analysis

Pathogenic amino acid substitutions can result in changes in free energy values, thereby directly impacting protein stability. We analyzed the impact of 16 *LILRB1* (G52S, Q53L, R59H, P86L, Y99N, Y99F, R107L, Q130L, S327F, Q331R, W351G, Q353H, D365Y, D365A, Q376H, and 2 G385S) and *PRSS3* (F4L and V25I) variants on protein stability by MAESTRO. MAESTRO is a robust tool for predicting stability changes following point mutations by providing predicted free energy change (ΔΔG) values and a corresponding prediction confidence estimation (c_{pred}). For the *LILRB1* protein, out of the 16 variants, only five had a destabilizing effect on the protein (Q53L, Y99N, W351G, D365Y, and D365A) with ΔΔG of 0.074, 0.85, 0.380, 0.083, and 0.086, respectively. The c_{pred} scores were 0.088, 0.923, 0.875, 0.885, and 0.872, respectively. The two *PRSS3* (F4L and V25I) variants had a destabilizing effect with ΔΔG of 0.016 and 0.799 and c_{pred} of 0.885 and 0.857 (Table 8). We used the YASARA tool to analyze the native and mutant *LILRB1* and *PRSS3* structures to evaluate their structural drifts (in terms of RMSD at residue and whole protein levels). The RMSD value is utilized to quantify the structural similarity between two atomic coordinates when they are superimposed. When there is divergence at the polypeptide chain level, impact of substitution mutations on amino acid structures can be determined. For the *LILRB1* protein, the five substitutions with destabilizing effects on the protein (Q53L, Y99N, W351G, D365Y, and D365A) had RMSDs at residue levels of 1.8395, 2.0688, 1.5186, 2.0098, and 2.0351. The two *PRSS3* (F4L and V25I) variants with destabilizing effects had RMSD at residue levels of 2.1465 and 2.0270, respectively (Figure 5 and Table 9).

4. Discussion

Most genetic studies on IBD have largely concentrated on identifying common variants with small effect sizes through GWAS studies (9). However, rare and highly penetrant variations identified through population-specific cohorts or family-focused research have immense potential to catch the variants with high effect size on complex diseases such as IBD (8, 12). Although, studying the familial cases may uncover rare causal variants, their heritability of disease in unrelated patient cohorts is still uncertain (23). Unlike VEO-IBD, which has a causal monogenic factor, late-onset is a complex and multifactorial disorder that cannot be explained by classical genetic segregation methods (16, 24, 25). Large-scale sporadic case–control studies on WES-based rare variant burden analysis (RVB) have previously identified several strong risk loci for complex diseases, such as Schizophrenia (26), Alzheimer's disease (27), epilepsy (2019), autism (28), and Crohn's disease (12).

According to a recent systematic review and meta-analysis of IBD in the Arab World, the consanguinity rate in Saudi Arabian IBD patients is as high as 32.6% (4). Consanguinity acts as a prerequisite risk factor for several autosomal recessive immune disorders (29, 30). Therefore, identifying the actual genetic causes underlying familial IBD is expected to aid in early detection, therapy optimization, carrier screening, and genetic counseling for extended families. In this context, we have sequenced the exomes of three consanguineous Saudi families with more than one IBD-affected sibling. We performed segregation analyses of the variants in the respective IBD families. However, this did not result in identifying any causal rare variant fitting into the classical autosomal recessive, compound heterozygote, or *de novo* inheritance patterns. Since the classical Mendelian segregation analysis does not apply to all forms of IBD, single-gene models often fail to explain the complex molecular etiology of the disease. For example, in other gastrointestinal diseases, such as celiac disease (CeD), a recent study of two rare Arab families with CeD concluded that the genetic variability cannot be explained by classical genetic segregation techniques, because the single gene model is incapable of dissecting the disease's molecular elements (24). It has adopted multidimensional computational analysis to identify and characterize the potential autoimmunity risk genes for Celiac disease (19). Therefore, following a similar strategy, we searched and identified potential IBD genes based on the rare variant burden analysis using a combination of artificial intelligence approaches, bioinformatic tools, and multi-dimensional, large-scale next-generation sequence datasets. This novel approach at a large scale is likely to provide some valuable clues to novel biomarkers or drug targets for many complex diseases in the future (24, 31–34).

We prioritized from thousands of rare variants of WES to potential two candidate genes, *LILRB1* and *PRSS3*, owing to their strong involvement in the innate immune system. Both genes are linked to inflammation, a process in which multiple pathways interact to contribute to this complex function. The *LILRB1* gene is a member of the leukocyte immunoglobulin-like receptor (*LILRs*; or *ILT, LIR,* and *CD85*) family, which are the most conserved genes located within the leukocyte receptor cluster on human chromosome 19 (35, 36). The family consists of 13 members with activating or inhibitory properties: *LILRs* with long cytoplasmic tails that contain inhibitory motifs based on tyrosine act as inhibitory receptors (*LILRBs*), whereas *LILRs* with short cytoplasmic tails act as activators (*LILRAs*). *LILRs* are two pseudogenes and 11 functional genes encoding five activating (*LILRA1, 2, 4–6*), five inhibitory (*LILRB1–5*), and one soluble form (*LILRA3*). Moreover, *LILRs* are classified into two classes based on the amino acid sequence similarity of the region that binds to *HLA*. *LILRB1, B2, A1, A2,* and *A3* are classified as members of group 1 that are highly similar in sequence and are likely to interact with *HLA* class I molecules (*HLAIs*). Furthermore, *LILRB1* has been shown to inhibit the combination of *CD8* and *HLA I* molecules, hence regulating *CD8+* T cells (37, 38).

From our results, we found that the three families shared 10 rare variants (six missense and four novel frameshift variants) in the *LILRB1* gene. However, 11 unique missense variants were shared only between families B and C. Furthermore, two unique missense variants were shown in families B and C, respectively. Of

10.3389/fmed.2023.1164305

TABLE 7 Conserved domains and their amino acid locations in *LILRB1* and *PRSS3*.

Gene	cDNA position	Amino acid location	Domain	Domain range
LILRB1	c.154G > A	p.Gly52Ser	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.158A > T	p.Gln53Leu	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.176G > A	p.Arg59His	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.257C > T	p.Pro86Leu	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.295 T > A	p.Tyr99Asn	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.296A > T	p.Tyr99Phe	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.320G > T	p.Arg107Leu	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.389A > T	p.Gln130Leu	IgC2_D1_LILR_KIR_like	28–118
LILRB1	c.1098_1099delAT	p.Trp367fs	Ig super family	327–419
LILRB1	c.1100_1101insCT	p.Trp367fs	Ig super family	327–419
LILRB1	c.1114_1115insAG	p.Thr372fs	Ig super family	327–419
LILRB1	c.980C > T	p.Ser327Phe	Ig super family	327–419
LILRB1	c.992A > G	p.Gln331Arg	Ig super family	327–419
LILRB1	c.1051 T > G	p.Trp351Gly	Ig super family	327–419
LILRB1	c.1059A > C	p.Gln353His	Ig super family	327–419
LILRB1	c.1093G > T	p.Asp365Tyr	Ig super family	327–419
LILRB1	c.1094A > C	p.Asp365Ala	Ig super family	327–419
LILRB1	c.1117_1118delTA	p.Tyr373fs	Ig super family	327–419
LILRB1	c.1128A > T	p.Gln376His	Ig super family	327–419
LILRB1	c.1153G > A	p.Gly385Ser	Ig super family	327–419
PRSS3	c.10 T > C	p.Phe4Leu	Trypsin-like serine protease	38–256
PRSS3	c.244G > A	p.Val82Ile	Trypsin-like serine protease	38–256

TABLE 8 MAESTRO program protein stability prediction on *LILRB1* and *PRSS3* variants.

Substitutions	Gene name	$\Delta\Delta G_{pred}$ (kcal/mol)	C_{pred} (kcal/mol)
G52.A(S)	*LILRB1*	−0.170	0.909
Q53.A(L)	*LILRB1*	0.074	0.880
R59.A(H)	*LILRB1*	−0.154	0.919
P86.A(L)	*LILRB1*	−0.363	0.885
Y99.A(N)	*LILRB1*	0.85	0.923
Y99.A(F)	*LILRB1*	−0.112	0.903
R107.A(L)	*LILRB1*	−0.607	0.864
Q130.A(L)	*LILRB1*	−0.132	0.916
S327.A(F)	*LILRB1*	−0.438	0.860
Q331.A(R)	*LILRB1*	−0.063	0.873
W351.A(G)	*LILRB1*	0.380	0.875
Q353.A(H)	*LILRB1*	−0.047	0.911
D365.A(Y)	*LILRB1*	−0.224	0.881
D365.A(A)	*LILRB1*	0.083	0.885
Q376.A(H)	*LILRB1*	0.086	0.872
G385.A(S)	*LILRB1*	−0.169	0.865
F4.A(L)	*PRSS3*	0.016	0.885
V25.A(I)	*PRSS3*	0.799	0.857

ΔΔG Positive score, Destabilizing; Negative score, Stabilizing.

TABLE 9 YASARA program RMSD at residue and whole level.

Gene name	Substitutions	Calpha RMSD (Å)	RMSD (Å)
LILRB1	Q53.A(L)	0.055	1.8395
LILRB1	Y99.A(N)	0.055	2.0688
LILRB1	W351.A(G)	0.055	1.5186
LILRB1	D365.A(A)	0.055	2.0098
LILRB1	Q376.A(H)	0.055	2.0351
PRSS3	F4.A(L)	0.121	2.1456
PRSS3	V25.A(I)	0.120	2.0270

Cutoff RMSD at the polypeptide chain > 0.2, residue levels > 2.

these variants, five (p.Gln53Leu; p.Tyr99Asn; p.Trp351Gly; p.Asp365Ala; and p.Gln376His) were seen to have a destabilizing effect on the corresponding protein with ΔΔG upon mutations of 0.074, 0.85, 0.380, 0,083, and 0.086 (kcal/mol), respectively, and the c_{pred} upon mutations of 0.088, 0.923, 0.875, 0.885, and 0.872 (kcal/mol) respectively. All these variants were rare and not present in public databases such as the Greater Middle East (GME), the KAIMRC Genomic Database (KGD), and the Genome Aggregation Database (gnomAD) (16, 39). Various *LILRB1* rare variants seen in these families might be dysregulating several immune pathways, such as adaptive immunity, that normally prevent pathogens from growing by specialized, systemic cells and processes (40). Another important pathway is the

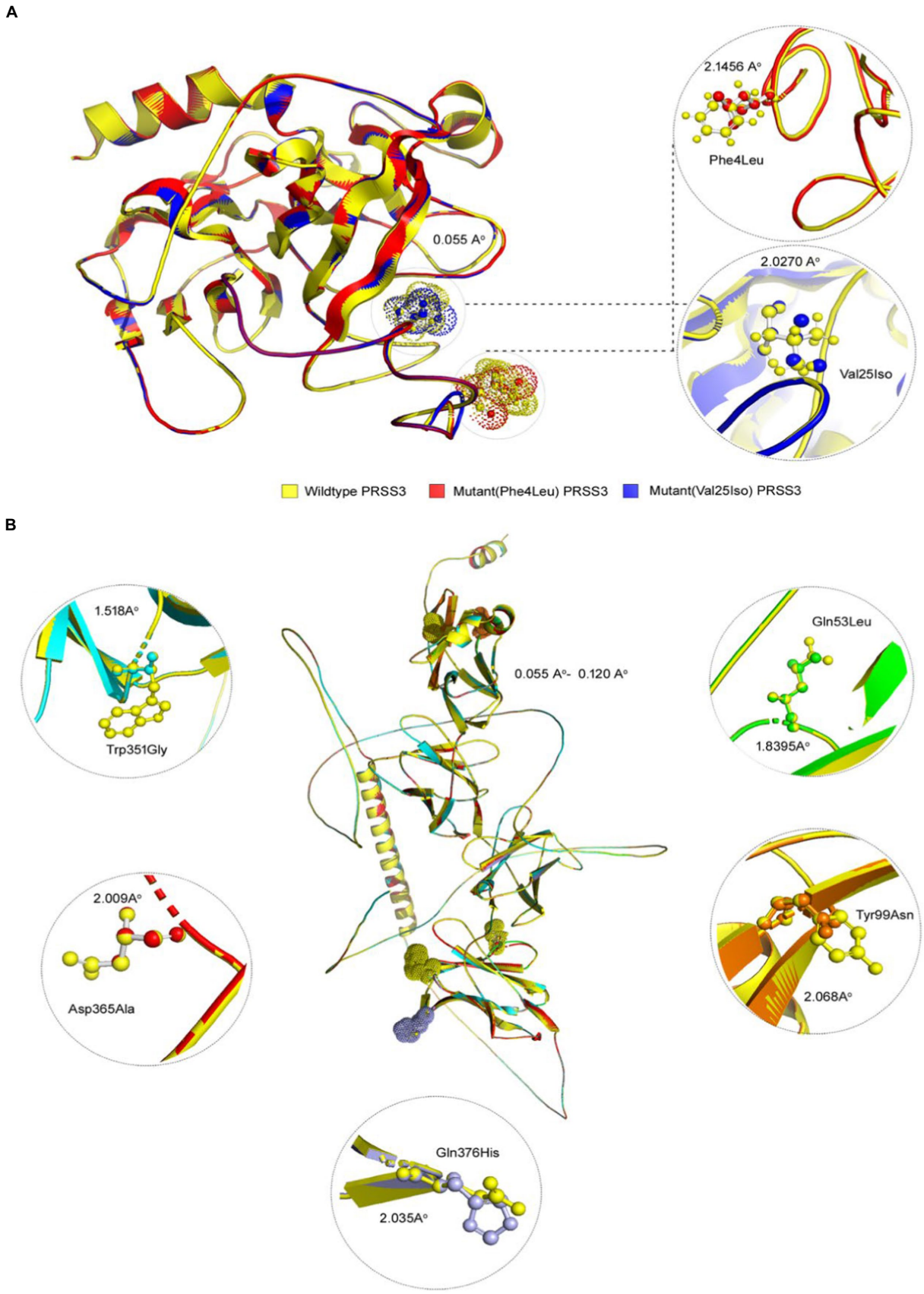


FIGURE 5
3D structures of *PRSS3* and *LILRB1* wild and mutant protein models. Structures of **(A)** *PRSS3* wild type in yellow, and mutant (p.Phe4Leu and p.Val25Ile) in red and blue, respectively **(B)** *LILRB1* wild type in yellow, and mutant (p.Gln53Leu) in green (p.Tyr99Asn) orange (p.Trp351Gly) blue (p.Asp365Ala), red, and (p.Gln376His) purple.

immunoregulatory interactions between a lymphoid and a non-lymphoid cell. A variety of receptors and cell adhesion molecules play important roles in modifying the response of lymphoid cells (such as B-, T-, and NK cells) to self, tumor antigens, and pathogenic organisms (41). Since the innate immune system detects microbial infections, any defect in this system could lead to microbial imbalance that could trigger IBD development.

The second gene, *PRSS3,* is a member of the trypsin family of serine proteases (synonyms: *PRSS4, TRY3*, and *TRY4*). This enzyme is found in the brain and pancreas, and it is resistant to common trypsin inhibitors. It acts on peptide bonds containing the carboxyl group of lysine or arginine. This gene is located on chromosome 9 at the locus of the T cell receptor beta variable orphans. *The PRSS3* gene has four transcripts encoding distinct isoforms. Furthermore, this gene is a known contributor to the initiation and progression of malignant tumors, but its significance in gastric cancer (GC) remains unknown (42). This is the first report linking the novel potential role of the *PRSS3* gene to IBD through shared rare variant burden analysis in three families from Saudi Arabia presenting late-onset IBD.

In the present study, we found that both families A and C shared the same missense variant for *PRSS3* (c.244G > A; rs76740888). Family B had a missense variant for *PRSS3* (c.10 T > C; rs772714741). The frequency of the *PRSS3*, c.244G > A variant in the GME variome project is 11%, which might be seen only among the Arab population. However, this variant is not present in gnomAD. Moreover, two prediction tools, the Mutation Taster and the likelihood ratio test (LRT), show that this variant is damaging. The frequency of the c.10 T > C variant is rare and not present among GME, KGD, and gnomAD. Interestingly, both variants have a destabilizing effect on the protein structure, with $\Delta\Delta G$ of 0.016 and 0.799 (kcal/mol) and c_{pred} of 0.885 and 0.857 (kcal/mol) (43). Destabilizing mutations reduce the stability of a protein and may lead to its misfolding, aggregation, and degradation (44).

Different rare variants of the *PRSS3* gene might be perturbing several immune pathways, such as the innate immune system and neutrophil degranulation (45). Any defect in these important pathways could harm autoimmunity, which will lead to the development of any disease linked to autoimmunity, such as IBD. Our findings suggest a novel strategy for deciphering the complex genetic basis of IBD through the whole exome sequence (WES) analysis of familial cases combined with computational analysis. This study was performed on three consanguineous Saudi families with IBD with each family having more than one affected sibling.

We sincerely acknowledge some limitations of this study. First, our findings were limited to three families with IBD, and studying more familial cases will help establish the role of the *LILRB1* and *PRSS3* and other potential causal genes, biomarkers, and drug targets for IBD. But our findings could be a proof of concept that rare variant burden (RVB) can assist in unraveling the genetic complexity of IBD, where classical Mendelian segregation models are of limited use. Second, while our study was conducted on humans, studying the role of LILRB1 and PRSS3 genetic variants on intestinal cell lines and animal models could aid in understanding how mutant proteins modulate autoimmune responses at the tissue level. Third, computational methods often show variable predictions; hence, their results should be interpreted in the context of subsequent biological experiment-based verifications.

5. Conclusion

This study proposes a novel strategy for understanding the genetic complexity of IBD by combining WES and computational multi-dimensional biological data analysis to identify potential IBD key proteins. Our findings suggest that the rare and novel variants identified in two potential key proteins (*LILRB1* and *PRSS3*) are likely to contribute to IBD pathogenesis *via* several important immune pathways, such as the innate and adaptive immune system pathways and neutrophil degranulation.

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 national policy on genomic data sharing in the public domain outside the country. Allowed data under the above mentioned restrictions of the IRB and participants requirements is presented in the article as well in the supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) protocols at King Abdulaziz University Hospital. The patients/participants provided their written informed consent to participate in this study.

Author contributions

RJ, RE, and NS: conceptualization and writing—original draft preparation. RJ, ZA, BB, and RE: methodology. RJ and BB: software and visualization. RJ, BB, RE, and NS: formal analysis. BB and NS: resources. RJ, HA-N, ZA, NA-T, NA, HA, MA, BA, NS, YQ, OS, BB, MM, and RE: writing—review and editing. NS, OS, BB, and RE: supervision. RE: project administration and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number (IFPHI-130-140-2020) and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

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

References

1. Coward S, Clement F, Benchimol EI, Bernstein CN, Avina-Zubieta JA, Bitton A, et al. Past and future burden of inflammatory bowel diseases based on modeling of population-based data. *Gastroenterology*. (2019) 156:1345–1353.e4. doi: 10.1053/j.gastro.2019.01.002

2. Horowitz J, Warner N, Staples J, Crowley E, Gosalia N, Murchie R, et al. Mutation spectrum of NOD2 reveals recessive inheritance as a main driver of early onset Crohn's disease. *Sci Rep*. (2021) 11:5595. doi: 10.1038/s41598-021-84938-8

3. Wijmenga C. Expressing the differences between Crohn disease and ulcerative colitis. *PLoS Med*. (2005) 2:e230; quiz e304. doi: 10.1371/journal.pmed.0020230

4. Mosli M, Alawadhi S, Hasan F, Abou Rached A, Sanai F, Danese S. Incidence, prevalence, and clinical epidemiology of inflammatory bowel disease in the Arab world: A systematic review and meta-analysis. *Inflamm Intest Dis*. (2021) 6:123–31. doi: 10.1159/000518003

5. Mosli M, Alzahrani A, Showlag S, Alshehri A, Hejazi A, Alnefaie M, et al. A cross-sectional survey of multi-generation inflammatory bowel disease consanguinity and its relationship with disease onset. *Saudi J Gastroenterol*. (2017) 23:337–40. doi: 10.4103/sjg.SJG_125_17

6. Cho JH, Brant SR. Recent insights into the genetics of inflammatory bowel disease. *Gastroenterology*. (2011) 140:1704–1712.e2. doi: 10.1053/j.gastro.2011.02.046

7. Bianco AM, Girardelli M, Tommasini A. Genetics of inflammatory bowel disease from multifactorial to monogenic forms. *World J Gastroenterol*. (2015) 21:12296–310. doi: 10.3748/wjg.v21.i43.12296

8. Cordero RY, Cordero JB, Stiemke AB, Datta LW, Buyske S, Kugathasan S, et al. Trans-ancestry, Bayesian meta-analysis discovers 20 novel risk loci for inflammatory bowel disease in an African American, east Asian and European cohort. *Hum Mol Genet*. (2023) 32:873–82. doi: 10.1093/hmg/ddac269

9. Lee JC, Biasci D, Roberts R, Gearry RB, Mansfield JC, Ahmad T, et al. Genome-wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat Genet*. (2017) 49:262–8. doi: 10.1038/ng.3755

10. Manolio TA. Genomewide association studies and assessment of the risk of disease. *N Engl J Med*. (2010) 363:166–76. doi: 10.1056/NEJMra0905980

11. Moran CJ, Klein C, Muise AM, Snapper SB. Very early-onset inflammatory bowel disease: gaining insight through focused discovery. *Inflamm Bowel Dis*. (2015) 21:1166–75. doi: 10.1097/MIB.0000000000000329

12. Sazonovs A, Stevens CR, Venkataraman GR, Yuan K, Avila B, Abreu MT, et al. Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. *Nat Genet*. (2022) 54:1275–83. doi: 10.1038/s41588-022-01156-2

13. Alharbi RS, Shaik NA, Almahdi H, Elsokary HA, Jamalalail BA, Mosli MH, et al. Genetic association study of NOD2 and IL23R amino acid substitution polymorphisms in Saudi inflammatory bowel disease patients. *J King Saud Univ*. (2022) 34:101726. doi: 10.1016/j.jksus.2021.101726

14. Uniken Venema WT, Voskuil MD, Dijkstra G, Weersma RK, Festen EA. The genetic background of inflammatory bowel disease: From correlation to causality. *J Pathol*. (2017) 241:146–58. doi: 10.1002/path.4817

15. Al-Abbasi FA, Mohammed K, Sadath S, Banaganapalli B, Nasser K, Shaik NA. Computational protein phenotype characterization of IL10RA mutations causative to early onset inflammatory bowel disease (IBD). *Front Genet*. (2018) 9:146. doi: 10.3389/fgene.2018.00146

16. Al-Numan HH, Jan RM, Al-Saud NBS, Rashidi OM, Alrayes NM, Alsufyani HA, et al. Exome sequencing identifies the extremely rare ITGAV and FN1 variants in early onset inflammatory bowel disease patients. *Front Pediatr*. (2022) 10:895074. doi: 10.3389/fped.2022.895074

17. 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

18. Prescott NJ, Lehne B, Stone K, Lee JC, Taylor K, Knight J, et al. Pooled sequencing of 531 genes in inflammatory bowel disease identifies an associated rare variant in BTNL2 and implicates other immune related genes. *PLoS Genet*. (2015) 11:e1004955. doi: 10.1371/journal.pgen.1004955

19. Danese S, Fiorino G, Michetti P. Changes in biosimilar knowledge among European Crohn's colitis organization [ECCO] members: an updated survey. *J Crohn's Colitis*. (2016) 10:1362–5. doi: 10.1093/ecco-jcc/jjw090

20. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic Acids Res*. (2013) 41:D1228–33. doi: 10.1093/nar/gks1147

21. Khan MM, Mohsen MT, Malik MZ, Bagabir SA, Alkhanani MF, Haque S, et al. Identification of potential key genes in prostate cancer with gene expression, pivotal pathways and regulatory networks analysis using integrated bioinformatics methods. *Genes*. (2022) 13:655. doi: 10.3390/genes13040655

22. Malik MZ, Chirom K, Ali S, Ishrat R, Somvanshi P, Singh RKB. Methodology of predicting novel key regulators in ovarian cancer network: a network theoretical approach. *BMC Cancer*. (2019) 19:1129. doi: 10.1186/s12885-019-6309-6

23. Ben-Yosef N, Frampton M, Schiff ER, Daher S, Abu Baker F, Safadi R, et al. Genetic analysis of four consanguineous multiplex families with inflammatory bowel disease. *Gastroenterol Rep*. (2021) 9:521–32. doi: 10.1093/gastro/goab007

24. Mansour H, Banaganapalli B, Nasser KK, Al-Aama JY, Shaik NA, Saadah OI, et al. Genome-wide association study-guided exome rare variant burden analysis identifies IL1R1 and CD3E as potential autoimmunity risk genes for celiac disease. *Front Pediatr*. (2022) 10:837957. doi: 10.3389/fped.2022.837957

25. Stange EF. Current and future aspects of IBD research and treatment: the 2022 perspective. *Front Gastroenterol*. (2022) 1:914371. doi: 10.3389/fgstr.2022.914371

26. Singh T, Poterba T, Curtis D, Akil H, Al Eissa M, Barchas JD, et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. *Nature*. (2022) 604:509–16. doi: 10.1038/s41586-022-04556-w

27. Sims R, Van Der Lee SJ, Naj AC, Bellenguez C, Badarinarayan N, Jakobsdottir J, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nat Genet*. (2017) 49:1373–84. doi: 10.1038/ng.3916

28. Balicza P, Varga N, Bolgár B, Pentelényi K, Bencsik R, Gál A, et al. Comprehensive analysis of rare variants of 101 autism-linked genes in a Hungarian cohort of autism Spectrum disorder patients. *Front Genet*. (2019) 10:434. doi: 10.3389/fgene.2019.00434

29. Al-Herz W, Aldhekri H, Barbouche MR, Rezaei N. Consanguinity and primary immunodeficiencies. *Hum Hered*. (2014) 77:138–43. doi: 10.1159/000357710

30. Romdhane L, Mezzi N, Hamdi Y, El-Kamah G, Barakat A, Abdelhak S. Consanguinity and inbreeding in health and disease in north African populations. *Annu Rev Genomics Hum Genet*. (2019) 20:155–79. doi: 10.1146/annurev-genom-083118-014954

31. Awan Z, Alrayes N, Khan Z, Almansouri M, Ibrahim Hussain Bima A, Almukadi H, et al. Identifying significant genes and functionally enriched pathways in familial hypercholesterolemia using integrated gene co-expression network analysis. *Saudi J Biol Sci*. (2022) 29:3287–99. doi: 10.1016/j.sjbs.2022.02.002

32. Banaganapalli B, Mallah B, Alghamdi KS, Albaqami WF, Alshaer DS, Alrayes N, et al. Integrative weighted molecular network construction from transcriptomics and genome wide association data to identify shared genetic biomarkers for COPD and lung cancer. *PLoS One*. (2022) 17:e0274629. doi: 10.1371/journal.pone.0274629

33. Bima AI, Elsamanoudy AZ, Alamri AS, Felimban R, Felemban M, Alghamdi KS, et al. Integrative global co-expression analysis identifies key microRNA-target gene networks as key blood biomarkers for obesity. *Minerva Med*. (2022) 113:532–41. doi: 10.23736/S0026-4806.21.07478-4

34. Bima AIH, Elsamanoudy AZ, Albaqami WF, Khan Z, Parambath SV, Al-Rayes N, et al. Integrative system biology and mathematical modeling of genetic networks identifies shared biomarkers for obesity and diabetes. *Math Biosci Eng*. (2022) 19:2310–29. doi: 10.3934/mbe.2022107

35. 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

36. Kumar SU, Balasundaram A, Cathryn RH, Varghese RP, Siva R, Gnanasambandan R, et al. Whole-exome sequencing analysis of NSCLC reveals the pathogenic missense variants from cancer-associated genes. *Comput Biol Med*. (2022) 148:105701. doi: 10.1016/j.compbiomed.2022.105701

37. Lan X, Liu F, Ma J, Chang Y, Lan X, Xiang L, et al. Leukocyte immunoglobulin-like receptor A3 is increased in IBD patients and functions as an anti-inflammatory modulator. *Clin Exp Immunol*. (2021) 203:286–303. doi: 10.1111/cei.13529

38. Oliveira MLG, Castelli EC, Veiga-Castelli LC, Pereira ALE, Marcorin L, Carratto TMT, et al. Genetic diversity of the LILRB1 and LILRB2 coding regions in an admixed Brazilian population sample. *HLA*. (2022) 100:325–48. doi: 10.1111/tan.14725

39. Alharthi AM, Banaganapalli B, Hassan SM, Rashidi O, Al-Shehri BA, Alaifan MA, et al. Complex inheritance of rare missense variants in PAK2, TAP2, and PLCL1 genes in a consanguineous Arab family with multiple autoimmune diseases including celiac disease. *Front Pediatr*. (2022) 10:895298. doi: 10.3389/fped.2022.895298

40. Iwasaki A, Medzhitov R. Control of adaptive immunity by the innate immune system. *Nat Immunol*. (2015) 16:343–53. doi: 10.1038/ni.3123

41. Symowski C, Voehringer D. Interactions between innate lymphoid cells and cells of the innate and adaptive immune system. *Front Immunol*. (2017) 8:1422. doi: 10.3389/fimmu.2017.01422

42. Wang F, Hu YL, Feng Y, Guo YB, Liu YF, Mao QS, et al. High-level expression of PRSS3 correlates with metastasis and poor prognosis in patients with gastric cancer. *J Surg Oncol*. (2019) 119:1108–21. doi: 10.1002/jso.25448

43. Backwell L, Marsh JA. Diverse molecular mechanisms underlying pathogenic protein mutations: beyond the loss-of-function paradigm. *Annu Rev Genomics Hum Genet*. (2022) 23:475–98. doi: 10.1146/annurev-genom-111221-103208

44. Kumar SU, Kumar DT, Siva R, Doss CGP, Zayed H. Integrative bioinformatics approaches to map potential novel genes and pathways involved in ovarian cancer. *Front Bioeng Biotechnol*. (2019) 7:391. doi: 10.3389/fbioe.2019.00391

45. Gierlikowska B, Stachura A, Gierlikowski W, Demkow U. Phagocytosis, degranulation and extracellular traps release by neutrophils-the current knowledge, pharmacological modulation and future prospects. *Front Pharmacol*. (2021) 12:666732. doi: 10.3389/fphar.2021.666732

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 07 June 2023
DOI 10.3389/fmed.2023.1151046

OPEN ACCESS

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

REVIEWED BY
Prabhash Kumar Jha,
Brigham and Women's Hospital,
Harvard Medical School,
United States
Konda Mani Saravanan,
Bharath Institute of Higher Education and Research,
India

*CORRESPONDENCE
S. Sajitha Lulu
ssajithalulu@vit.ac.in

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

RECEIVED 25 January 2023
ACCEPTED 20 March 2023
PUBLISHED 07 June 2023

CITATION
Premkumar T and Sajitha Lulu S (2023) Molecular crosstalk between COVID-19 and Alzheimer's disease using microarray and RNA-seq datasets: A system biology approach. *Front. Med.* 10:1151046.
doi: 10.3389/fmed.2023.1151046

Molecular crosstalk between COVID-19 and Alzheimer's disease using microarray and RNA-seq datasets: A system biology approach

T. Premkumar and S. Sajitha Lulu*

Department of Biotechnology, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Objective: Coronavirus disease 2019 (COVID-19) is an infectious disease caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). The clinical and epidemiological analysis reported the association between SARS-CoV-2 and neurological diseases. Among neurological diseases, Alzheimer's disease (AD) has developed as a crucial comorbidity of SARS-CoV-2. This study aimed to understand the common transcriptional signatures between SARS-CoV-2 and AD.

Materials and methods: System biology approaches were used to compare the datasets of AD and COVID-19 to identify the genetic association. For this, we have integrated three human whole transcriptomic datasets for COVID-19 and five microarray datasets for AD. We have identified differentially expressed genes for all the datasets and constructed a protein–protein interaction (PPI) network. Hub genes were identified from the PPI network, and hub genes-associated regulatory molecules (transcription factors and miRNAs) were identified for further validation.

Results: A total of 9,500 differentially expressed genes (DEGs) were identified for AD and 7,000 DEGs for COVID-19. Gene ontology analysis resulted in 37 molecular functions, 79 cellular components, and 129 biological processes were found to be commonly enriched in AD and COVID-19. We identified 26 hub genes which includes *AKT1, ALB, BDNF, CD4, CDH1, DLG4, EGF, EGFR, FN1, GAPDH, INS, ITGB1, ACTB, SRC, TP53, CDC42, RUNX2, HSPA8, PSMD2, GFAP, VAMP2, MAPK8, CAV1, GNB1, RBX1*, and *ITGA2B*. Specific miRNA targets associated with Alzheimer's disease and COVID-19 were identified through miRNA target prediction. In addition, we found hub genes-transcription factor and hub genes-drugs interaction. We also performed pathway analysis for the hub genes and found that several cell signaling pathways are enriched, such as PI3K-AKT, Neurotrophin, Rap1, Ras, and JAK–STAT.

Conclusion: Our results suggest that the identified hub genes could be diagnostic biomarkers and potential therapeutic drug targets for COVID-19 patients with AD comorbidity.

KEYWORDS

COVID-19, Alzheimer's disease, regulatory networks, comorbidity, biomarkers

1. Introduction

SARS-CoV-2 (Severe Acute Respiratory Syndrome-Corona Virus Disease 2019) become a major health issue and highest prevalence rate (1). According to the world health organization (WHO) report worldwide, the COVID-19 outbreak affected over 600 million people and 6.8 million of them died, as of 6 march 2023 a total of 1.3B vaccine doses have been administrated.[1] SARS-CoV-2 genome consists of 29,811 nucleotides of enveloped positive-stranded ssRNA; as a result, SARS-CoV-2 appears to bind exclusively to angiotensin-converting enzyme 2 (ACE2) (2). This causes severe acute respiratory distress. ACE2 expression levels are highest in the small intestine, testis, heart, kidneys, and thyroid and the lowest in the brain, bone marrow, spleen, blood, blood vessels, and muscle (3). COVID-19 vaccines were developed and deployed rapidly, successfully controlled the pandemic, and reduced the risk of associated death and severe illness (4–6). COVID-19 poses a greater risk of death for patients with pre-existing neurological conditions (7). Virus RNA transcripts and viral proteins were also found in brain tissues of COVID-19 patients during an autopsy (8, 9). Neurological symptoms have been reported in COVID-19 cases more notably in recovered patients from COVID-19 challenged memory loss and cognitive disability (10). Clinical studies have proven the possibility of COVID-19 pathogenesis in the brain, and, some studies pointed out that COVID-19 might accelerate the neurodegeneration of Alzheimer's Disease (AD) and Parkinson's Disease (5, 11–15). As a result of COVID-19, cognitive impairment may be caused by the following mechanisms like Direct COVID-19 infection in CNS, Systematic hyperinflammatory response to COVID-19, Peripheral organ dysfunction, Severe coagulopathy, Cerebrovascular ischemia due to endothelial dysfunction, and Mechanical ventilation due to severe disease conditions (16, 17).

Alzheimer's Disease is a neurodegenerative disorder more than 50 m people are affected worldwide and this count is expected 150 m in 2050 (18). The major reason for AD is a breakdown of amyloid precursor protein (APP) in the brain which generates beta-amyloid (Aβ) in extracellular neural space (19–21). Several enzymes reported for the breakdown of APP importantly three secretase enzymes such as alpha-secretase, beta-secretase, and gamma-secretase play crucial roles in the cleavage process (22–24). Another possible mechanism of AD is an intracellular hyperphosphorylated tau protein (25). The tau protein plays a vital role in the stabilization and assembly of microtubules, as well as in regulating plasticity and synaptic function. Tau protein hyper phosphorylates under certain physiological conditions, resulting in the destabilization of associated microtubules, synaptic damage, and other complications (26, 27). A higher permeability of BBB might permit viruses and bacteria to enter the brain (28). Several pathogens are implicated in the development of AD, including viruses, bacteria, fungi, and parasites (29). COVID-19 crosses the BBB and induces an inflammatory response within microvascular endothelial cells leading to BBB dysfunction (16, 30). In previous studies, integrated bioinformatics and system biology approaches also investigated the impact of SARS-CoV-2 on neurological disease progression (31–33). Systems biology provides a comprehensive interpretation of high-throughput platforms including genomics, proteomics, and metabolomics for analysis, display, compatibility, and accessibility. Comorbidity analysis for diverse diseases has become possible with the availability of high-throughput data and system biology bioinformatics approaches also provides a better way to unravel the biological complexity of these multifactorial diseases influenced by multiple pathogenic determinants (34, 35). To investigate the molecular factors that influence the development of SARS-CoV-2 and neurological comorbidities, we investigated multiple gene expression datasets from AD and SARS-CoV-2 which includes microarray data and transcriptome data from various human brain tissue and blood samples. We proposed a network-based systems biology approach to explore the relationship between AD and SARS-CoV-2.

2. Materials and methods

2.1. Data collection

We have used gene expression datasets such as transcriptome datasets and microarray datasets to find the differentially expressed genes. This collection of datasets was extracted from gene expression omnibus (GEO) at the National Center for Biotechnology Information[2] (36, 37).

For our analysis, we used the following inclusion criteria:

1. Dataset which contains samples from the disease group and the control group in original experimental studies.
2. Expression profiling by array used for AD with GEO2R tool support.
3. Expression profiling by high throughput sequencing with raw counts data used for COVID-19.
4. Only homo-sapiens datasets were included.
5. A dataset containing at least eight samples included.

The keywords used for AD include "Alzheimer's Disease" and further the results were filtered by the term "homo-sapiens," and we selected the study type "expression profiling by array" which resulted in five datasets for AD. Among the five datasets, three of them were associated with peripheral blood mononuclear cells (PBMCs), and two of them were brain tissue-based. For COVID-19 we used the keywords "SARS-CoV-2" to narrow down the results and further filtered them by "homo-sapiens," and "expression profiling by high-throughput sequencing." We retrieved three datasets related to COVID-19, including two PBMC datasets and one brain tissue dataset. Both control (non-diseased) and diseased samples are included in all the datasets Table 1.

1 https://covid19.who.int/

2 https://www.ncbi.nlm.nih.gov/

TABLE 1 Microarray datasets obtained from the GEO database with the search key terms "Alzheimer's Disease" and "SARS-CoV-2" with a filter restricting to "Homo Sapiens."

S. No	Accession ID	Platform	Sample count (case/control)	Analysis methods
1	GSE4226	GPL1211, NIA MGC, Mammalian Genome Collection	AD;14/14	GEO2R
2	GSE4229	GPL1211, NIA MGC, Mammalian Genome Collection	AD;12/28	GEO2R
3	GSE18309	GPL570, Affymetrix Human Genome U133 Plus 2.0 Array	AD;6/3	GEO2R
4	GSE97760	GPL16699, Agilent-039494 Sure Print G3 Human GE v2 8x60K Microarray	AD;9/10	GEO2R
5	GSE36980	GPL6244, Affymetrix Human Gene 1.0 ST Array	AD;33/47	GEO2R
6	GSE152418	GPL24676, Illumina NovaSeq 6,000	COVID;17/17	DESeq2
7	GSE166190	GPL20301, Illumina HiSeq 4,000	COVID;11/11	DESeq2
8	GSE174745	GPL24676, Illumina NovaSeq 6,000	COVID;6/3	DESeq2

Expression type microarray and RNA-Seq to Alzheimer's Disease and SARS-CoV-2, respectively.

2.2. Preprocessing and identification of differentially expressed genes

To classify genes with significantly different expression levels between samples, differential gene expression analysis is necessary. GEO2R was used to identify DEGs from microarray data, the selected microarray datasets have two groups control and disease (37, 38). The (Linear Models for Microarray Data) limma Bioconductor package is also available in GEO2R online tool for finding the differentially expressed genes (39). As part of the normalization process, outliers were removed using the log2 transform, and the Benjamin Hackenberg methods are used by default to correct p value (40). To perform DEGs analysis, we selected false discovery rate (FDR) p values adjusted for multiple testing. We downloaded the full table with the following columns for further analysis value of p, adjusted value of p, log fold change, gene symbol, and title (41). Following DEGs, we plotted a volcano plot using the pheatmap package in R, genes with p value <0.05, and log FC | > 1 was considered (42).

For transcriptomics datasets, we have used a DESeq2 Bioconductor package (version 3.16) in RStudio version 2022. The transcriptome profile of COVID-19 tissues and blood samples was compared with control tissues and blood samples. DESeq2 is a statistical model designed to identify differentially expressed genes between two or more conditions, it is often used in the analysis of RNA-Seq data, to identify the genes which change in expression between different biological samples or conditions (43, 44). The DESeq2 model uses a negative binomial distribution to model the count data obtained from RNA-Seq experiments and variance for each gene across all samples. The model accounts for technical variability such as differences in sequencing depth, and for biological variabilities such as differences in cell size or the presence of outliers (44).

Once the mean and variance for each gene are estimated, the DESeq2 model uses a hypothesis testing framework to determine which genes are significantly differentially expressed between the conditions of interest. The resulting p value and log fold changes are then used to rank the genes based on their level of differential expression (45, 46).

2.3. Identification of common gene ontology terms and identification of overlapped genes among COVID-19 and Alzheimer's disease

Followed by preprocessing and DEGs identification of COVID-19 and AD datasets, we classified them into four different groups AD-PBMC, AD-Tissue, COVID-19-PBMC, and COVID-19 -Tissue (47). To identify the overlapped gene among these four groups, a Venn diagram was created using an online Venn diagram tool Interactive Venn.[3] Then the identified common genes were taken for constructing a (Module 1) PPI network for further analysis. Web-based database for annotation visualization and integrated discovery (DAVID)[4] tool was used to perform a gene ontology analysis for DEGs for Alzheimer's disease and COVID-19 independently (48). We have taken only those genes with common GO terms among AD and COVID-19 for further analysis and constructed a PPI network (Module 2).

2.4. Protein–protein interaction analysis and hub genes prediction

The biological functions and possible associations are mainly carried out by the PPI and we constructed two PPI networks. The

3 http://www.interactivenn.net/

4 https://david.ncifcrf.gov/

first protein interaction network (Module 1) was constructed using the common differentially expressed genes between the four groups and on other hand, the PPI network (module 2) was constructed using the genes with common GO terms. The protein interactions were constructed using STRING version 11.5[5] online tool then the PPI network was analyzed and visualized through Cytoscape software[6] (49). The protein interaction networks are large networks and every node is connected with an edge, the highly interconnected genes (edges) in the PPI network consider hub genes. After constructing the two PPI networks we used the CytoHubba plugin version 0.1 in Cytoscape to identify the highly connected genes (50). Four topological features or ranking methods such as maximal clique centrality (MCC), Degree, Closeness, and Betweenness were employed to identify the hub genes. We have collected the top 20 genes from every method, and the gene present in at least three ranking methods were considered hub genes (51).

2.5. Analysis of transcription factor and microRNAs of hub genes

The interaction between hub genes-transcription factors (TFs) and hub genes-microRNAs (miRNA) has been conducted. Transcription factors play a crucial role, it binds with specific genes and regulates the rate of transcription of genetic information. Bioinformatically and/or *in vitro* assessment is possible of some of the mechanistic functions of candidate miRNAs prior to conducting preclinical animal tests (52). Cytoscape iRegulon plugin version 1.3 was used to predict the potential interactions between hub genes and TFs. In iRegulon, the enriched motifs were ranked depending on the direct targets using the position weight matrix (53). Therefore, AD and COVID-19 associated hub genes miRNA targets were predicted by using miRDB (MicroRNA Target Prediction Database).[7] The miRNA targets predictive score (rank) >80 was considered a reliable score (54). The identified miRNAs were further plotted using Cytoscape software. For a better understanding of the role of miRNAs in disease mechanisms, we identified the hub miRNAs using four ranking methods (Degree, betweenness, closeness, and stress) of the CytoHubba plugin in Cytoscape (55, 56).

2.6. Drug-gene interaction analysis of hub genes

The drug-gene interaction was identified using Drug Gene Interaction Database (DGIdb) (57). DGIdb interface provides a search for genes against a database of drug-gene interactions and druggable targets. FDA approval status was confirmed through the drug bank database for shortlisted drugs in the interaction (Figure 1).

5 https://string-db.org/

6 https://cytoscape.org/

7 https://mirdb.org/

2.7. Gene ontology and pathway analysis of hub genes

Cluster Profiler (Version 4.1.0) Bioconductor package in R was used for creating Gene ontology of the hub genes (58). The top gene-ontology of molecular function (MF), cellular component (CC), and biological process (BP) were plotted using a bubble plot, and biochemical pathways associated with hub genes were identified using the KEGG database (Kyoto encyclopedia genes and genomes) (59).

3. Statistical analysis

3.1. DEGs

DEGs were identified for each data set by using adjusted *p*-values based on the moderated t-statistic (adj P) <0.05 along with an absolute value of logFC (log foldchange) of >1. The logFC ≥ 1 was considered as upregulated genes and logFC ≤ -1 was considered as downregulated genes.

3.2. Gene set enrichment analysis

The enrichment analysis of the gene ontology terms was confirmed using the "cluster Profiler" package, the analysis was performed separately for each comparison with applied hypergeometric statistical test, through the below equation,

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

p-values were adjusted for multiple comparisons, and *q*-values were also calculated for FDR control as well. *p*-values <0.05 were considered to be significantly enriched terms (58).

3.3. Gene ontology and pathway analysis

In DAVID, Fisher's Exact test is adopted to measure the gene enrichment in annotation terms. Fisher's Exact *p*-values are computed by summing probabilities P over defined sets of tables ($\text{Prob} = \sum \text{Ap}$). The modified Fisher exact *p*-value (EASE score) ≤ 0.05 and $\text{FDR} < 0.05$ are considered strongly enriched (60, 61).

3.4. Protein interaction network constructions

Protein interactions are assessed and integrated using the STRING database which includes direct (physical) and indirect (functional) associations. PPI networks can be constructed by calculating the distance 'D' between pairs of proteins (u,v),

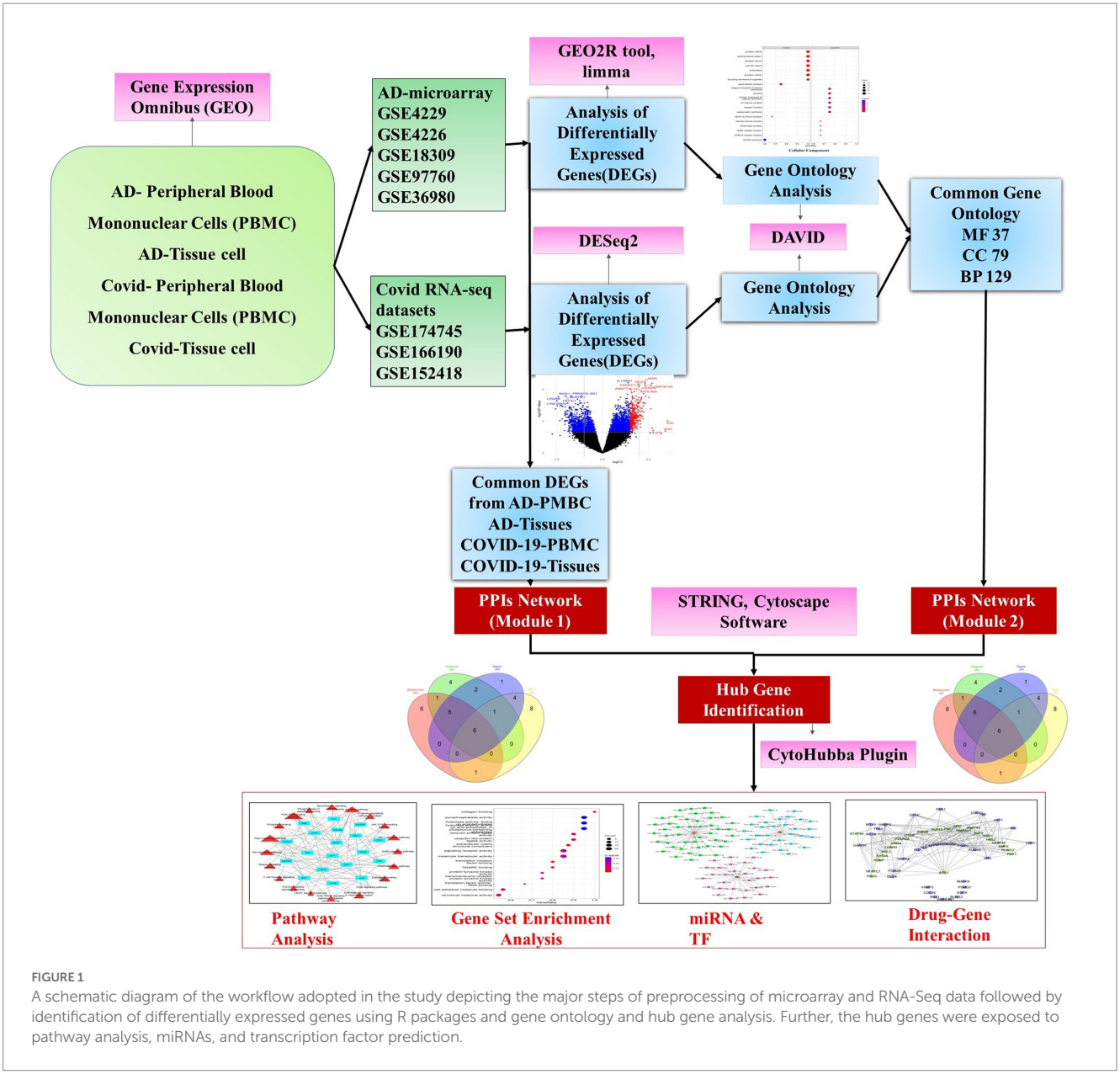


FIGURE 1
A schematic diagram of the workflow adopted in the study depicting the major steps of preprocessing of microarray and RNA-Seq data followed by identification of differentially expressed genes using R packages and gene ontology and hub gene analysis. Further, the hub genes were exposed to pathway analysis, miRNAs, and transcription factor prediction.

$$D(u,v) = \frac{2|Nu \cap Nv|}{|Nu| + |Nv|}$$

STRING tool provides four thresholds as a default including low (0.15), medium (0.40), high (0.70), and highest (0.90) and, we created a network using a medium threshold value (61).

4. Results

4.1. Analysis of microarray and transcriptome datasets

We retrieved five microarray datasets for AD and three transcriptome datasets for COVID-19 which includes disease and healthy samples. The AD microarray datasets were GSE4226, GSE4229, GSE18309, GSE97760, and GSE36980 analyzed through GEO2R. The transcriptome-based COVID-19 datasets GSE152418, GSE166190, and GSE174745 were analyzed through the DESeq2 Bioconductor package in R software. The datasets were analyzed individually and identified the DEGs (Supplementary Tables S1, S2). The overall upregulated and downregulated DEGs were tabulated in Table 2. Followed by DEGs the datasets were classified to four different groups such as AD-PBMC, AD-Tissue, COVID-PBMC, and COVID-Tissue in order to identify a common gene. Figure 2 demonstrates the volcano plots of the AD and SARS-CoV-2 datasets, where the red dot represents a gene that has been upregulated, and the blue dot represents a gene that has been downregulated.

4.2. Identification of common genes

The overlapped genes among the four groups are depicted in the Venn diagram Figure 3 for better understanding. Only 9 (*HST6, POLR3G, SLC6A20, ITGA2B, HOMER3, GMPR, AGBL1, CRABP2,*

TABLE 2 Differentially expressed genes of Alzheimer's disease and COVID-19 datasets with details of upregulated and downregulated genes and total counts after deletion of duplication.

Sample groups	Datasets	Up regulated	Down regulated	Total DEGs	Duplication removed
	GSE4226	2,560	656	18,550	7,944
AD- PBMC	GSE4229	16	318		
	GSE18309	983	886		
	GSE97760	4,733	8,398		
AD-Tissue	GSE36980	1,612	1,121	2,733	1,611
COVID-19-PBMC	GSE152418	1,115	2,545	8,840	5,165
	GSE166190	206	4,974		
COVID-19 Tissue	GSE174745	1867	534	2,401	1864

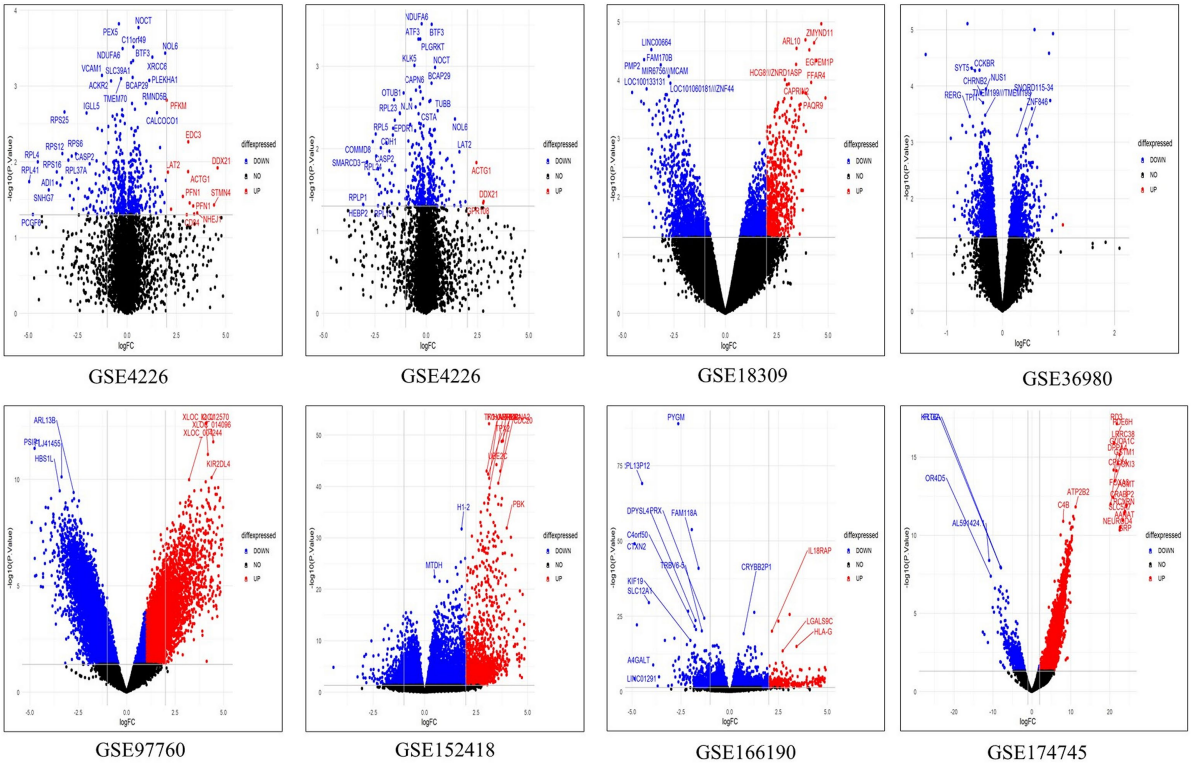


FIGURE 2
The multiple volcano plot showing differentially expressed genes of COVID-19 and AD (upregulated genes in red and downregulated genes in blue). The *x*-axis depicts the log fold change in gene expression between different samples and the *y*-axis depicts FDR-adjusted *p* values.

OLFML2B) genes have been found to be shared between AD-PBMC, AD-Tissue, COVID-19-PBMC, and COVID-19-Tissue. We identified the genes which were present in at least 3 groups and tabulated them (Table 3) for further analysis and construct a (Module 1) PPI network.

4.3. Identification of common gene ontology terms among COVID-19 and Alzheimer's disease datasets

DAVID analysis was performed to understand the biological significance of AD and COVID-19 DEGs. We found 164 MF, 175 CC, and 581 BP were enriched in Alzheimer's disease and 146 MF, 196 CC, and 545 BP were enriched in COVID-19 datasets and 37 MF, 79CC and 129 BP were found to be commonly enriched between Alzheimer's disease and the COVID-19 dataset. For this study, we have considered only the common GO terms for further analysis and (module 2) protein interaction network construction. Supplementary Table S4 gives the details of the commonly enriched GO terms.

4.4. Protein interaction network construction and analysis

The STRING database was used to construct the protein interaction network then visualized *via* Cytoscape software. The edges

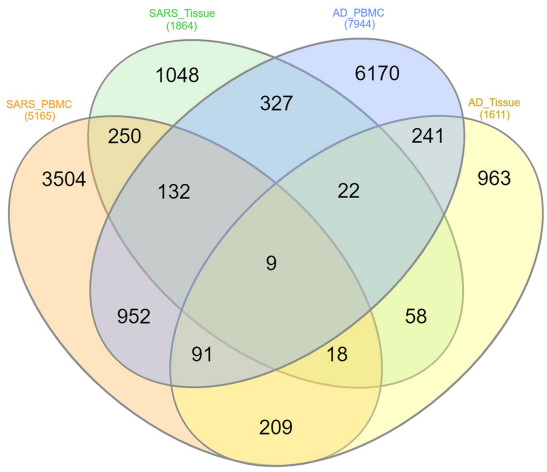


FIGURE 3
Venn diagram of shared differentially expressed genes, where each ellipse represents AD-PBMC, AD-Tissue, COVID-19-PBMC, and COVID-19-Tissue with Nine (*HST6, POLR3G, SLC6A20, ITGA2B, HOMER3, GMPR, AGBL1, CRABP2, OLFML2B*) genes common among the four groups.

TABLE 3 Common genes identified among AD-PBMC, AD-Tissues, COVID-19-PBMC, and COVID-19-Tissues.

S. No	Datasets	Common Genes
1.	AD-PBMC, AD-Tissue, COVID-19-PBMC, COVID-19-Tissue	9
2.	AD-PBMC, COVID-19-PBMC, COVID-19-Tissue	132
3.	AD-Tissue, COVID-19-PBMC, COVID-19-Tissue	22
4	AD-PBMC, AD-Tissue, COVID-19-PBMC	327

represent the interactions between the genes, and the nodes represent the genes. Figure 4 illustrates the (Module 1) PPI network of common genes with 823 edges and 373 nodes. Figure 5 illustrates the (Module 2) PPI network of GO sources with 2,674 nodes and 50,719 edges established according to the results.

4.5. Hub genes identification

Using the CytoHubba plugin of Cytoscape, we identified the highly interacting hub genes for the progression of AD and SARS-CoV-2. Four different algorithms, namely MCC, Degree, Betweenness, and Closeness were utilized to extract the hub genes from module 1 and module 2. We obtained the top 20 genes from both modules based on these four ranking methods and tabulated them in module 1 (Table 4) and module 2 in (Table 5). The gene present in at least 3 ranking methods are considered as hub genes. As a result, Figure 4 displays the list of hub genes (*ACTB, CDC42, RUNX2, HSPA8, PSMD2, GFAP, VAMP2, MAPK8, CAV1, GNB1, RBX1, ITGA2B*) obtained from common genes (module 1) PPI network. A group of 17 (*AKT1, ALB, BDNF, CAV1, CD4, CDC42, CDH1, DLG4, EGF, EGFR, FN1, GAPDH, INS, ITGB1, ACTB, SRC, TP53*) overlapping genes was obtained through gene ontology (module 2) PPI network (Figures 5A,B). We identified that *CAV1, CDC42*, and *ACTB* genes are common among the two sets of hub genes. The expression of Caveolin-1 (Cav-1) has been associated with aging in both senescent cells and aged tissues *in vitro* and *in vivo*. In murine embryonic fibroblasts, Cav-1 knockout accelerates premature senescence, while loss of Cav-1 accelerates neurodegeneration and aging. In most cell types, ACTB (Actin-Beta) is abundantly and stably expressed and is commonly used to normalize gene expression as an internal control (62). ACTB variant rs852423 has been found to be associated with increased susceptibility to AD (63). The identified module 1 and module 2 hub genes and their major roles are tabulated in Supplementary File 2.

4.6. MicroRNAs network of hub genes

The regulatory networks such as miRNAs and TFs of the hub genes were identified. MicroRNAs (miRNA) and transcription factors (TFs) are involved in the development and progression of COVID-19 and its comorbid conditions. Based on the analysis of the hub genes-miRNA and hub genes-Transcription factors, we have obtained a clear network of interactions. The results revealed that the miRNAs regulate 26 hub genes, which could be a possible target of the comorbidity. All the hub genes have targeted a total of 839 miRNAs of which 27 miRNAs were targeted in more than three hub genes (Figure 6A; Supplementary Table S5).

Also, we have identified the hub miRNAs using four ranking methods (Degree, betweenness, closeness, and stress) of the CytoHubba plugin in Cytoscape. We extracted the top 40 nodes from each ranking method and the overlapped miRNAs were identified using a Venn diagram (Figure 6B; Supplementary Table S6). The miRNAs present at least three ranking methods considered as hub-miRNAs and we found five hub-miRNAs including hsa-miR-6,867-5p, hsa-miR-548c-3p, hsa-miR-6,828-3p, hsa-miR-545-5p, and hsa-miR-5,011-5p.

4.7. Transcription factor network of hub genes

iRegulon predicted 85 TFs for the hub genes and importantly four TFs HAND2, GATA1, GATA2, and GATA6 interacted with 23 hub genes (Figure 7; Supplementary Table S7). The heart-and neural crest derivatives expressed protein-2 (HAND2) play a crucial role in neural crest development (64). The synergistic activation between HAND2 and GATA4 TFs is causally linked to congenital heart diseases (CHD). Severe CDH may contribute to delayed brain development, thromboembolism, and pulmonary hypertension. The transcription factors might play a major role in different cell types. GATA family TFs are zinc finger DNA binding proteins, GATA1 and GATA2 play an essential role in developing and maintaining the hematopoietic system (65). Jin Chu et al. reported that GATA1 acts as a transcription repressor for gamma-secretase activating protein (gsap) gene expression (66). Interestingly previous studies suggested that GATA1 is a transcription repressor for synapse-related genes. In neurological

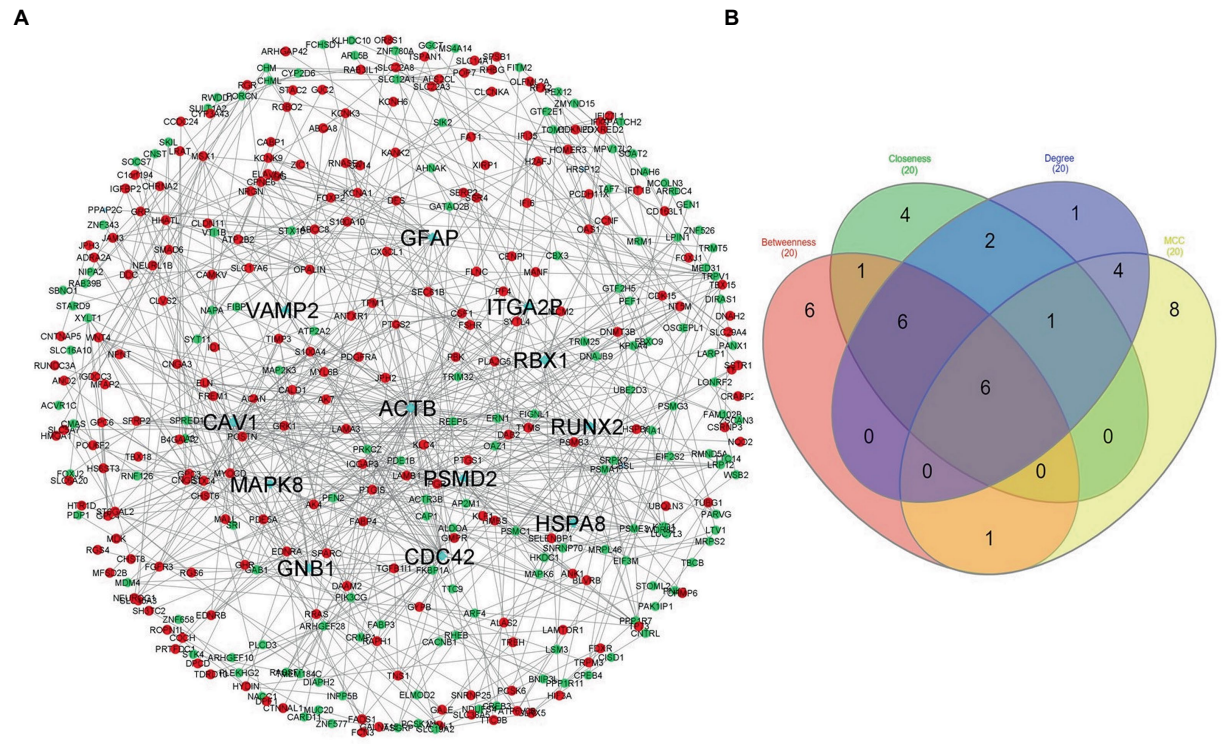


FIGURE 4
Network of protein–protein interaction and detected hub genes (from genes common among AD-PBMC, AD-Tissue, COVID-19-PBMC, and COVID-19-Tissue, module 1). **(A)** The up-regulated and down-regulated genes in red and green colors and hub genes in aqua. **(B)** Venn diagram representing the genes commonly shared among the topological features of MCC, Betweenness, Closeness, and Degree.

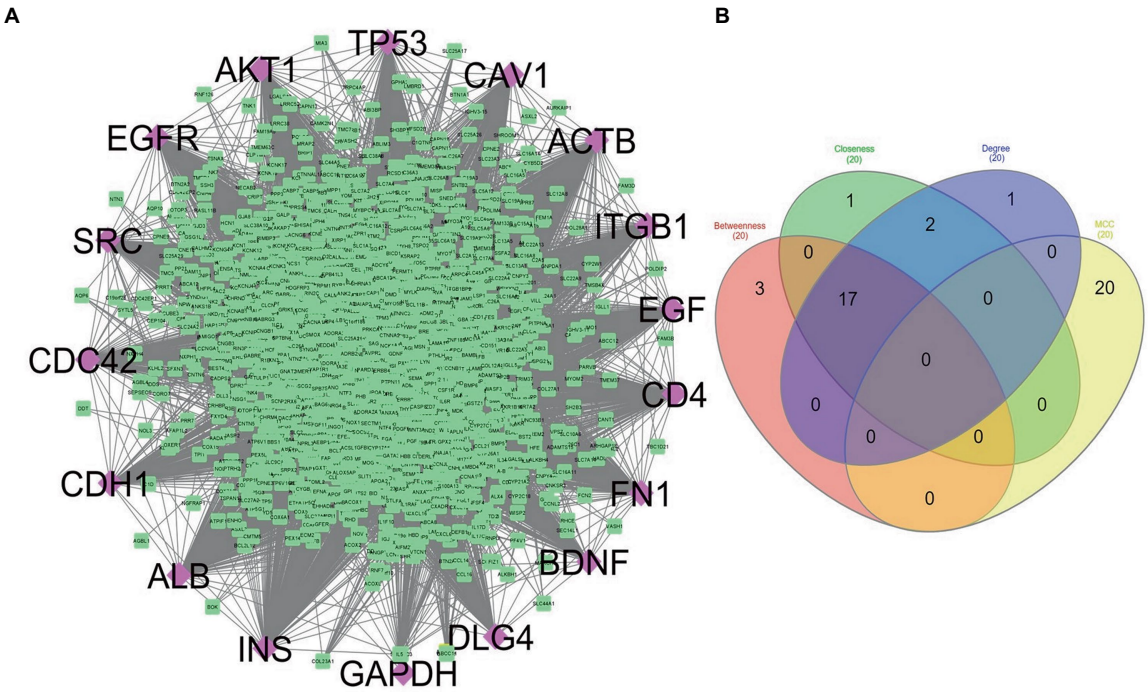


FIGURE 5
(A) Network constructed to represent the common genes shared by ontology terms of Alzheimer's disease and COVID-19 gene ontology terms (module 2). Purple diamonds represent the hub genes of this network. **(B)** Venn diagram showing the genes commonly shared among the topological features of MCC, Betweenness, Closeness, and Degree.

TABLE 4 The top 20 genes from module 1 of (common genes of Alzheimer's disease and COVID-19 tissues and blood) protein–protein interaction network analyzed using four different topological analysis methods such as MCC, Closeness, Betweenness, and Degree through CytoHubba plugin.

S. No	Betweenness	Closeness	Degree	MCC
1.	ACTB	ACTB	ACTB	PSMA1
2.	CDC42	CDC42	CDC42	PSMD2
3.	RUNX2	HSPA8	RUNX2	PSMC1
4.	HSPA8	RUNX2	GFAP	PSME3
5.	GFAP	CAV1	HSPA8	PSMB3
6.	ITGA2B	GFAP	CAV1	ACTB
7.	CAV1	MAPK8	GNB1	RUNX2
8.	SNRNP70	PTGS2	PSMD2	POSTN
9.	RBX1	VAMP2	MAPK8	ELN
10.	GNB1	PRKCZ	ITGA2B	SPARC
11.	VAMP2	PIK3CG	PSMA1	ACAN
12.	PIK3CG	ITGA2B	PTGS2	SPRED1
13.	MAPK8	WNT4	ACAN	TP73
14.	FKBP1A	RBX1	PSME3	TIMP3
15.	MYL6B	ACAN	RBX1	OAZ1
16.	PSMD2	PGR	TRPV1	MAPK6
17.	SLC12A1	PSMD2	PSMB3	GFAP
18.	ABCC8	GNB1	PSMC1	CDC42
19.	OAZ1	TRPV1	KCNA1	HSPA8
20.	HMBS	MAP2K3	VAMP2	SDC4

TABLE 5 The identified top 20 genes from module 2 (common gene ontology terms between Alzheimer's disease and COVID-19) of protein–protein interaction network analyzed using four topological analysis methods such as MCC, Closeness, Betweenness, and Degree through CytoHubba plugin.

S. No	Betweenness	Closeness	Degree	MCC
1.	SRC	STAT3	STAT3	NDUFA6
2.	CFTR	DLG4	CDH1	UQCRH
3.	CAV1	CAV1	BDNF	ATP5MF
4.	ACTB	ACTB	MMP9	NDUFB7
5.	EGF	ERBB2	EGFR	NDUFV2
6.	BDNF	BDNF	ALB	ATP5PO
7.	ALB	ALB	AKT1	NDUFC2
8.	ITGB1	ITGB1	ITGB1	NDUFB6
9.	TP53	TP53	TP53	P13073
10.	INS	INS	INS	UQCRC1
11.	CDC42	CDC42	CD4	ATP5PD
12.	CYCS	EGF	DLG4	COX5A
13.	AKT1	AKT1	ACTB	NDUFB9
14.	CDH1	CDH1	CDC42	ATP5ME
15.	FN1	FN1	FN1	UQCR10
16.	SNCA	SRC	SRC	ATP5PF
17.	EGFR	ESR1	ERBB2	NDUFA12
18.	GAPDH	GAPDH	GAPDH	NDUFA8
19.	DLG4	EGFR	EGF	NDUFV1
20.	CD4	CD4	CAV1	ATP5MG

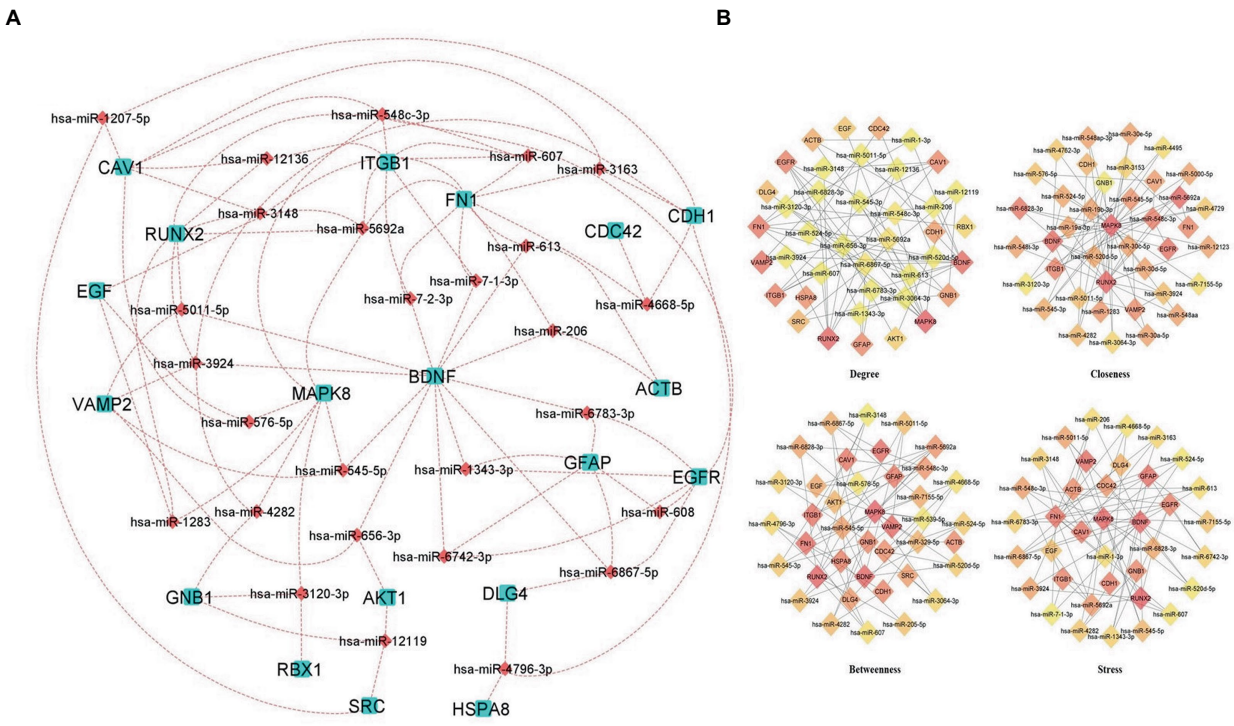


FIGURE 6
Hub Genes-miRNAs Network **(A)** miRNAs interacting with more than three hub genes, aqua color squares representing the hub genes and the maroon color diamonds representing the miRNAs. **(B)** Predicted hub miRNAs using four topological features of CytoHubba including Betweenness, Closeness, Degree, and Stress.

conditions such as AD, NGB may have therapeutic and disease-preventing properties that can be explored experimentally (67).

4.8. Identification of drug-gene interaction

We investigated the drug interactions of hub genes using the DGIdb. A total of 26 hub genes were explored through the drug-gene interactions network. The network result shows that a total of 106 were interacting with the hub genes (Figure 8; Supplementary Table S8). Some of the drugs were already approved by the food and drug administration (FDA) which makes this drug more possible to treat AD and COVID-19 comorbidity. There are potential therapeutics for COVID-19 comorbidities associated with the dysregulation of the proteins.

4.9. Gene set enrichment analysis of hub genes

Functional enrichment analysis results showed that hub genes are involved in several biological functions. We identified hub genes related gene ontology using cluster profiler package in r, and we plotted the significantly enriched terms based on adjusted p value <0.05, as illustrated in Figure 9. There are several pathways were enriched in KEGG analysis including the PI3K-AKT, Neurotrophin, Rap1, Ras, and JAK–STAT signaling pathways, and the top 20 signaling pathways are depicted in Figure 10 (Supplementary Table S9). The gene set enrichment results clearly show that the hub genes are majorly involved in the signaling pathways which might be closely linked to COVID-19 and AD.

5. Discussion

High-throughput sequencing technologies, bioinformatics, and systems biology analysis methods could identify and reveals the changes in the expression level of genes and also assists to identify the potential biomarkers for several diseases importantly neurodegenerative diseases. In this study, the focus is on understanding how AD and COVID-19 disease are related through pathogenetic processes and molecular crosstalks. We followed systems biology approaches including DEGs identification, PPI network construction, hub genes identification, gene set enrichment analysis, and pathway analysis. Also, we explored and identified the regulatory network and drug-genes interaction of the hub genes. To investigate the relationship between AD and COVID-19 we performed gene set enrichment analysis using AD and COVID-19 DEGs discretely. The datasets were further classified into four different groups such as AD-PBMC, AD-Tissue, COVID-19-PBMC, and COVID-19-Tissue. We collected the common DEGs from among the four groups for constructing a Protein–Protein interaction network (module 1). While only 9 DEGs (*HST6, POLR3G, SLC6A20, ITGA2B, HOMER3, GMPR, AGBL1, CRABP2,* and *OLFML2B*) were commonly expressed between these groups. In addition, we performed Gene Set Enrichment Analysis for the DEGs of Alzheimer's disease and

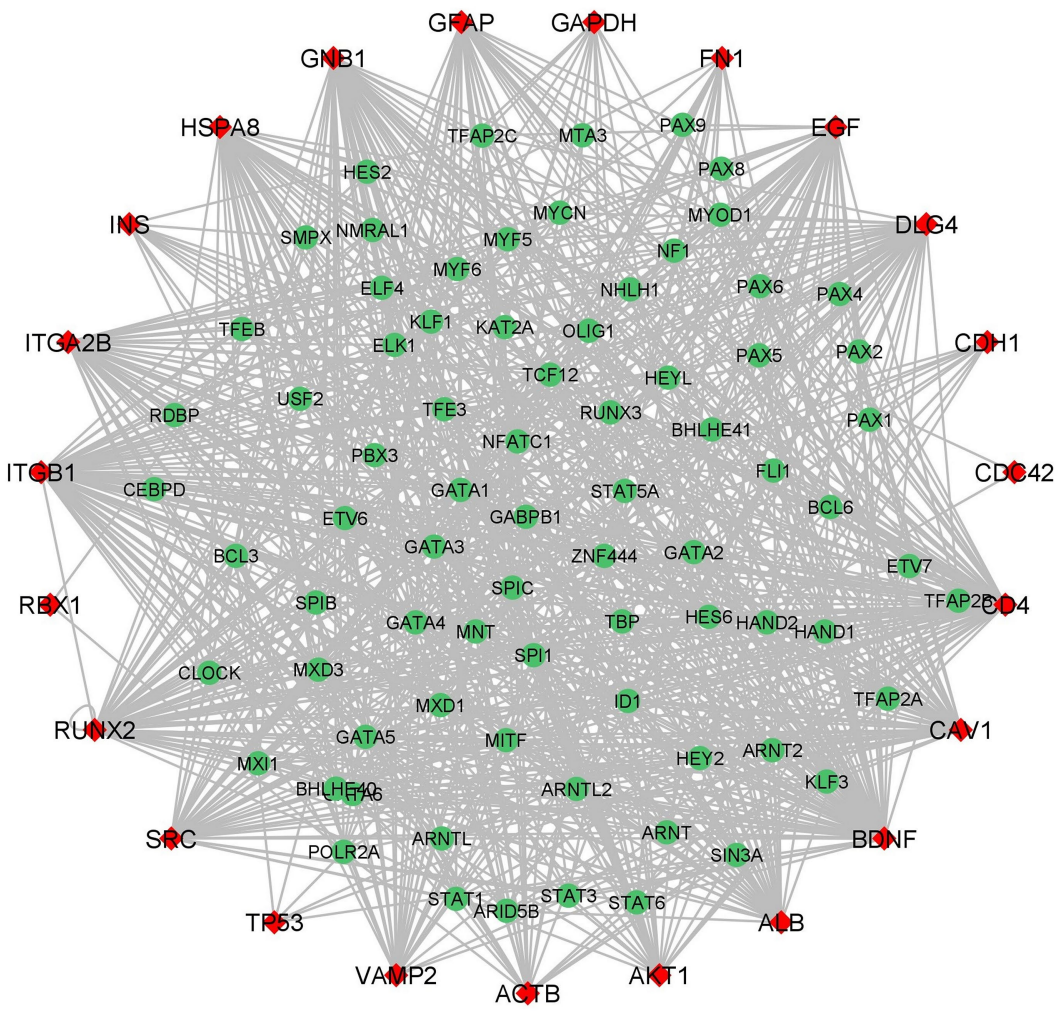


FIGURE 7
Hub Genes-Transcription Factors network (red color diamond designates the hub genes and the green color circulars designate the Transcription Factors). The edges between the two genes indicates the interaction between TFs and hub genes.

SARS-CoV-2 DEGs, then we retrieved the genes with common gene ontology terms for constructing a PPI network (module 2).

The *HST6, ITGA2B, HOMER3*, and *CRABP2* genes have not been reported in AD or COVID-19 related articles. In the extracellular matrix, Olfactomedin Like 2B (OLFML2B) is the olfactomedin domain protein photomedin-2, with an important role in neural crest development and neurogenesis, cell–cell adhesion, and cell cycle regulation. The OLFML2B gene may contribute to the treatment of bladder cancer in the future based on individual prognostic markers (68). Hongde Liu proposed that GMPR's (Guanosine Monophosphate Reductase) GMPR1 is associated with Tau phosphorylation in AD *via* the AMPK (AMP-activated protein kinase) and adenosine receptor pathways (69). A therapeutic strategy of inhibiting GMPR1 with lumacaftor has been proposed to treat AD based on the elevated expression of GMPR in this disease. Wei Dong et al. explored the common initiative molecular pathways in AD and ischemic stroke and they found that AGBL1 is a common risk gene (70). SLC6A20 appears to be a novel regulator of glycine and proline levels in the brain according to the research of Mihyun Bae. Further, pharmacologically inhibiting SLC6A20 may contribute to the treatment of brain disorders *via* an increase in glycine levels in the brain and N-Methyl-D-Aspartate receptors (NMDAR) activity (71). Some important biological processes, including spliceosome genes, were dysregulated by POLR3B genes. A number of transcription factors, including FOXC2 and GATA1, play a role in neuronal dysfunction and intellectual disability, which are affected by impaired protein synthesis and splicing (72).

miRNAs as biomarkers: miRNA subsets have shown clinical relevance as biomarkers according to a growing number of reports. There are emerging miRNA therapeutics that are used to determine the presence of pathology, as well as the progression, genetic links, and stage of the disease. miRNAs have been translated into clinical medicine faster than ever because of the bioinformatic approach to identifying miRNA-binding sites and their related biological pathways in target genes, as well as the expanding availability of *in vitro* and *in vivo* preclinical research models (73). The miRNA helps to understand the development and progression of COVID-19 and AD comorbidity. In the miRNAs network *BDNF, MAPK8, ITGB1, FN1, EGFR*, and *RUNX2* hub genes are associated with most of the miRNAs. The co-expression network revealed that hsa-miR-6,867-5P regulates

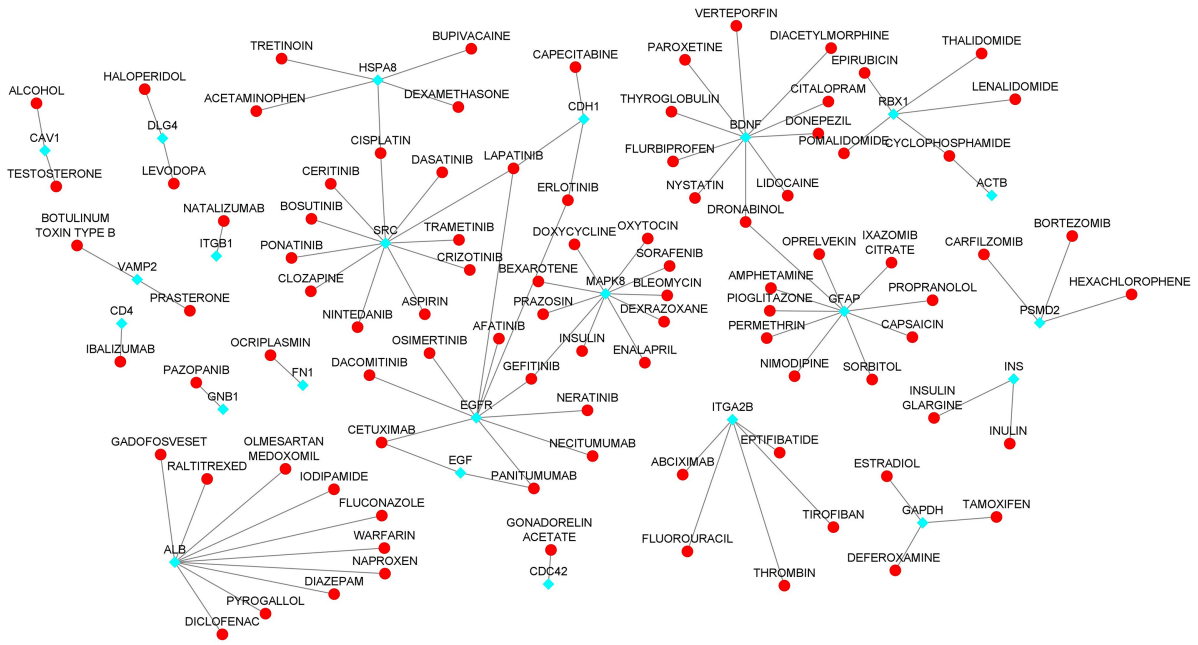


FIGURE 8
Drug-Hub Gene Network (aqua color indicating the hub genes and red color indicating the drugs).

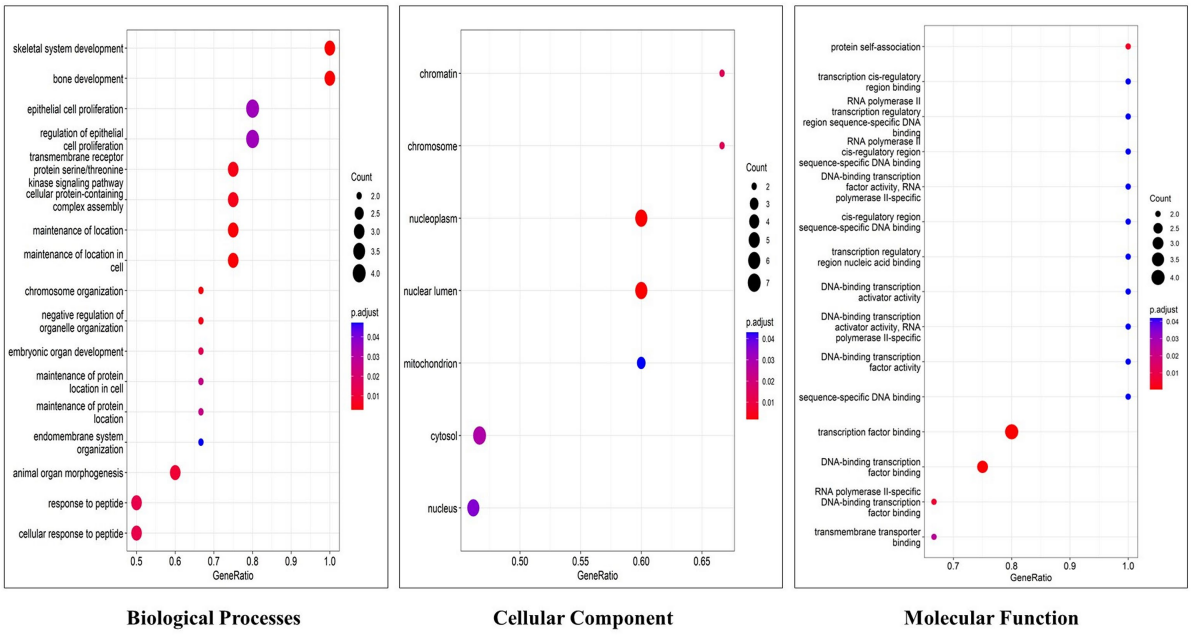


FIGURE 9
Top 20 gene ontology terms of hub genes (The *x*-axis label represents the gene ratio and the *y*-axis label represents gene ontology terms).

EGFR, DLG4, GFAP, BDNF and hsa-miR-548C-3p regulates EGFR, MAPK8, ITGB1, CAV1 and hsa-miR-5692a regulates ITGB1, FN1, MAPK8, EGF, RUNX2. Research suggested that hsa-miR6867-5P and 6,867-5P were associated with platelet apoptosis and adhesion in an autoimmune disease like immune thrombocytopenia (74). Recent studies exhibited that hypothalamic miRNAs including miR-548C-3p are potential contributors to different neurodegenerative diseases, also this author identified 29 novel hypothalamic MicroRNAs as a propitious therapeutic regimen for SARS-CoV-2 by regulating ACE2 and TMPRSS2 expression (75). Cosin et al. studied a multiple linear regression model for predicting amyloid beta levels in Cerebrospinal fluid, for this they used four validated miRNAs for AD including miR-545-5p, miR-142-3p, miR-34a-5p, and miR-15b-5p. The results revealed that miR-34a-5p is the best-predicting miRNA for amyloid

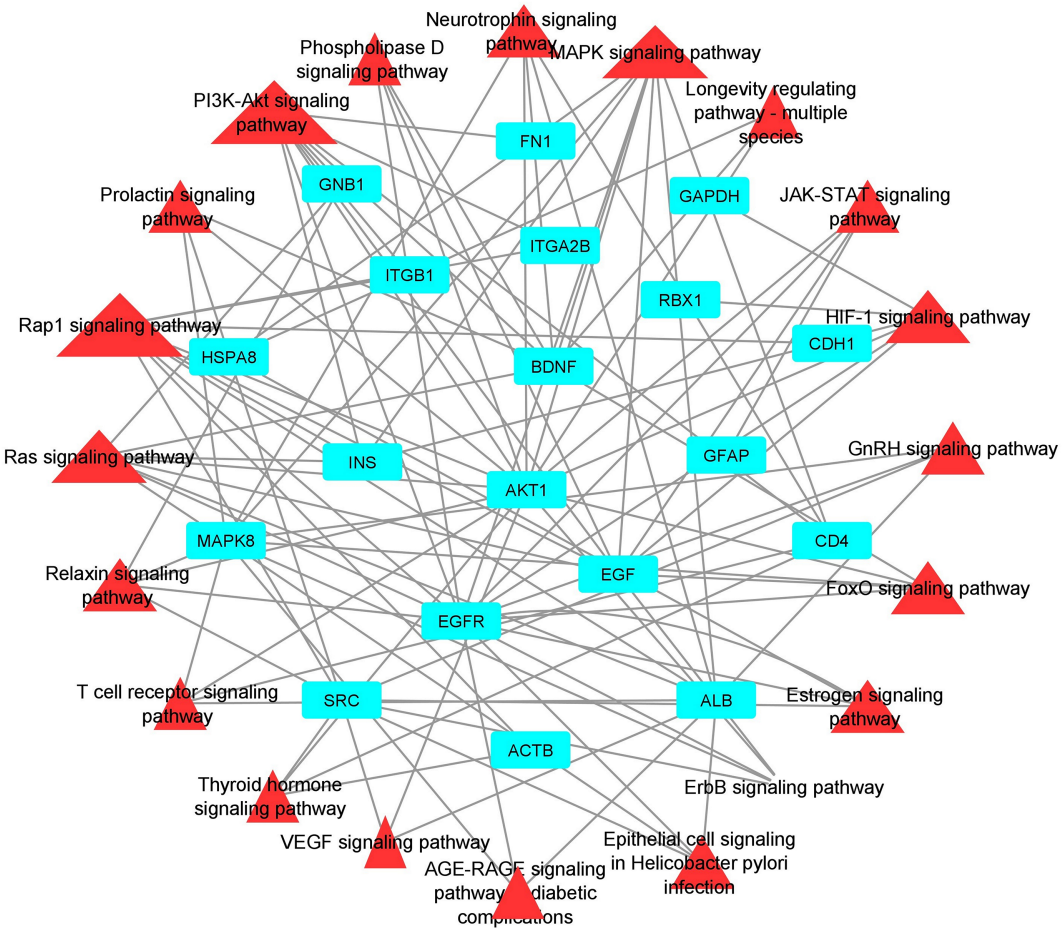


FIGURE 10
Pathway-Hub Gene Network (aqua color indicating the hub genes and the red color indicating the signal pathways).

beta levels in cerebrospinal fluid (Cosín-Tomás et al., 2017). The miR-545-3p, and miR-34a-5p could be potential biomarkers for the early detection of AD (Cosín-Tomás et al., 2017).

To illustrate the mechanisms of hub genes we performed enrichment analysis including GO and pathway analysis. We found various cell signaling pathways are enriched including RAP1, MAPK, PI3K-AKT, RAS, and HIF-1 signaling pathways, etc. The signaling pathway of RAP1 was found to be a crucial regulator of cellular functions such as the formation and control of cell adhesion and junction and, also plays a major role during cell invasion and metastasis in different cancers (76). MAPK pathway responds to numerous extracellular stimulations including inflammatory cytokines, stress, and viral infection. Furthermore, COVID-19 infection activated MAPK and the downstream signaling possibly leading to cell death. Intense work is in progress to develop a compound to target MAPK pathways to treat neurodegenerative and inflammatory diseases (77). Proliferation, apoptosis, and angiogenesis, the Renin-angiotensin signaling pathway (RAS) has been shown to play a role in tumorigenesis through complex interactions (78). Krishna Sriram et al. reported that RAS has a great tendency to cause comorbidities and mortality and they proposed a model to predict effective drugs to target RAS (79). RAS–ERK signaling induces amyloid precursor protein and tau protein hyperphosphorylation which are enhanced in AD brains, and inhibition of RAS-MAPK activation prevents tau and amyloid precursor protein hyperphosphorylation (80). HIF-1α (hypoxia-inducible factor) plays a crucial role in inflammatory responses, regulating metabolic pathways and regulating the aging process. Dysregulations of the pathway HIF-1α lead to several diseases including cardiovascular disease, cancer, and AD. HIF-1α is a key activator for COVID-19 and inflammatory responses and it could be a therapeutic target for virus-induced inflammatory diseases and COVID-19 (81). As part of the immune response and virus entry into the cell, Phosphatidylinositol 3-kinase (PI3K)/AKT signaling plays a significant role also this pathway is involved in several aspects of neurological disease development (82). Patients with COVID-19 have been found to have an increased risk of lung tissue fibrosis following activation of the PI3K-AKT signaling pathway (83). Cancers and diabetes are associated with excessive activation of the PI3K-AKT pathway also cardiovascular diseases and neurological conditions such as AD and PD might also be affected by the deregulation of the pathway (84). Enriched BP of hub genes has primarily participated in the cellular response to peptides, animal organ morphogenesis, endomembrane system organization, maintenance of protein location, and embryonic organ development. The top enriched terms of CC were nucleus, cytosol, mitochondrion, nuclear lumen, and nucleoplasm. The top five

terms in MF were mainly enriched transmembrane transporter binding, RNA polymerase II-specific DNA-binding transcription factor binding, DNA binding transcription factor binding, sequence-specific DNA binding and transcription factor binding. We constructed a drug-gene network for hub genes and investigated the relationship between the chemical and the disease. Through this drug-gene network, we found several drugs including diacetylmorphine, donepezil, dronabinol, levodopa, haloperi, deferoxamine, raltitrexed, diazepam, and warfarin. These drugs are already reported for treating AD and Parkinson's disease (85–89). Recent studies reported repurposing of CNS drugs are potential to treat SARS-CoV-2-infected individuals (90). We have found an interaction between DEGs-miRNAs-TFs which are plays key roles in the pathogenesis of neurological disorders.

It is necessary to acknowledge that the study has some limitations because it only relies on bioinformatics and network biology. One of the limitations of the study is the potential confounding effects associated with the variations in transcriptome profiles from different tissues (brain vs. blood). Also selecting overlapping DEGs from separate analyses of tissues and blood samples may not completely eliminate the confounding effect of sample variation. Additionally, the large number of DEGs identified in the study may have caused a potential for false positive results. While we attempted to address these issues by performing additional analyses including hub genes and pathway analysis.

6. Conclusion

The present study aims to understand the molecular crosstalk between COVID-19 and Alzheimer's Disease, including discovering the gene expression signatures, TFs, Drug-gene interaction, miRNAs associations, and dysregulated molecular pathways. As a result of integrated analyses of microarrays and transcriptomics of PBMC cells and tissue cells, we were able to identify AD and COVID-19 DEGs. Through PPI network analysis twenty-three (*AKT1, ALB, BDNF, CAV1, CD4, CDC42, CDH1, DLG4, EGF, EGFR, FN1, GAPDH, INS, ITGB1, ACTB, SRC, TP53, RUNX2, HSPA8, PSMD2, GFAP, VAMP2, MAPK8, GNB1, RBX1, ITGA2B*) hub genes were identified. Transcription factor network analyses revealed that several TFs play a crucial role in post-transcriptional and transcriptional regulators of the differentially expressed genes. The identified shared pathways between AD and COVID-19 provide there are several similar underlying mechanisms play in both diseases. Our findings could lead to identifying a potential biomarker to predict the highest risk of neurological complications with COVID-19. Also, the identified transcription factor might be a potential therapeutic drug target for both diseases.

Data availability statement

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

Author contributions

TP analyzed the data and wrote the manuscript. SS conceptualized and designed the work, revised, and edited the manuscript. All authors contributed to the article and approved the submitted version.

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.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1151046/full#supplementary-material

References

1. Hu B, Guo H, Zhou P, Shi ZL. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol*. (2020) 19:141–54. doi: 10.1038/s41579-020-00459-7

2. Holmes EC, Goldstein SA, Rasmussen AL, Robertson DL, Crits-Christoph A, Wertheim JO, et al. The origins of SARS-CoV-2: a critical review. *Cells*. (2021) 184:4848–56. doi: 10.1016/j.cell.2021.08.017

3. Li MY, Li L, Zhang Y, Wang XS. Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues. *Infect Dis Poverty*. (2020) 9:45. doi: 10.1186/s40249-020-00662-x

4. Ferini-Strambi L, Salsone M. COVID-19 and neurological disorders: are neurodegenerative or neuroimmunological diseases more vulnerable? *J Neurol*. (2021) 268:409–19. doi: 10.1007/s00415-020-10070-8

5. Wouk J, Rechenchoski DZ, Rodrigues BCD, Ribelato EV, Faccin-Galhardi LC. Viral infections and their relationship to neurological disorders. *Arch Virol*. (2021) 166:733–53. doi: 10.1007/s00705-021-04959-6

6. Lukiw WJ, Pogue A, Hill JM. SARS-CoV-2 infectivity and neurological targets in the brain. *Cell Mol Neurobiol*. (2022) 42:217–24. doi: 10.1007/s10571-020-00947-7

7. Ahmed SSSJ, Paramasivam P, Kamath M, Sharma A, Rome S, Murugesan R. Genetic exchange of lung-derived exosome to brain causing neuronal changes on COVID-19 infection. *Mol Neurobiol*. (2021) 58:5356–68. doi: 10.1007/s12035-021-02485-9

8. Prasad K, Yousef AlOmar S, Awad Alqahtani SM, Zubbair Malik M, Kumar V. Brain disease network analysis to elucidate the neurological manifestations of COVID-19. *Mol Neurobiol*. (2021) 58:1875–93. doi: 10.1007/s12035-020-02266-w

9. Douaud G, Lee S, Alfaro-Almagro F, Arthofer C, Wang C, McCarthy P, et al. SARS-CoV-2 is associated with changes in brain structure in UK biobank. *Nature*. (2022) 604:697–707. doi: 10.1038/s41586-022-04569-5

10. Tavares-Júnior JWL, de Souza ACC, Borges JWP, Oliveira DN, Siqueira-Neto JI, Sobreira-Neto MA, et al. COVID-19 associated cognitive impairment: a systematic review. *Cortex*. (2022) 152:77–97. doi: 10.1016/j.cortex.2022.04.006

11. Alshebri MS, Alshouimi RA, Alhumidi HA, Alshaya AI. Neurological complications of SARS-CoV, MERS-CoV, and COVID-19. *SN Compr Clin Med*. (2020) 2:2037–47. doi: 10.1007/s42399-020-00589-2

12. Ellul MA, Benjamin L, Singh B, Lant S, Michael BD, Easton A, et al. Neurological associations of COVID-19. *Lancet Neurol*. (2020) 19:767–83. doi: 10.1016/S1474-4422(20)30221-0

13. Gordon MN, Heneka MT, le Page LM, Limberger C, Morgan D, Tenner AJ, et al. Impact of COVID-19 on the onset and progression of Alzheimer's disease and related dementias: a roadmap for future research. *Alzheimer's Dementia*. (2022) 18:1038–46. doi: 10.1002/alz.12488

14. Rhodus EK, Aisen P, Grill JD, Rentz DM, Petersen RC, Sperling RA, et al. Alzheimer's disease clinical trial research adaptation following COVID-19 pandemic onset: National sample of Alzheimer's clinical trial consortium sites. *J Prev Alzheimers Dis*. (2022) 9:665–71. doi: 10.14283/jpad.2022.79

15. Snider BJ, Holtzman DM. Effects of COVID-19 on preclinical and clinical research in neurology: examples from research on neurodegeneration and Alzheimer's disease. *Neuron*. (2021) 109:3199–202. doi: 10.1016/j.neuron.2021.08.026

16. Ciaccio M, Lo Sasso B, Scazzone C, Gambino CM, Ciaccio AM, Bivona G, et al. COVID-19 and Alzheimer's disease. *Brain Sci*. (2021) 11:1–10. doi: 10.3390/brainsci11030305

17. Liu L, Ni SY, Yan W, Lu QD, Zhao YM, Xu YY, et al. Mental and neurological disorders and risk of COVID-19 susceptibility, illness severity and mortality: a systematic review, meta-analysis and call for action. *E Clin Med*. (2021) 1:101111. doi: 10.1016/j.eclinm.2021.101111

18. Zhang XX, Tian Y, Wang ZT, Ma YH, Tan L, Yu JT. The epidemiology of Alzheimer's disease modifiable risk factors and prevention. *J Prevent Alzheimer's Dis*. (2021) 3:1–9. doi: 10.14283/jpad.2021.15

19. Khalifa N, Ben TD, Marinangeli C, Depuydt M, Courtoy PJ, Christophe RJ, et al. Structural features of the KPI domain control APP dimerization, trafficking, and processing. *FASEB J*. (2012) 26:855–67. doi: 10.1096/fj.11-190207

20. Asionowski MAJ, Aass CHH, Ahrenholz FALKF. Constitutive and regulated secretase cleavage of Alzheimer' s amyloid precursor protein by a disintegrin metalloprotease. *Proc Natl Acad Sci U S A*. (1999) 96:3922–7. doi: 10.1073/pnas.96.7.3922

21. Premkumar T, Sajitha LS. Molecular mechanisms of emerging therapeutic targets in Alzheimer's disease: a systematic review. *Neurochem J*. (2022) 16:443–55. doi: 10.1134/S1819712422040183

22. Lichtenthaler SF. Alpha-secretase in Alzheimer's disease: molecular identity, regulation and therapeutic potential. *J Neurochem*. (2011) 116:10–21. doi: 10.1111/j.1471-4159.2010.07081.x

23. Vassar R. BACE1: the β-secreiase enzyme in Alzheimer's disease. *J Mol Neurosci*. (2004) 23:105–14. doi: 10.1385/JMN:23:1-2:105

24. Krishnaswamy S, Verdile G, Groth D, Kanyenda L. The structure and function of Alzheimer' s gamma secretase enzyme complex. *Crit Rev Clin Lab Sci*. (2009) 46:282–301. doi: 10.3109/10408360903335821

25. Kolarova M, García-Sierra F, Bartos A, Ricny J, Ripova D. Structure and pathology of tau protein in Alzheimer disease. *Int J Alzheimers Dis*. (2012) 2012:1–13. doi: 10.1155/2012/731526

26. Mandelkow EM, Mandelkow E. Biochemistry and cell biology of tau protein in neurofibrillary degeneration. *Cold Spring Harb Perspect Biol*. (2012) 4:1–25. doi: 10.1101/cshperspect.a006247

27. Pizzarelli R, Pediconi N, di Angelantonio S. Molecular imaging of tau protein: new insights and future directions. *Front Mol Neurosci*. (2020) 13:1–6. doi: 10.3389/fnmol.2020.586169

28. Sait A, Angeli C, Doig AJ, Day PJR. Viral involvement in Alzheimer's disease. *ACS Chem Neurosci*. (2021) 12:1049–60. doi: 10.1021/acschemneuro.0c00719

29. Vigasova D, Nemergut M, Liskova B, Damborsky J. Multi-pathogen infections and Alzheimer's disease. *Microbial Cell Fact*. (2021) 20:25. doi: 10.1186/s12934-021-01520-7

30. Hardan L, Filtchev D, Kassem R, Bourgi R, Lukomska-Szymanska M, Tarhini H, et al. Covid-19 and alzheimer's disease: A literature review. *Medicine*. (2021) 57:159. doi: 10.3390/medicina57111159

31. Jha PK, Vijay A, Halu A, Uchida S, Aikawa M. Gene expression profiling reveals the shared and distinct transcriptional signatures in human lung epithelial cells infected with SARS-CoV-2, MERS-CoV, or SARS-CoV: potential implications in cardiovascular complications of COVID-19. *Front Cardiovasc Med*. (2021) 7:7. doi: 10.3389/fcvm.2020.623012

32. Jha PK, Vijay A, Sahu A, Ashraf MZ. Comprehensive gene expression meta-analysis and integrated bioinformatic approaches reveal shared signatures between thrombosis and myeloproliferative disorders. *Sci Rep*. (2016) 6:1–13. doi: 10.1038/srep37099

33. Rahman MH, Peng S, Hu X, Chen C, Uddin S, Quinn JMW, et al. Bioinformatics methodologies to identify interactions between type 2 diabetes and neurological comorbidities. *IEEE Access*. (2019) 7:183948–70. doi: 10.1109/ACCESS.2019.2960037

34. del Prete E, Facchiano A, Liò P. Bioinformatics methodologies for coeliac disease and its comorbidities. *Brief Bioinform*. (2018) 21:355–67. doi: 10.1093/bib/bby109

35. Diaz-Beltran L, Cano C, Wall DP, Esteban FJ. Systems biology as a comparative approach to understand complex gene expression in neurological diseases. Behav Sci, 3. (2013). 253–273.

36. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets - update. *Nucleic Acids Res*. (2013) 41:D991–5. doi: 10.1093/nar/gks1193

37. Sean D, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics*. (2007) 23:1846–7. doi: 10.1093/bioinformatics/btm254

38. Sean D, Meltzer PS. GEOquery: a bridge between the gene expression omnibus (GEO) and BioConductor. *Bioinformatics*. (2007) 23:1846–7. doi: 10.1093/bioinformatics/btm254

39. 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

40. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc B*. (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

41. Wang M, Wang L, Wu S, Zhou D, Wang X. Identification of key genes and prognostic value analysis in hepatocellular carcinoma by integrated bioinformatics analysis. *Int J Genomics*. (2019) 2019:3518378. doi: 10.1155/2019/3518378

42. Dalman MR, Deeter A, Nimishakavi G, Duan ZH. Fold change and p-value cutoffs significantly alter microarray interpretations. *BMC Bioinformatics*. (2012) 13:1–4. doi: 10.1186/1471-2105-13-S2-S11

43. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. (2010) 11:1–12. doi: 10.1186/gb-2010-11-10-r106

44. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. (2014) 15:1–21. doi: 10.1186/s13059-014-0550-8

45. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. (2012) 31:46–53. doi: 10.1038/nbt.2450

46. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol*. (2014) 24, 32:896–902. doi: 10.1038/nbt.2931

47. Ghahramani N, Shodja J, Rafat SA, Panahi B, Hasanpur K. Integrative systems biology analysis elucidates mastitis disease underlying functional modules in dairy cattle. *Front Genet*. (2021) 12:12. doi: 10.3389/fgene.2021.712306

48. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. (2022) 50:W216–21. doi: 10.1093/nar/gkac194

49. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. (2013) 41:D808–15. doi: 10.1093/nar/gks1094

50. Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol*. (2014) 8:S11. doi: 10.1186/1752-0509-8-S4-S11

51. Yu D, Lim J, Wang X, Liang F, Xiao G. Enhanced construction of gene regulatory networks using hub gene information. *BMC Bioinformat*. (2017) 18:1576. doi: 10.1186/s12859-017-1576-1

52. Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The human transcription factors. *Cell*. (2018) 172:650–65. doi: 10.1016/j.cell.2018.01.029

53. Janky R, Verfaillie A, Imrichová H, van de Sande B, Standaert L, Christiaens V, et al. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS Comput Biol*. (2014) 10:e1003731. doi: 10.1371/journal.pcbi.1003731

54. Abbas SZ, Qadir MI, Muhammad SA. Systems-level differential gene expression analysis reveals new genetic variants of oral cancer. *Sci Rep*. (2020) 10:14667. doi: 10.1038/s41598-020-71346-7

55. Ishrat R, Ahmed MM, Tazyeen S, Alam A, Farooqui A, Ali R, et al. In Silico integrative approach revealed key MicroRNAs and associated target genes in Cardiorenal syndrome. *Bioinform Biol Insights*. (2021) 15:7396. doi: 10.1177/11779322211027396

56. Qiu X, Lin J, Liang B, Chen Y, Liu G, Zheng J. Identification of hub genes and MicroRNAs associated with idiopathic pulmonary arterial hypertension by integrated bioinformatics analyses. *Front Genet*. (2021) 12:544. doi: 10.3389/fgene.2021.636934

57. Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the drug–gene interaction database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res*. (2021) 49:D1144–51. doi: 10.1093/nar/gkaa1084

58. 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

59. 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

60. Golkar-Narenji A, Antosik P, Nolin S, Rucinski M, Jopek K, Zok A, et al. Gene ontology groups and signaling pathways regulating the process of avian satellite cell differentiation. *Genes (Basel)*. (2022) 13:242. doi: 10.3390/genes13020242

61. Alam MS, Sultana A, Reza MS, Amanullah M, Kabir SR, Mollah MNH. Integrated bioinformatics and statistical approaches to explore molecular biomarkers for breast cancer diagnosis, prognosis and therapies. *PLoS One*. (2022) 17:e0268967. doi: 10.1371/journal.pone.0268967

62. Vandesompele J, de Preter K, Ilip P, Poppe B, van Roy N, de Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. (2002) 3:34. doi: 10.1186/gb-2002-3-7-research0034

63. Talwar P, Silla Y, Grover S, Gupta M, Agarwal R, Kushwaha S, et al. Genomic convergence and network analysis approach to identify candidate genes in Alzheimer's disease. *BMC Genomics*. (2014) 15:199. doi: 10.1186/1471-2164-15-199

64. Marinić M, Mika K, Chigurupati S, Lynch VJ. Evolutionary transcriptomics implicates hand2 in the origins of implantation and regulation of gestation length. *elife*. (2021) 10:1–52. doi: 10.7554/eLife.61257

65. Gao J, Chen YH, Peterson LAC. GATA family transcriptional factors: Emerging suspects in hematologic disorders. *Exp Hematol Oncol*. (2015) 4:24. doi: 10.1186/s40164-015-0024-z

66. Chu J, Wisniewski T, Praticò D. GATA1-mediated transcriptional regulation of the γ-secretase activating protein increases Aβ formation in down syndrome. *Ann Neurol*. (2016) 79:138–43. doi: 10.1002/ana.24540

67. Tam KT, Chan PK, Zhang W, Law PP, Tian Z, Chan GCF, et al. Identification of a novel distal regulatory element of the human Neuroglobin gene by the chromosome conformation capture approach. *Nucleic Acids Res*. (2017) 45:115–26. doi: 10.1093/nar/gkw820

68. Lin J, Xu X, Li T, Yao J, Yu M, Zhu Y, et al. OLFML2B is a robust prognostic biomarker in bladder cancer through genome-wide screening: a study based on seven cohorts. *Front Oncol*. (2021) 15. doi: 10.3389/fonc.2021.650678

69. Liu H, Luo K, Luo D. Guanosine monophosphate reductase 1 is a potential therapeutic target for Alzheimer's disease. *Sci Rep*. (2018) 8:21256. doi: 10.1038/s41598-018-21256-6

70. Dong W, Huang Y. Is cerebral vascular pathology a bystander of Alzheimer's disease? Evidence from a genetic perspective. *Alzheimers Dement*. (2021) 1:e050716. doi: 10.1002/alz.050716

71. Bae M, Roh JD, Kim Y, Kim SS, Han HM, Yang E, et al. SLC6A20 transporter: a novel regulator of brain glycine homeostasis and NMDAR function. *EMBO Mol Med*. (2021) 13:e12632. doi: 10.15252/emmm.202012632

72. Saghi M, InanlooRahatloo K, Alavi A, Kahrizi K, Najmabadi H. Intellectual disability associated with craniofacial dysmorphism due to POLR3B mutation and defect in spliceosomal machinery. *BMC Med Genet*. (2022) 15:89. doi: 10.1186/s12920-022-01237-5

73. Hanna J, Hossain GS, Kocerha J. The potential for microRNA therapeutics and clinical research. *Front Genet*. (2019) 10:478. doi: 10.3389/fgene.2019.00478

74. Deng G, Yu S, He Y, Sun T, Liang W, Yu L, et al. MicroRNA profiling of platelets from immune thrombocytopeniaand target gene prediction. *Mol Med Rep*. (2017) 16:2835–43. doi: 10.3892/mmr.2017.6901

75. Mukhopadhyay D, Mussa BM. Identification of novel hypothalamic micrornas as promising therapeutics for sars-cov-2 by regulating ace2 and tmprss2 expression: an in silico analysis. *Brain Sci*. (2020) 10:1–11. doi: 10.3390/brainsci10100666

76. Zhang YL, Wang RC, Cheng K, Ring BZ, Su L. Roles of Rap1 signaling in tumor cell migration and invasion. *Cancer Biol Med Cancer Biol Med*. (2017) 14:90–9. doi: 10.20892/j.issn.2095-3941.2016.0086

77. Mohanta TK, Sharma N, Arina P, Defilippi P. Molecular insights into the MAPK Cascade during viral infection: Potential crosstalk between HCQ and HCQ analogues. *BioMed Res Int*. (2020) 2020:8827752. doi: 10.1155/2020/8827752

78. Iftikhar A, Islam M, Shepherd S, Jones S, Ellis I. Is RAS the link between COVID-19 and increased stress in head and neck cancer patients? *Front Cell Dev Biol*. (2021) 9:999. doi: 10.3389/fcell.2021.714999

79. Sriram K, Loomba R, Insel PA. Targeting the renin-angiotensin signaling pathway in COVID-19: unanswered questions, opportunities, and challenges. *Proc Natl Acad Sci U S A*. (2020) 117:29274–82. doi: 10.1073/pnas.2009875117

80. Kirouac L, Rajic AJ, Cribbs DH, Padmanabhan J. Activation of Ras-ERK signaling and GSK-3 by amyloid precursor protein and amyloid beta facilitates neurodegeneration in Alzheimer's disease. *eNeuro*. (2017) 4:ENEURO.0149–16.2017. doi: 10.1523/ENEURO.0149-16.2017

81. Tian M, Liu W, Li X, Zhao P, Shereen MA, Zhu C, et al. HIF-1α promotes SARS-CoV-2 infection and aggravates inflammatory responses to COVID-19. *Signal Transduct Target Ther*. (2021) 6:308. doi: 10.1038/s41392-021-00726-w

82. Rahman MH, Rana HK, Peng S, Kibria MG, Islam MZ, Mahmud SMH, et al. Bioinformatics and system biology approaches to identify pathophysiological impact of COVID-19 to the progression and severity of neurological diseases. *Comput Biol Med*. (2021) 1:138.

83. Khezri MR. PI3K/AKT signaling pathway: A possible target for adjuvant therapy in COVID-19. *Human Cell*. (2021) 34:700–1. doi: 10.1007/s13577-021-00484-5

84. Long HZ, Cheng Y, Zhou ZW, Luo HY, Wen DD, Gao LC. PI3K/AKT signal pathway: A target of natural products in the prevention and treatment of Alzheimer's disease and Parkinson's disease. *Front Pharmacol*. (2021) 12. doi: 10.3389/fphar.2021.648636

85. Ornstein TJ, Iddon JL, Baldacchino AM, Sahakian BJ, London M, Everitt BJ, et al. Profiles of cognitive dysfunction in chronic amphetamine and heroin abusers. *Neuropsychopharmacology*. (2000) 23:113–26. doi: 10.1016/S0893-133X(00)00097-X

86. Knowles J. Clinical impact review donepezil in Alzheimer's disease: An evidence-based review of its impact on clinical and economic outcomes. *Core Evid*. (2006) 1:195–219.

87. Aso E, Andrés-Benito P, Ferrer I. Delineating the efficacy of a cannabis-based medicine at advanced stages of dementia in a murine model. *J Alzheimers Dis*. (2016) 54:903–12. doi: 10.3233/JAD-160533

88. Tipples K, Kolluri RB, Raouf S. Encephalopathy secondary to capecitabine chemotherapy: a case report and discussion. *J Oncol Pharm Pract*. (2009) 15:237–9. doi: 10.1177/1078155209102511

89. Venti A, Giordano T, Eder P, Bush AI, Lahiri DK, Greig NH, et al. The integrated role of desferrioxamine and phenserine targeted to an iron-responsive element in the APP-mRNA 5′-untranslated region. *Ann N Y Acad Sci*. (2004) 1035:34–48. doi: 10.1196/annals.1332.003

90. Hashimoto K. Repurposing of CNS drugs to treat COVID-19 infection: targeting the sigma-1 receptor. *Europ Arch Psychiatry Clin Neurosci*. (2021) 271:249–58. doi: 10.1007/s00406-020-01231-x

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 09 June 2023
DOI 10.3389/fmed.2023.1108358

OPEN ACCESS

EDITED BY
Balu Kamaraj,
Imam Abdulrahman Bin Faisal University,
Saudi Arabia

REVIEWED BY
Animesh A. Sinha,
University at Buffalo, United States
Majji Rambabu,
REVA University, India

*CORRESPONDENCE
Baskaran Reena Rajkumari
b.reenarajkumari@vit.ac.in

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

RECEIVED 25 November 2022
ACCEPTED 22 March 2023
PUBLISHED 09 June 2023

CITATION
Premanand A and Reena Rajkumari B (2023)
Bioinformatic analysis of gene expression data
reveals Src family protein tyrosine kinases as
key players in androgenetic alopecia.
Front. Med. 10:1108358.
doi: 10.3389/fmed.2023.1108358

Bioinformatic analysis of gene expression data reveals Src family protein tyrosine kinases as key players in androgenetic alopecia

Adaikalasamy Premanand and Baskaran Reena Rajkumari *

Department of Integrative Biology, School of Bio Sciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Introduction: Androgenetic alopecia (AGA) is a common progressive scalp hair loss disorder that leads to baldness. This study aimed to identify core genes and pathways involved in premature AGA through an *in-silico* approach.

Methods: Gene expression data (GSE90594) from vertex scalps of men with premature AGA and men without pattern hair loss was downloaded from the Gene Expression Omnibus database. Differentially expressed genes (DEGs) between the bald and haired samples were identified using the *limma* package in R. Gene ontology and Reactome pathway enrichment analyses were conducted separately for the up-regulated and down-regulated genes. The DEGs were annotated with the AGA risk loci, and motif analysis in the promoters of the DEGs was also carried out. STRING Protein-protein interaction (PPI) and Reactome Functional Interaction (FI) networks were constructed using the DEGs, and the networks were analyzed to identify hub genes that play could play crucial roles in AGA pathogenesis.

Results and discussion: The *in-silico* study revealed that genes involved in the structural makeup of the skin epidermis, hair follicle development, and hair cycle are down-regulated, while genes associated with the innate and adaptive immune systems, cytokine signaling, and interferon signaling pathways are up-regulated in the balding scalps of AGA. The PPI and FI network analyses identified 25 hub genes namely CTNNB1, EGF, GNAI3, NRAS, BTK, ESR1, HCK, ITGB7, LCK, LCP2, LYN, PDGFRB, PIK3CD, PTPN6, RAC2, SPI1, STAT3, STAT5A, VAV1, PSMB8, HLA-A, HLA-F, HLA-E, IRF4, and ITGAM that play crucial roles in AGA pathogenesis. The study also implicates that Src family tyrosine kinase genes such as LCK, and LYN in the up-regulation of the inflammatory process in the balding scalps of AGA highlighting their potential as therapeutic targets for future investigations.

KEYWORDS

androgenetic alopecia, differential gene expression analysis, reactome functional interaction network, STRING protein-protein interaction network, gene ontology, motif analysis, Wnt/β-catenin signaling, Src family protein tyrosine kinases

Introduction

Androgenetic alopecia (AGA) is a complex genetic disorder characterized by a progressive loss of scalp hair leading to baldness. It is more prevalent in men than women, and the hair loss pattern differs between the sexes (1). In men, AGA, also known as male pattern hair loss, is defined by a distinct M-shaped pattern hair loss that begins with

a bi-temporal recession of the frontal hairline, followed by hair thinning at the frontal and vertex scalp region, which eventually converges resulting in complete baldness in the frontal and vertex scalp region (1, 2). Hair loss, particularly adolescent AGA, causes serious psychosocial ramifications in men affecting their self-esteem and quality of life (3).

Hair loss in AGA is attributed to the gradual transformation of thick pigmented large terminal hairs into non-pigmented small fine vellus hair through hair follicle miniaturization process driven by the androgen 5α-dihydrotestosterone (5α-DHT) (1). However, the mechanism of hair follicle miniaturization is poorly understood and the inadequate understanding of the pathobiology of AGA impedes the search for a permanent cure to hair loss (4). Molecular genetic studies have identified 12 genomic regions of interest and genes such as AR, EDA2R, PAX1, FOXA2, HDAC9, TARDBP, HDAC4, AUTS2, IMP5, SETBP1, SUCNR, MBBL1, EBF1, WNT10A, SSPN, and ITPR2 associated with AGA (2). However, these identified genes explain only a limited proportion of the pathogenesis and genetic variance of AGA since most of the identified genetic variants reside in the non-coding region of the genome for which no clear functional effect has been established yet (2). Hence, the identification of additional genetic loci for AGA is warranted to understand the pathobiology and to aid drug discovery.

Recently, Michel et al. (5) performed a microarray gene expression analysis between hairless or bald vertex scalp from young men with premature AGA and haired scalp from control men to identify dysregulated genes in AGA. The identification of differentially expressed genes (DEGs) was carried out by analysis of variance test and Tukey's *post-hoc* tests. After Benjamini-Hochberg correction they, found 184 down-regulated and 149 up-regulated genes in the AGA group compared with the healthy group. In this study, we utilized the same data of Michel et al. (5) to identify DEGs in the AGA pathology employing a different method and threshold criteria. We constructed biological networks, such as the STRING protein-protein interaction (PPI) and Reactome Functional Interactome (FI) networks, using the DEGs obtained. We then focused on the hub nodes in both the PPI and FI networks and identified the hub genes that were common to both networks as worthy of further investigation into the signaling pathways involved in AGA development.

Materials and methods

Microarray data

The raw dataset of the gene expression profile GSE90594 generated by Michel et al. (5) was downloaded from the GEO database (6). The data was obtained from scalp biopsies taken from the vertex region of 14 young males with premature alopecia (age 29.4 ± 3.4 years, stage V–VII as per Hamilton-Norwood classification) and 14 healthy volunteers with less than 2% white hairs (age 26.1 ± 3.6 years, Stage I or II according to Hamilton-Norwood classification). Both the alopecia and healthy group did not have any other skin involvement, autoimmune disorders, and systemic diseases (5).

Data preprocessing and differential gene expression analysis

limma v3.50.3 (Linear Models for Microarray Data) package, a R/Bioconductor software package, which provides an integrated solution for analyzing gene expression data from microarray technologies was utilized for data analysis (7). The Data pre-processing included background correction using normexp method and quantile normalization. Boxplot and cluster analyses were performed to identify and remove outliers in the samples. Then the control probes and the unexpressed probes are filtered out while the probes that are expressed above background are retained for further analysis. In addition, for multiple probes corresponding to the same genes in the arrays their average expression value was computed by *avereps* function in *limma*. Then the DEGs for the alopecia samples compared to the healthy samples were mined using the single-channel design matrix provided in the *limma* package. Benjamini and Hochberg's method was utilized to compute the adjusted *p*-values (False Discovery Rate, FDR) (8). The probes with adjusted *p*-value (FDR) < 0.05 were selected as differentially expressed.

Gene ontology and pathway enrichment analysis

ToppGene Suite[1] (updated: Mar 2021) was employed to perform gene ontology (GO) functional and pathway analysis to identify items in gene lists that may have relevance to the biological question being investigated (9, 10). The ToppFun function in the ToppGene Suite was utilized to carry out GO (biological process and molecular function), gene family (source: genenames.org), and pathway enrichment (source: Biosystems-Reactome) analyses for the DEGs. All genes that are detected in the microarray analysis were used as background gene set in the ToppGene Suite for these analyses. The probability density function which is the default method for *p*-value and FDR calculation was selected. Gene count >2 and FDR B&H *q*-value < 0.05 were chosen as the cut-off criteria for the analyses.

Annotation of differentially expressed genes with AGA risk loci

Windows of 50 kb, 100 kb and 500 kb flanking the 107 lead SNPs associated from 8 genome wide association studies [Study Accession IDs: GCST000250 (11), GCST000251 (12), GCST001548 (13), GCST001297 (14), GCST005116 (15), GCST90043616 (16), GCST003983 (17), and GCST90043619 (16)] for the trait androgenetic alopecia indexed in the NHGRI-EBI GWAS catalog were prepared by calculating coordinates of 50 kb, 100 kb and 500 kb distance on either sides from the SNP position. Gene coordinates of DEGs transcript(s) were annotated using RefSeq Identifiers (Hg38). The flanking coordinates of SNPs were overlapped with the coordinates of DEGs utilizing the intersect

1 https://ToppGene.cchmc.org/

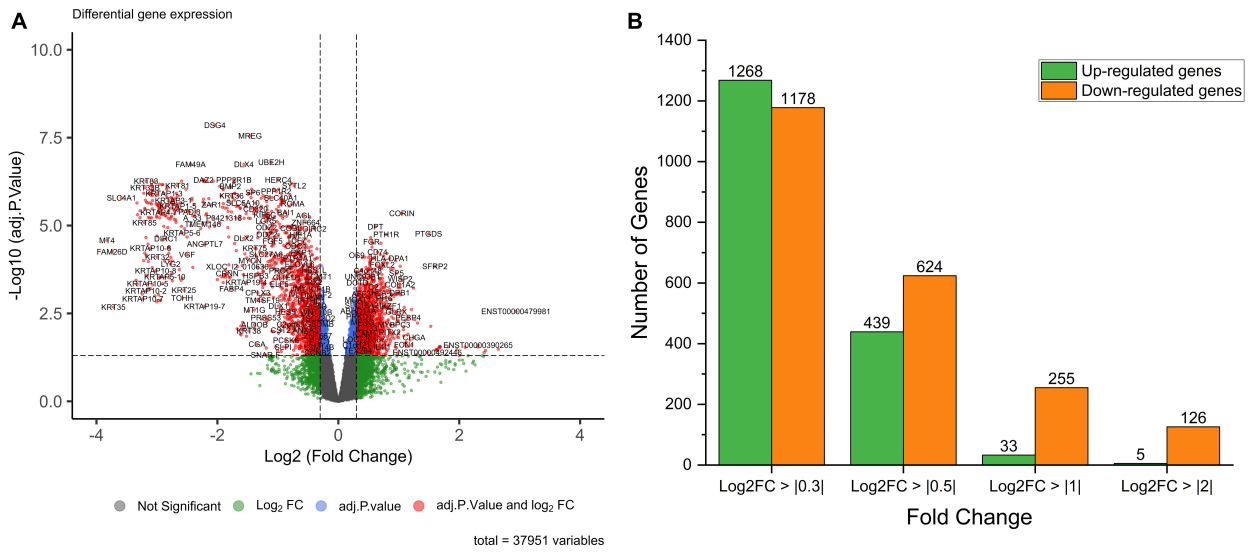


FIGURE 1
(A) Volcano plot of the differentially expressed genes in AGA with $\log_2$FC > | 0.3| and adj. p-value < 0.05 as cut off value. **(B)** Bar chart depicting the number of differentially expressed genes for the $\log_2$FC values.

function in Bedtools v2.30.0 (18). An overlap is only considered when there is a minimum of 1 bp overlap between the coordinates of DEG transcripts and the flanking coordinates of the lead SNPs (19).

Motif analysis in the promoter regions of differentially expressed genes

The promoter regions of up and down-regulated DEGs were separately subjected to motif analysis utilizing the gene-based analysis method in HOMER v4.11 software[2] (20). 2,000 bp upstream and 200 bp downstream relative to the transcriptional start site of the genes were considered as promoter regions (19) and the promoter sets for the DEGs were constructed based on RefSeq genes (Hg38). Motifs of length up to 12 bases were probed with Benjamini-Hochberg-corrected p-value ≤ 0.05 as cut-off value.

STRING protein-protein interaction network

The protein-protein interaction (PPI) interaction network for the DEGs were computed through the STRING database. The online web resource STRING v11.5[3] is a biological database that includes direct (physical) and indirect (functional) protein-protein association data which are both specific and biologically meaningful (21). The PPI interaction network for the DEGs were computed through the stringApp plugin v1.7.1 in Cytoscape v3.9.1 (22). An interaction score of 0.900 (highest confidence) was used as the cut off criterion for constructing the PPI network.

Reactome functional interaction network

Reactome functional interaction (FI) network was constructed for the DEGs utilizing the Cytoscape application ReactomeFIViz v8.0.4 which probe for disease-related pathways and network patterns using the Reactome functional interaction (FI) network (23, 24) created based on the well-known biological pathway database Reactome[4] (25, 26). Reactome FI network 2021 version was used to construct the FI network for the DEGs. Gene ontology biological process and pathway enrichment analysis for the nodes (genes) mapped in the network was carried out through the inbuilt Reactome FI network analysis tool.

Network analysis and hub gene identification

The topological properties of the PPI and FI network were analyzed through the Cytoscape pre-installed network analyzer v4.4.8 tool (27). Cytohubba v0.1 plugin was used to identify hub proteins in the PPI and FI network and rank them based on topological algorithms and centralities such as Maximal Clique Centrality (MCC), Maximum Neighborhood component (MNC), Density of Maximum Neighborhood Component (DMNC), Degree, Closeness, and betweenness (28). The clusters in the networks were determined using the MCODE plugin with specific parameters including a degree cut-off of 2, fluff node density cut-off of 0.1, node score cut-off of 0.2, K-core of 2,

2 http://homer.ucsd.edu/homer/microarray/index.html
3 https://string-db.org/
4 https://reactome.org/

and max depth of 100 to determine the highly interconnected nodes (29).

Results

Data processing and screening of differentially expressed genes

The GSE90594 dataset contained 28 samples of which 14 samples are from men with premature AGA and 14 samples from healthy men without hair loss. Cluster analysis of the samples after background correction and normalization of the arrays revealed 9 samples (5 alopecia and 4 healthy samples) as outliers (**Supplementary I-1, 2**). The outlier samples were removed, and differential gene expression analysis was carried out between 9 alopecia and 10 healthy samples. The probes were annotated with Entrez Gene ID, Gene Symbol, and Gene names using the clusterProfiler 4.0 v4.4.3 package in R by querying the Reference Seq ID (30). From this list, the probes that have a valid Entrez Gene ID are selected for further analysis. Subsequently probes with similar Probe IDs (Probe Names) are averaged using avereps function in *limma* and the probes with different probe ID for same genes are kept as such.

The analysis returned a total of 289 DEGs (33 up-regulated and 256 down-regulated DEGs) for a threshold cut off of value $\log_2 FC > |1|$ and q-value < 0.05 (**Supplementary II-6**). Further to construct a big and detailed STRING PPI and reactome FI networks a total of 2,439 unique DEGs (1,261 up-regulated, 1,171 down-regulated, and 7 genes with probes expressed in both directions) that falls within a cut off value of $\log_2 FC > |0.3|$ and q-value < 0.05 were mined (**Figure 1**). The gene family enrichment analysis of these 2,439 DEGs are given in **Figure 2** and **Supplementary II-7**. GO functional analyses revealed that the up-regulated genes enriched for immune system mediated GO terms implying a heightened immune response in hairless scalp, while the down-regulated genes enriched for hair growth related GO terms as expected (**Table 1** and **Supplementary II-8**). The Reactome pathway enrichment analysis also enriched pathways such as keratinization, developmental biology G2/M DNA replication checkpoint for down-regulated genes, wherein for up-regulated genes innate immune system, Cytokine signaling, interferon signaling, adaptive immune system, and antigen processing cross presentation pathways were enriched (**Table 2** and **Supplementary II-9**). The DEG list was inspected for genes known to be involved in various signaling pathways such as Wnt, NF-κB, TGF-β, BMP, and Vitamin D metabolism and the mapped DEGs for these signaling pathways were provided in the **Supplementary I-3**.

Annotation of differentially expressed genes with AGA risk loci

Mapping of SNPs identified through GWAS to DEGs annotates genetic variants located in or near the gene regions that are differentially expressed, helps to understand the functional role of DEGs and their association with disease. To identify genes that potentially contribute to AGA pathology, we annotated the coordinates of 107 genomic loci associated with AGA risk in men identified through GWAS with our DEGs. The analysis identified 51 DEGs within the window of 500 kb of 73 AGA risk SNPs and 14 DEGs within the window of 50 kb of 16 lead SNPs (**Table 3**). Some of the DEGs were mapped to reported AGA risk SNPs including MEMO1 at loci 2p22.3, SRD5A2 at 2p23.1, FOXL2NB at 3q23, FGF5 at 4q21.21, DKK2 at 4q25, EBF1 at 5q33.3, IRF4 at 6p25.3, CENPW at 6q22.32, PAGE2 at Xp11.21, highlighting their association with AGA pathology. Our analysis also identified other DEGs, such as HOXD9 at 2q31.1, LHPP at 10q26.13, CRHR1 at 17q21.31, STH at 17q21.31, and PAGE2B at Xp11.21, as more likely to be candidate gene risks for AGA than the mapped genes in GWAS studies. MEMO1 was down-regulated, and it plays a crucial role in regulating cell proliferation, survival, and differentiation in the hair follicle (30). SRD5A2, whose product inhibits hair growth, was up-regulated (31). HOXD9, a member of the HOX family of genes that plays a crucial role in the development and patterning of various tissues and organs in the body, was up-regulated in our analysis, although its role in hair growth is unknown (32). FGF5, which inhibits hair growth and is involved in the transition of hair follicles from anagen to catagen phase, was down-regulated (33). DKK2, a Wnt inhibitor that leads to hair growth inhibition was up-regulated (1). FOXL2NB, IRF4, CENPW, EBF1, LHPP, CRHR1, and STH located within the AGA risk loci warrant further investigation. The mapped DEGs within 500 kb of lead SNPs may also be considered for future investigation of their association with hair growth and AGA (**Table 3**).

Motif analysis in the promoter regions of AGA differentially expressed genes

The transcription factor motif enrichment analysis on the promoter regions of the differentially expressed genes was carried to identify potential transcription factors involved in the AGA pathology. The top transcription factor motifs enriched for the down-regulated genes are LEF1 (Lymphoid Enhancer binding Factor 1), HOXB13 (Homeobox B13), NEUROD1 (Neuronal Differentiation 1), ZNF189 (Zinc Finger protein 189), and MEF2C (MADS Box Transcription factor 2, Polypeptide C) (**Figure 3** and **Supplementary II-10**). The transcription factor LEF1 actively participates in the Wnt signaling pathway by activating the transcription of target genes in the presence of β-catenin. Wnt/β-catenin Signaling plays a crucial role in hair follicle differentiation and morphogenesis (31). The transcription factor HOXB13 belongs to HOX gene family which plays a crucial role in regulating embryonic development including hair formation. HOXB13 is implicated in skin development and low level of its expression is associated with telogen hair follicle (32, 34). The transcription factor NEUROD1 is primarily involved in the development and differentiation of the nervous system. NEUROD1 acts by controlling the expression of genes involved in neuronal development and in the formation of axons and dendrites (35). ZNF189 belongs to the zinc finger protein family which play important roles in various biological processes including transcriptional regulation, DNA repair, and cellular signaling. MEF2C belongs to the MADS box transcription factor 2 (MEF2) family of transcription factors and is involved in myogenesis (32).

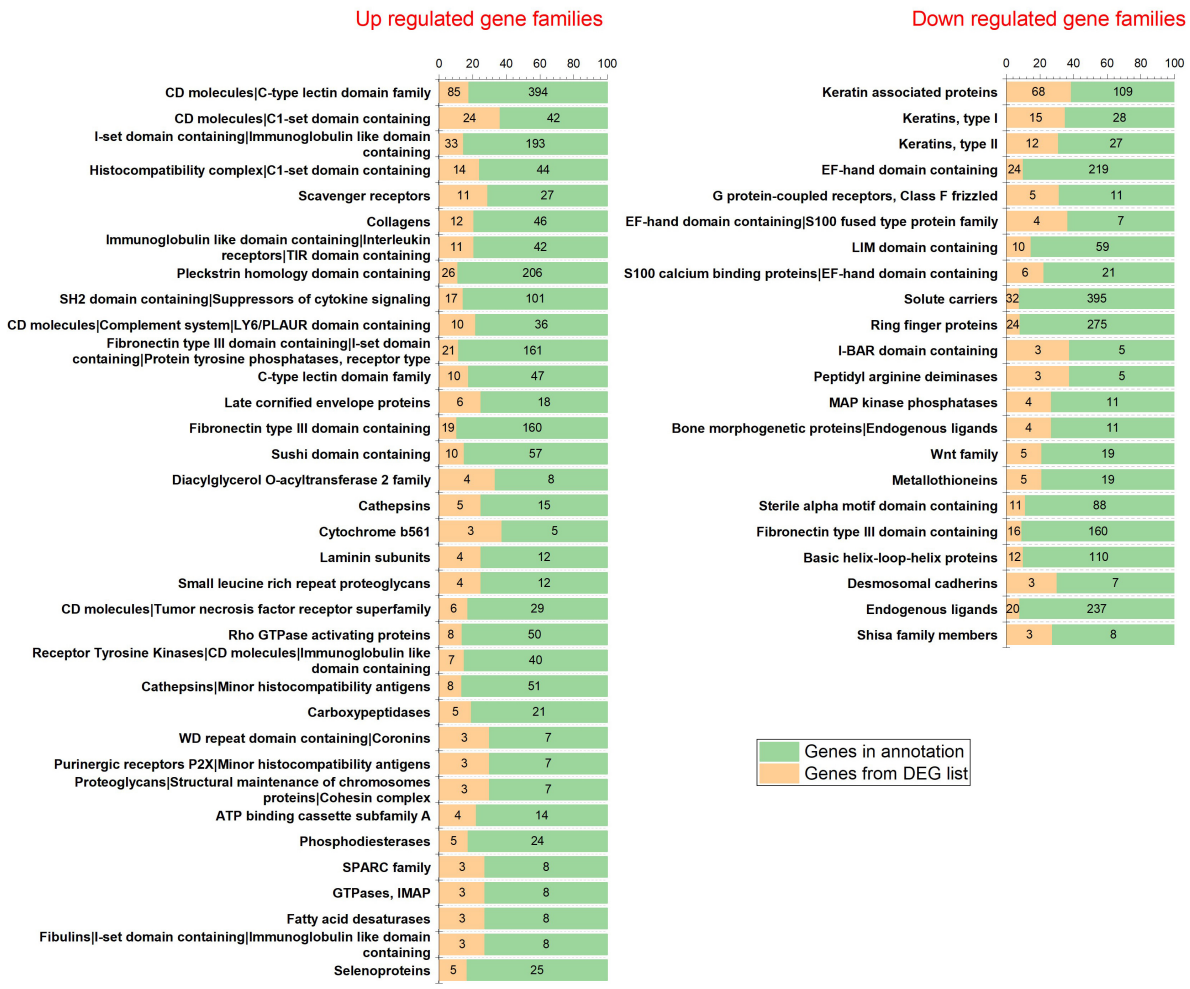


FIGURE 2
Gene family enrichment analysis of the differentially expressed genes.

Many transcription factor motifs belonging to the SMAD, HOX, STAT, ZNF, NEURO, FOX, and FOS gene families (Figure 3) are enriched for the down-regulated genes indicating their role in hair growth which has to be studied further.

The motifs for IRF3 (Interferon Regulatory Factor 3), PRDM1 (PR/SET Domain 1), IRF8 (Interferon Regulatory Factor 8), SPI1 (Spi-1 Proto-Oncogene), SPI1:IRF8, ISRE (Interferon-sensitive response element), IRF2 (Interferon Regulatory Factor 2), IRF1 (Interferon Regulatory Factor 1) and SF1 transcription factors were enriched as the top motifs for the up-regulated genes (Figure 3 and Supplementary I-11). The Interferon regulatory factors (IRFs) are a family of transcription factors that regulate various aspects of the immune system from promoting immune cell development to immune cell differentiation. They play a central role in controlling the innate and adaptive immune responses to pathogens (33). IRF1 and IRF2 are important in regulating dendritic cells which participates in antigen presentation and bridge the innate and adaptive immune system. IRF3 involves in type I interferon production and IRF8 regulate myeloid cell development (33). PRDM1 coordinates several important functions in the adaptive immune system that support the key effector functions of B and T lymphocytes (36). SPI-1 encodes an ETS-domain transcription factor that control gene expression involving in the development of myeloid and B-lymphoid immune cells (37). The enrichment of these transcription factor motifs as the top motifs in the up-regulated genes of bald scalp implies a state of heightened immune response in AGA.

STRING protein-protein interaction network analysis and identification of hub genes

The stringApp generated 1967 PPI pairs for the submitted DEGs. The main PPI network, which consisted of 749 nodes (447 up-regulated genes, 273 down-regulated genes, and 29 linker genes) and 1,856 edges, was selected for further analysis while disconnected nodes and small isolated PPI pairs were discarded (Figure 4 and Supplementary II-12). The PPI network had a clustering coefficient of 0.334, a characteristic path length of 5.683, a network diameter of 19, a network density of 0.007, and an average of 4.956 neighbors. The functional enrichment analysis performed using the inbuilt STRING tool on the Reactome and Wikipathway databases revealed that the PPI network was enriched for several

TABLE 1 Result of gene ontology analysis of DEGs from ToppGene Suite (FDR < 0.05).

	Gene ontology of up-regulated genes			Gene ontology of down-regulated genes		
	Gene ontology term	Number of genes from DEG list	Number of genes in annotation	Gene ontology term	Number of genes enriched	Number of genes in annotation
Molecular function	Extracellular matrix structural constituent	44	195	Structural constituent of skin epidermis	16	44
	Signaling receptor binding	181	1,813	Structural molecule activity	87	892
	Protein-containing complex binding	166	1,726			
	Oxidoreductase activity	97	834			
	integrin binding	35	171			
	Immune receptor activity	34	165			
	Carbohydrate binding	46	315			
	Antigen binding	32	189			
	MHC protein Complex binding	14	43			
	MHC class II protein complex binding	11	27			
Biological process	Regulation of immune system process	239	1,821	Intermediate filament organization	29	74
	Cell activation	208	1,464	Molting cycle	38	149
	leukocyte activation	184	1,277	Hair cycle	38	149
	Immune effector process	144	895	Intermediate filament cytoskeleton organization	30	96
	Regulation of immune response	159	1,088	Intermediate filament-based process	30	98
	Positive regulation of immune system process	165	1,164	Epithelium development	180	1,979
	Lymphocyte activation	154	1,058	Skin development	60	387
	Cell adhesion	211	1,742	Epidermis development	67	500
	Leukocyte mediated immunity	105	594	Hair follicle development	27	120
	T cell activation	115	704	Hair cycle process	27	123
Cellular component	Cell surface	162	1,178	Intermediate filament	101	229
	External side of plasma membrane	104	599	Keratin filament	73	108
	Side of membrane	124	853	Intermediate filament cytoskeleton	103	271
	Extracellular matrix	106	678	Polymeric cytoskeletal fiber	156	889
	External encapsulating structure	106	680	Supramolecular polymer	177	1,181
	Collagen-containing extracellular matrix	90	541	Supramolecular fiber	176	1,172
	Intrinsic component of plasma membrane	202	1,992	Supramolecular complex	195	1,549
	Integral component of plasma membrane	192	1,893	Anchoring junction	109	1,419
	Secretory granule	116	987	Cell-cell junction	53	590
	MHC protein complex	16	26	Extracellular matrix	58	678

immune response-related pathways (Supplementary II-13). The pathway terms related to Cytokine Signaling in the Immune System, Interferon Signaling, T cell receptor signaling, Signaling by Interleukins, Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell, Adaptive and Innate immune systems, and TCF-dependent signaling in response to WNT were enriched

TABLE 2 Result of reactome pathway enrichment analysis of DEGs from ToppGene Suite (FDR < 0.05).

Reactome ID	Pathway name	Genes from DEG list	Genes in annotation
Up-regulated genes			
1269203	Innate immune system	155	1,302
1269310	Cytokine signaling in immune system	102	760
1270244	Extracellular matrix organization	53	297
1269314	Interferon gamma signaling	27	94
1269201	Immunoregulatory interactions between a Lymphoid and a non-lymphoid cell	31	135
1457780	Neutrophil degranulation	69	492
1269318	Signaling by interleukins	70	528
1269340	Hemostasis	78	639
1269311	Interferon Signaling	33	202
1269173	Phosphorylation of CD3 and TCR zeta chains	10	25
1270246	Collagen biosynthesis and modifying enzymes	17	70
1269171	Adaptive immune system	88	823
1269350	Platelet activation, signaling and aggregation	40	282
1470923	Interleukin-4 and 13 signaling	22	114
1269174	Translocation of ZAP-70 to immunological synapse	9	22
1270260	Integrin cell surface interactions	16	68
1270001	Metabolism of lipids and lipoproteins	84	816
1269182	PD-1 signaling	9	26
1269195	Antigen processing-cross presentation	19	101
1270245	Collagen formation	18	93
Down-regulated genes			
1457790	Keratinization	103	214
1270302	Developmental biology	158	1,078
1269756	G2/M DNA replication checkpoint	4	5
1269570	Class B/2 (Secretin family receptors)	16	93

from the Reactome database. Additionally, the Wikipathways database identified significant pathway terms related to the Inflammatory response pathway, Development of pulmonary dendritic cells and macrophage subsets, B cell receptor signaling pathway, and the Vitamin D receptor pathway (Supplementary II-13). These results confirm the credibility of the PPI network and reinforce the observation of immune response-related and hair follicle-related pathways.

The top 20 ranking hub nodes (genes) in the PPI network were identified using the Cytoscape plugin Cytohubba based on four topological analysis methods and two centralities (MCC, DMNC, MNC, Degree, Closeness, and Betweenness) and are listed in Table 4. Out of these, a total of 15 hub genes that appeared in at least three of these categories were considered as significant hub genes, and the frequently appeared genes are highlighted in the Table 4. The MCODE cluster analysis performed on the String PPI network revealed 8 clusters when using the 15 hub genes as roots for clustering. The 8 clusters were comprised of 53, 63, 101, 59, 37, 41, 53, and 9 nodes, respectively (Supplementary II-14). The top 3 highly interconnected clusters were selected for further analysis (Supplementary 1–4). Cluster 1 had 10 hub genes, cluster 2 contained 5 hub genes, and cluster 3 had 6 hub genes. Our analysis revealed that two hub genes LCK and STAT5A appeared in all 3 clusters strongly suggesting their putative role in AGA.

Reactome protein functional interaction network analysis and identification of hub genes

The ReactomeFIViz tool was utilized to construct the FI network for the DEGs resulting in an initial network of 1,092 connected nodes, 1,340 unconnected nodes, and 4,047 edges (Supplementary II-15). The unconnected nodes were discarded from the analysis and the final FI network consisted of 1,014 nodes (581 up-regulated and 433 down-regulated genes) with 3,980 edges as shown in Figure 5. The FI network had a clustering coefficient of 0.280, a network diameter of 11, a network density of 0.008, and an average number of neighbors of 7.850. The pathway enrichment and GO Biological process analyses were conducted using the inbuilt ReactomeFIViz – analysis network function tool (Supplementary II-16). Reactome pathway terms such as Extracellular matrix organization, Keratinization,

TABLE 3 Overlap of AGA risk loci from genome-wide significance studies with differentially expressed genes in premature AGA samples compared to normal men.

Chr	Cytogenetic region	SNP	Study accession	Mapped genes	500 kb		100 kb		50 kb		Log_2FC
					Diff genes	Up/Down	Diff genes	Up/Down	Diff genes	Up/Down	
Chr1	1p33	rs61784834	GCST005116	RPL21P24, FOXD2	TRABD2B	Down	–	–	–	–	–0.61
Chr1	1p36.11	rs11249243	GCST005116	RUNX3, MIR4425	RUNX3	Up	RUNX3	Up	–	–	0.38
					CLIC4	Down	–	–	–	–	–0.62
Chr1	1p36.11	rs9803723	GCST005116	IFITM3P7, SYF2	RUNX3	Up	–	–	–	–	0.38
					CLIC4	Down	–	–	–	–	–0.62
Chr1	1p36.11	rs2064251	GCST005116	IFITM3P7, SYF2	RUNX3	Up	–	–	–	–	0.38
					CLIC4	Down	–	–	–	–	–0.62
Chr1	1p36.11	rs7534070	GCST003983	SYF2, IFITM3P7	RUNX3	Up	–	–	–	–	0.38
					CLIC4	Down	–	–	–	–	–0.62
Chr1	1p36.22	rs12565727	GCST001548	C1orf127	ANGPTL7	Down	–	–	–	–	–2.21
Chr1	1p36.22	rs2095921	GCST003983	C1orf127	ANGPTL7	Down	–	–	–	–	–2.21
Chr1	1p36.22	rs7542354	GCST005116	C1orf127	ANGPTL7	Down	–	–	–	–	–2.21
Chr1	1q24.2	rs78003935	GCST003983	GORAB-AS1, HAUS4P1	PRRX1	Up	–	–	–	–	0.54
Chr1	1q24.2	rs11578119	GCST005116	GORAB, GORAB-AS1	PRRX1	Up	–	–	–	–	0.54
Chr2	2p14	rs6546334	GCST003983	LINC01812	CNRIP1	Up	–	–	–	–	0.37
Chr2	2p14	rs62146540	GCST005116	FBXL12P1	CNRIP1	Up	–	–	–	–	0.37
					PLEK	Up	–	–	–	–	0.64
Chr2	2p21	rs11694173	GCST003983	THADA	ZFP36L2	Up	–	–	–	–	0.31
Chr2	2p22.3	rs13021718	GCST005116	DPY30, MEMO1	MEMO1	Down	MEMO1	Down	MEMO1	Down	–0.32
					SRD5A2	Up	–	–	–	–	0.71
Chr2	2p23.1	rs9282858	GCST003983	SRD5A2	SRD5A2	Up	SRD5A2	Up	SRD5A2	Up	0.71
					GALNT14	Down	–	–	–	–	–0.59
					MEMO1	Down	–	–	–	–	–0.32
					EHD3	Down	–	–	–	–	–1.49
					CAPN14	Down	–	–	–	–	–0.92
Chr2	2q13	rs3827760	GCST003983, GCST90043616	EDAR	GCC2	Down	–	–	–	–	–0.35
Chr2	2q31.1	rs13405699	GCST005116, GCST003983	’–	MAP3K20	Down	–	–	–	–	–0.42
Chr2	2q31.1	rs71421546	GCST005116	HOXD-AS2	HOXD9	Up	HOXD9	Up	HOXD9	Up	0.34
Chr2	2q35	rs74333950	GCST003983	WNT10A	CYP27A1	Up	CYP27A1	Up	–	–	0.42

(Continued)

10.3389/fmed.2023.1108358

TABLE 3 (Continued)

Chr	Cytogenetic region	SNP	Study accession	Mapped genes	500 kb		100 kb		50 kb		Log_2FC
					Diff genes	Up/Down	Diff genes	Up/Down	Diff genes	Up/Down	
Chr2	2q35	rs7349332	GCST005116	WNT10A	CYP27A1	Up	CYP27A1	Up	–	–	0.42
Chr2	2q37.3	rs9287638	GCST001548	TWIST2, LINC01937	TWIST2	Up	TWIST2	Up	–	–	0.48
Chr2	2q37.3	rs11684254	GCST005116, GCST003983	LINC01937, TWIST2	TWIST2	Up	TWIST2	Up	–	–	0.48
Chr3	3q23	rs6788232	GCST005116	PRR23A, FOXL2NB	FOXL2NB	Up	FOXL2NB	Up	FOXL2NB	Up	0.73
					FOXL2	Up	FOXL2	Up	–	–	0.95
Chr3	3q23	rs7642536	GCST005116, GCST003983	MRPS22	FOXL2NB	Up	–	–	–	–	0.73
					FOXL2	Up	–	–	–	–	0.95
Chr3	3q25.1	rs4679956	GCST003983	AADACL2-AS1	IGSF10	Up	–	–	–	–	0.45
Chr3	3q25.1	rs16863765	GCST005116	AADACL2-AS1	IGSF10	Up	–	–	–	–	0.45
Chr4	4q21.21	rs7680591	GCST005116	FGF5	FGF5	Down	FGF5	Down	FGF5	Down	−1.09
Chr4	4q21.21	rs4690116	GCST003983	FGF5	FGF5	Down	FGF5	Down	FGF5	Down	−1.09
Chr4	4q25	rs78311490	GCST003983	DKK2	DKK2	Up	DKK2	Up	DKK2	Up	0.34
Chr5	5q33.3	rs1422798	GCST005116	EBF1	EBF1	Up	EBF1	Up	EBF1	Up	0.34
					RNF145	Down	–	–	–	–	−0.40
Chr5	5q33.3	rs62385385	GCST003983	EBF1	EBF1	Up	EBF1	Up	EBF1	Up	0.34
					RNF145	Down	–	–	–	–	−0.40
Chr6	6p25.3	rs12203592	GCST005116, GCST003983	IRF4	IRF4	Up	IRF4	Up	IRF4	Up	0.49
Chr6	6q21	rs12214131	GCST005116	’–	PREP	Down	–	–	–	–	−0.37
Chr6	6q22.32	rs9398803	GCST005116	CENPW	CENPW	Down	CENPW	Down	CENPW	Down	−0.32
Chr6	6q22.32	rs1262557	GCST003983	RPS4XP9	CENPW	Down	–	–	–	–	−0.32
Chr7	7p21.1	rs2073963	GCST001548	HDAC9	TWIST1	Up	–	–	–	–	0.57
Chr7	7p21.1	rs71530654	GCST005116	HDAC9	TWIST1	Up	–	–	–	–	0.57
Chr7	7p21.1	rs7801037	GCST003983	HDAC9	TWIST1	Up	–	–	–	–	0.57
Chr7	7q11.22	rs939963	GCST005116	RNU6-832P	AUTS2	Up	–	–	–	–	0.31
Chr7	7q11.22	rs34991987	GCST003983	RNU6-832P	AUTS2	Up	–	–	–	–	0.31
Chr7	7q11.22	rs6945541	GCST001548	RNU6-832P	AUTS2	Up	–	–	–	–	0.31
Chr7	7q11.22	rs4718886	GCST005116	Y_RNA, RNU6-229P	AUTS2	Up	–	–	–	–	0.31
Chr7	7q32.3	rs9719620	GCST005116	MKLN1, MKLN1-AS	LINC-PINT	Up	–	–	–	–	0.40

(Continued)

10.3389/fmed.2023.1108358

TABLE 3 (Continued)

Chr	Cytogenetic region	SNP	Study accession	Mapped genes	500 kb		100 kb		50 kb		Log_2FC
					Diff genes	Up/Down	Diff genes	Up/Down	Diff genes	Up/Down	
Chr10	10q22.3	rs11593840	GCST005116, GCST003983	LRMDA	KCNMA1	Up	–	–	–	–	0.46
Chr10	10q26.13	rs3781458	GCST003983	FAM53B	LHPP	Up	LHPP	Up	LHPP	Up	0.30
Chr10	10q26.13	rs3781452	GCST005116	FAM53B	LHPP	Up	LHPP	Up	LHPP	Up	0.30
Chr11	11p11.2	rs11037975	GCST005116, GCST003983	ALX4	CD82	Down	–	–	–	–	–0.33
					ACCS	Up	–	–	–	–	0.56
Chr12	12p11.22	rs7976269	GCST005116	FAR2	TMTC1	Down	–	–	–	–	–0.48
Chr12	12p12.1	rs9300169	GCST003983	SSPN	RASSF8-AS1	Up	–	–	–	–	0.38
Chr12	12p12.1	rs7974900	GCST005116	SSPN	RASSF8-AS1	Up	–	–	–	–	0.38
Chr12	12q13.13	rs180807105	GCST90043616	HOXC12	MAP3K12	Up	–	–	–	–	0.31
					NFE2	Up	–	–	–	–	0.71
Chr12	12q24.33	rs76972608	GCST005116, GCST003983	FZD10-AS1, LINC02419	FZD10	Down	FZD10	Down	–	–	–0.55
Chr13	13q12.3	rs9314998	GCST003983	LINC00385, KATNAL1	LINC00426	Up	–	–	–	–	0.44
Chr17	17q21.31	rs12373124	GCST001548	MAPT-AS1, SPPL2C	CRHR1	Down	CRHR1	Down	CRHR1	Down	–0.72
					STH	Up	STH	Up	STH	Up	0.36
Chr17	17q21.31	rs919462	GCST005116	MAPT	STH	Up	STH	Up	STH	Up	0.36
					CRHR1	Down	–	–	–	–	–0.72
Chr17	17q21.31	rs201408539	GCST003983	KANSL1	STH	Up	STH	Up	–	–	0.36
					CRHR1	Down	–	–	–	–	–0.72
Chr17	17q21.31	rs572756998	GCST005116	ARL17B	CRHR1	Down	–	–	–	–	–0.72
					STH	Up	–	–	–	–	0.36
					WNT3	Down	–	–	–	–	–0.68
Chr17	17q22	rs17833789	GCST005116	AKAP1	MSI2	Down	–	–	–	–	–0.33
					MTVR2	Up	–	–	–	–	0.38
Chr17	17q22	rs62060349	GCST003983	LINC02563, AKAP1	MSI2	Down	–	–	–	–	–0.33
					MTVR2	Up	–	–	–	–	0.38
Chr20	20p11.22	rs2180439	GCST000251, GCST001297	’–	PAX1	Up	–	–	–	–	0.81
Chr20	20p11.22	rs77410716	GCST005116	’–	PAX1	Up	–	–	–	–	0.81
Chr20	20p11.22	rs552649178	GCST005116	LINC01432	PAX1	Up	–	–	–	–	0.81

(Continued)

10.3389/fmed.2023.1108358

TABLE 3 (Continued)

Chr	Cytogenetic region	SNP	Study accession	Mapped genes	500 kb		100 kb		50 kb		Log_2FC
					Diff genes	Up/Down	Diff genes	Up/Down	Diff genes	Up/Down	
Chr20	20p11.22	rs201563	GCST003983	LINC01432	PAX1	Up	–	–	–	–	0.81
Chr20	20p11.22	rs6047844	GCST001548	LINC01432	PAX1	Up	–	–	–	–	0.81
Chr20	20p11.22	rs11087368	GCST005116	LINC01432	PAX1	Up	–	–	–	–	0.81
Chr20	20p11.22	rs1160312	GCST000250	LINC01432	PAX1	Up	–	–	–	–	0.81
ChrX	Xp11.21	rs185597083	GCST003983	FAM104B, PAGE2	PAGE2	Up	PAGE2	Up	PAGE2	Up	0.48
ChrX	Xp22.31	rs5933688	GCST003983	ANAPC15P1, NOLC1P1	PAGE2B	Up	PAGE2B	Up	PAGE2B	Up	0.69
					ANOS1	Down	–	–	–	–	−0.61
ChrX	Xp22.31	rs5934505	GCST005116	ANAPC15P1, NOLC1P1	ANOS1	Down	–	–	–	–	−0.61
ChrX	Xq12	rs200644307	GCST003983	,_	AR	Up	–	–	–	–	0.43
ChrX	Xq12	rs6625163	GCST000250	,_	AR	Up	–	–	–	–	0.43
ChrX	Xq12	rs2497938	GCST001548, GCST001297	,_	AR	Up	–	–	–	–	0.43
ChrX	Xq12	rs7061504	GCST005116	OPHN1	AR	Up	–	–	–	–	0.43

Interferon gamma signaling, Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins, Response to elevated platelet cytosolic Ca2 +, Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell, and WNT ligand biogenesis and trafficking were enriched for the genes present in the FI network. GO Biological process terms such as immune response, transmembrane receptor protein tyrosine kinase signaling pathway, canonical Wnt signaling pathway, and inflammatory response were also enriched.

The top 20 ranked hub genes in the FI network identified based on the six algorithms including MCC, DMNC, MNC, Degree, Closeness, and Betweenness are presented in the Table 5. Out of these, a total of 19 hub genes that appeared in at least three of the categories were considered significant hub genes, and the frequently appeared genes are highlighted in the Table 5. The MCODE cluster analysis of the FI network revealed 11 clusters when using the 19 hub genes as roots for clustering. The 11 clusters had node numbers of 189, 55, 216, 47, 34, 59, 153, 180, 29, 69, and 6, respectively (Supplementary II-17). The top 3 clusters ranked based on their cluster score were selected for further analysis (Supplementary I-5). Cluster 1 consisted of 14 hub genes, cluster 2 contained 1 hub genes, and cluster 3 had 9 hub genes. Our results showed that seven hub genes (HCK, GNAI3, RAC2, PDGFRB, EGF, NRAS, and STAT5A) were present in two of the selected clusters indicating their potential role in AGA.

Candidate genes in AGA pathology

The 25 hub genes identified from the analyses of the PPI and FI networks constructed based on the DEGs were considered key genes in the pathology of AGA (Table 6). Out of these 25 hub genes, 21 genes (BTK, ESR1, HCK, ITGB7, LCK, LCP2, LYN, PDGFRB, PIK3CD, PTPN6, RAC2, SPI1, STAT3, STAT5A, VAV1, PSMB8, HLA-A, HLA-F, HLA-E, IRF4, and ITGAM) were found to be up-regulated, while 4 genes (CTNNB1, EGF, GNAI3, and NRAS) were downregulated. The results of the GO biological process and pathway enrichment analysis, conducted using the Toppgene suite, revealed that the hub genes were associated with immune and inflammatory processes (Table 7). The significant biological terms enriched for the hub genes included regulation of immune system process, T cell activation, immune response-regulating cell surface receptor signaling pathways. Furthermore, the significant pathway terms enriched for the hub genes included cytokine signaling in the immune system, signaling by interleukins, signaling by the B cell receptor (BCR), and signaling by SCF-KIT, interleukin-3, 5, and GM-CSF signaling, and the innate immune system.

CTNNB1 (Catenin Beta 1) is a crucial downstream component of the Canonical Wnt Signaling Pathway. In the presence of Wnt ligand, β-catenin accumulates in the nucleus and functions as a coactivator for the transcription factors TCF/LEF, leading to the activation of Wnt responsive genes (35). The Wnt/β-catenin signaling pathway is essential for hair growth and its inhibition, driven by 5α-dihydrotestosterone through the androgen receptor, can result in hair loss in AGA (1). GNAI3 (G Protein Subunit Alpha I3) functions as a downstream transducer of G protein-coupled receptors (GPCRs) in various signaling pathways (38). GPCRs play a role in regulating skin homeostasis and maintaining

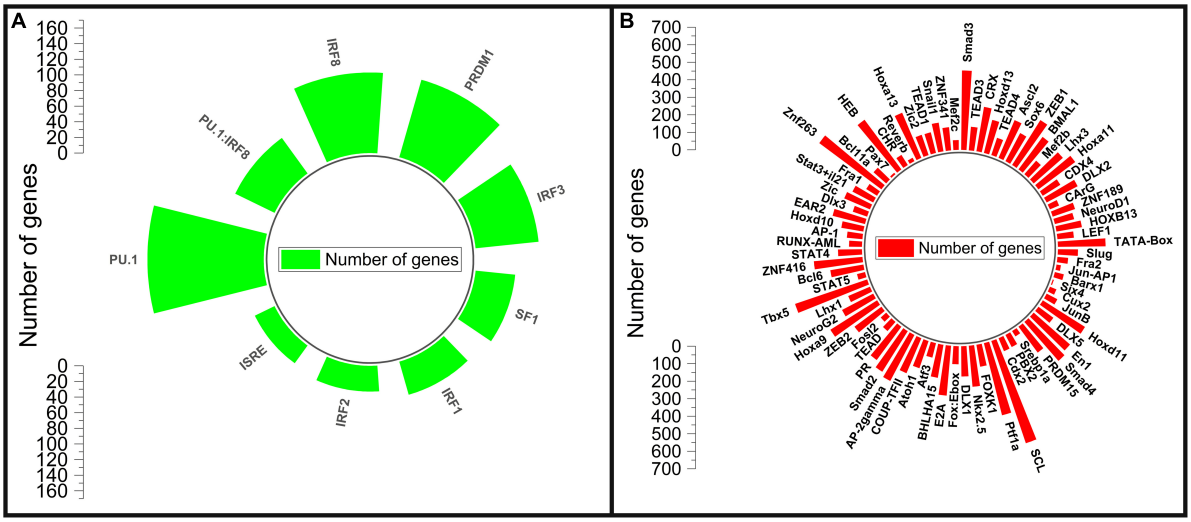


FIGURE 3
Motif enrichment analysis of the differentially expressed genes. **(A)** Motifs enriched in up-regulated genes, **(B)** motifs enriched in down regulated genes.

hair growth (39–41). NRAS (NRAS Proto-Oncogene, GTPase) is a membrane protein that travels between the plasma membrane and Golgi apparatus (42). EGF (Epidermal Growth Factor) acts as a switch in the hair growth cycle (43). It regulates the expression of hair follicle regulatory genes through Wnt//β-catenin signaling (44). Thus, these 4 downregulated hub genes which are involved in hair growth mechanisms are crucial and their downregulation in AGA is expected.

The LCK (LCK Proto-Oncogene, Src Family Tyrosine Kinase) gene, which encodes a non-receptor protein-tyrosine kinase, is a crucial signaling molecule in the selection and maturation of developing T cells and plays a key role in T cell receptor signal transduction pathways (25, 26). The up-regulation of the LCK gene is also associated with alopecia areata (27). The LYN (LYN Proto-Oncogene, Src Family Tyrosine Kinase) gene encodes a non-receptor tyrosine-protein kinase and is crucial for regulating innate and adaptive immune responses, integrin signaling, growth factor and cytokine responses, and hematopoiesis (24). BTK (Bruton Tyrosine Kinase) and plays a key role in B lymphocyte development and is a target for inflammatory diseases (45). Inhibition of BTK by inhibitors leads to changes in hair and nails texture (38). PIK3CD (Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Delta) is involved in immune system response (46). PTPN6 (Protein Tyrosine Phosphatase Non-Receptor Type 6) is critical for the function of lymphoid and myeloid cells (47). SPI1 (Spi-1 Proto-Oncogene) encodes a transcriptional activator specifically involved in the development of macrophages and B cells. This protein also regulates pre-mRNA splicing (23). STAT3 (Signal Transducer and Activator of Transcription 3) is activated by cytokines and growth factors. This gene plays an important role in maintaining the homeosis of skin (48). STAT5A (Signal Transducer and Activator of Transcription 5A) protein serves a dual function of signal transduction and activation of transcription in cells exposed to cytokine and other growth factors. This protein also mediates cellular responses to activated FGFR1, FGFR2, FGFR3 and FGFR4 (31). Also, STAT5 activation is important for hair growth phase induction in hair dermal papilla cells (DPCs) (34). VAV1 (Vav Guanine Nucleotide Exchange Factor 1) encoded protein is important in hematopoiesis and plays a role in the development and activation of T-cell and B-cell (32). PSMB8 (Proteasome 20S Subunit Beta 8) plays an important role in cellular homeostasis through selective destruction of ubiquitinated proteins. Mutations in this gene are associated with autoinflammatory responses (49). ESR1 (Estrogen Receptor 1) is a nuclear sex steroid hormone receptor which regulates many genes responsible for growth, metabolism and reproductive functions. This gene is known to express in hair follicle cells (50). HCK (HCK Proto-Oncogene, Src Family Tyrosine Kinase) participates in the regulation of innate immune responses by inducing monocyte, neutrophil, macrophage and mast cell functions. This gene is recently reported to play a role in hair regenerative potential of stem cells (51). ITGB7 (Integrin Subunit Beta 7) is an adhesion receptor which mediates signaling from the extra cellular matrix to the cell. They also function as a homing receptor for lymphocytes migration (46). LCP2 (Lymphocyte Cytosolic Protein 2) acts as a substrate for the T cell antigen receptor mediated intracellular tyrosine kinase pathway (46). PDGFRB (Platelet Derived Growth Factor Receptor Beta) gene encodes a cell surface tyrosine-protein kinase receptor for the members of the platelet-derived growth factor family. It plays an essential role in cell proliferation, differentiation, survival, chemotaxis, and migration (52). RAC2 (Rac Family Small GTPase 2) involve in phagocytosis of apoptotic cells and epithelial cell polarization (46). IRF4 (Interferon Regulatory Factor 4) regulates interferon signaling and negatively regulates Toll like receptor in the induction of innate and adaptive immune systems (42). ITGAM (Integrin Subunit Alpha M) functions as macrophage receptor and plays a key role in the adherence of monocytes and neutrophils (42). HLA-A (Major Histocompatibility Complex, Class I, A), HLA-F (Major Histocompatibility Complex, Class I, F) and HLA-E (Major Histocompatibility Complex, Class I, E) plays a central role in immune system by participating in cell presentation for recognition by T cell receptor (42). A majority of the hub genes namely PTPN6,

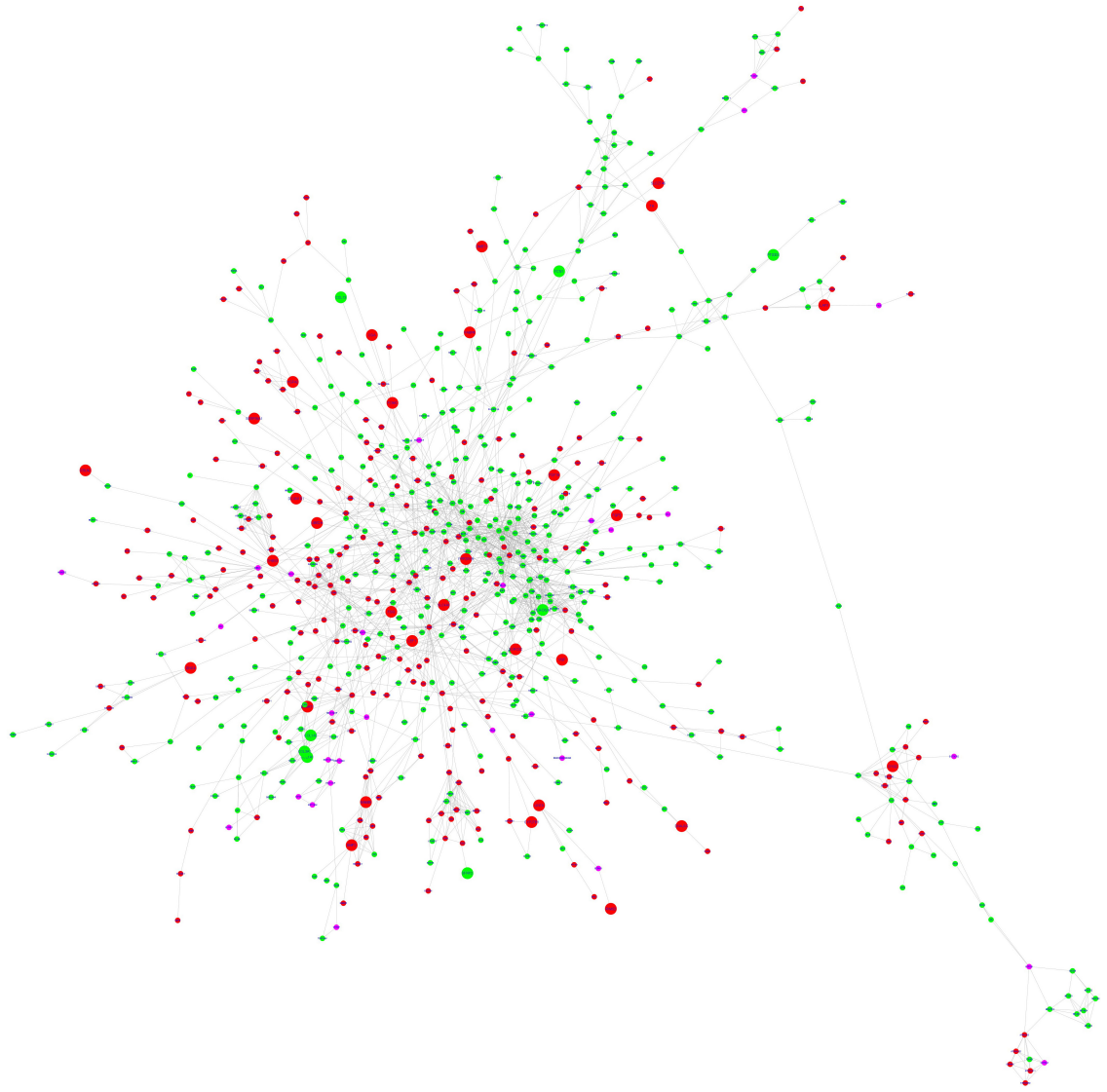


FIGURE 4
STRING protein-protein interaction network. The nodes are represented as circles and edges as lines. Red and green nodes indicate downregulated and upregulated genes, respectively. The larger nodes represent DEGs that comply with a $\log_2 FC > |1|$ value. The pink nodes indicate linker genes.

LCK, LCP2, LYN, HCK, VAV1, STAT3, STAT5A, and BTK belongs to the Src homology 2 (SH2) domain containing tyrosine kinases and participate in the immune system process.

We conducted ClueGO reactome pathway enrichment analysis for the genes that were identified by at least two algorithms of the Cytohubba analysis of the biological networks (PPI and FI) as well as 289 DEGs that met the cut-off value of $\log_2 FC > |1|$ using the ClueGo plugin v2.5.9 in Cytoscape (53). The results were presented as a network of pathways with genes participating in the pathways, which are illustrated in Figure 6. The analysis revealed pathways such as keratinization, formation of the cornified envelope, developmental biology, interferon alpha/beta signaling, cytokine signaling in the immune system, receptor tyrosine kinase signaling, PI5P, PP2A, and IER3 regulation of PI3K/AKT signaling, immunoregulatory interactions between lymphoid and non-lymphoid cells, costimulation by the CD28 family, and the GPVI-mediated activation cascade are the predominant pathways for our input genes. The enrichment of pathways involved in immune system function are consistent with our findings suggesting that immune system dysregulation plays a role in AGA pathology.

Validation of DEGs with other datasets

To validate the results of our analysis of the GEO dataset GSE90594, we compared the DEGs obtained with other datasets available in the GEO database. As of November 1, 2022, we found that no profile in the database contained samples from men with AGA and from normal haired men, except for the profile we analyzed in this study. The few available datasets related to AGA lacked control samples from normal men and the quality of the microarray and RNA-Seq data was questionable. Despite these limitations, we selected two datasets, GSE66663 (which

TABLE 4 Top 20 hub proteins identified by different topological algorithms and centralities utilizing cytohubba plugin in the STRING PPI network.

Topological algorithms				Centralities	
MCC	MNC	DMNC	Degree	Betweenness	Closeness
PSMB8	LYN	IFITM3	NRAS	CTNNB1	CTNNB1
IRF9	LCK	IFITM1	CTNNB1	SDC1	LCK
ISG15	HLA-A	ISG15	STAT3	COL4A4	STAT3
EGR1	STAT3	IFITM2	LYN	ENPP1	EGF
IFITM3	NRAS	OAS1	LCK	NRAS	LYN
IFITM1	HLA-DRB1	EGR1	HLA-A	HGF	NRAS
IFITM2	PTPN6	IRF8	PTPN6	STAT3	PTPN6
OAS1	STAT5A	HLA-G	STAT5A	PPARA	STAT5A
HLA-A	PSMB8	IRF1	HLA-DRB1	ITGAM	PDGFRB
HLA-F	HLA-F	POFUT2	EGF	PKLR	VAV1
HLA-E	CDK1	SPON1	ITGAM	LYN	HGF
IRF4	HLA-E	THSD4	B2M	H2AX	HCK
IRF1	CTNNB1	ADAMTS7	CDK1	THBS1	ITGAM
IRF8	HCK	ADAMTS1	PSMB8	TNF	ESR1
HLA-G	CCNB1	CFP	HLA-F	EGF	AR
CFP	FGR	ADAMTS17	HLA-E	HIF1A	HIF1A
POFUT2	VAV1	ADAMTS10	CCNB1	ACSL1	CXCL12
SPON1	CCNA2	THBS2	PTPRC	PPARG	HLA-A
THSD4	IRF4	THBS1	VAV1	RACK1	PTPRJ
ADAMTS7	LCP2	IRF9	IRF4	QPRT	SFN

The highlighted genes are present in more than two columns, as indicated by the color code: Violet denotes presence in 4 columns, blue denotes presence in 3 columns, and green denotes presence in 2 columns.

includes hTERT-immortalized DPCs derived from balding frontal and non-balding occipital scalp samples from men with AGA) and GSE212301 (which contains RNA-Seq data from balding vertex and non-balding occipital scalp samples of 10 men with AGA), and performed differential gene expression analyses. The common DEGs between the datasets are presented in the **Supplementary II-18**. We discovered 490 DEGs were common between GSE90594 and GSE66663 dataset in which 190 genes were differentially regulated in same directions. Whereas 180 DEGs were common between GSE90594 and GSE212301 dataset in which 44 genes were differentially regulated in same directions.

Discussion

Differential gene expression analysis is a technique used to identify genes whose expression levels change significantly between two or more experimental conditions or samples using the data generated from microarray or RNA sequencing experiments. This approach helps to determine which genes are upregulated or downregulated in response to a specific condition, such as a disease state or treatment, which facilitates understanding of the underlying molecular mechanisms of diseases (54). In this study we analyzed gene expression data from the scalps of 9 individuals with premature AGA and 10 normal volunteers from the GEO database profile GSE90594 to identify core genes associated with AGA (5). In Michel et al. (5) analysis report, the authors performed differential gene expression analysis on all 28 samples (14 alopecia and 14 normal samples) using ANOVA and Tukey's *post-hoc* tests. After applying the Benjamini-Hochberg correction for multiple testing, they identified 333 DEGs consisting of 184 downregulated and 149 upregulated genes. The authors selected the DEGs using a cut-off of fold change $\geq \pm 1.5$ ($\log_2 FC \geq \pm 0.58$) and $p \leq 0.05$ for significance (5). In our analysis, we normalized the microarrays, removed the outlier samples, and performed the differential gene expression analysis using the single-channel design matrix provided in the *limma* package. We used Benjamini and Hochberg's method to compute the adjusted p-values (FDR or q-value) and considered probes with $q \leq 0.05$ to be significant. In our analysis, the fold change values of AGA-associated genes known to play a crucial role in disease pathology, such as AR ($\log_2 FC = 0.33$), CTNNB1 ($\log_2 FC = -0.58$), TGFB2 ($\log_2 FC = -0.58$), and SRD5A2 ($\log_2 FC = 0.56$) between the AGA patients and healthy group were lower. To thoroughly examine the pathology of AGA, we adopted a stringent criterion of $\log_2 FC \geq \pm 0.3$ with a strict FDR value ($q \leq 0.05$) and obtained 2,439 DEGs, taking into account that subtle differences in gene expression can have a significant biological impact and that some genes are more sensitive to changes in dosage (55, 56).

To shed light on the biological roles and processes associated with the 2,439 DEGs, we performed gene family enrichment, GO (biological process, molecular function, and cellular component)

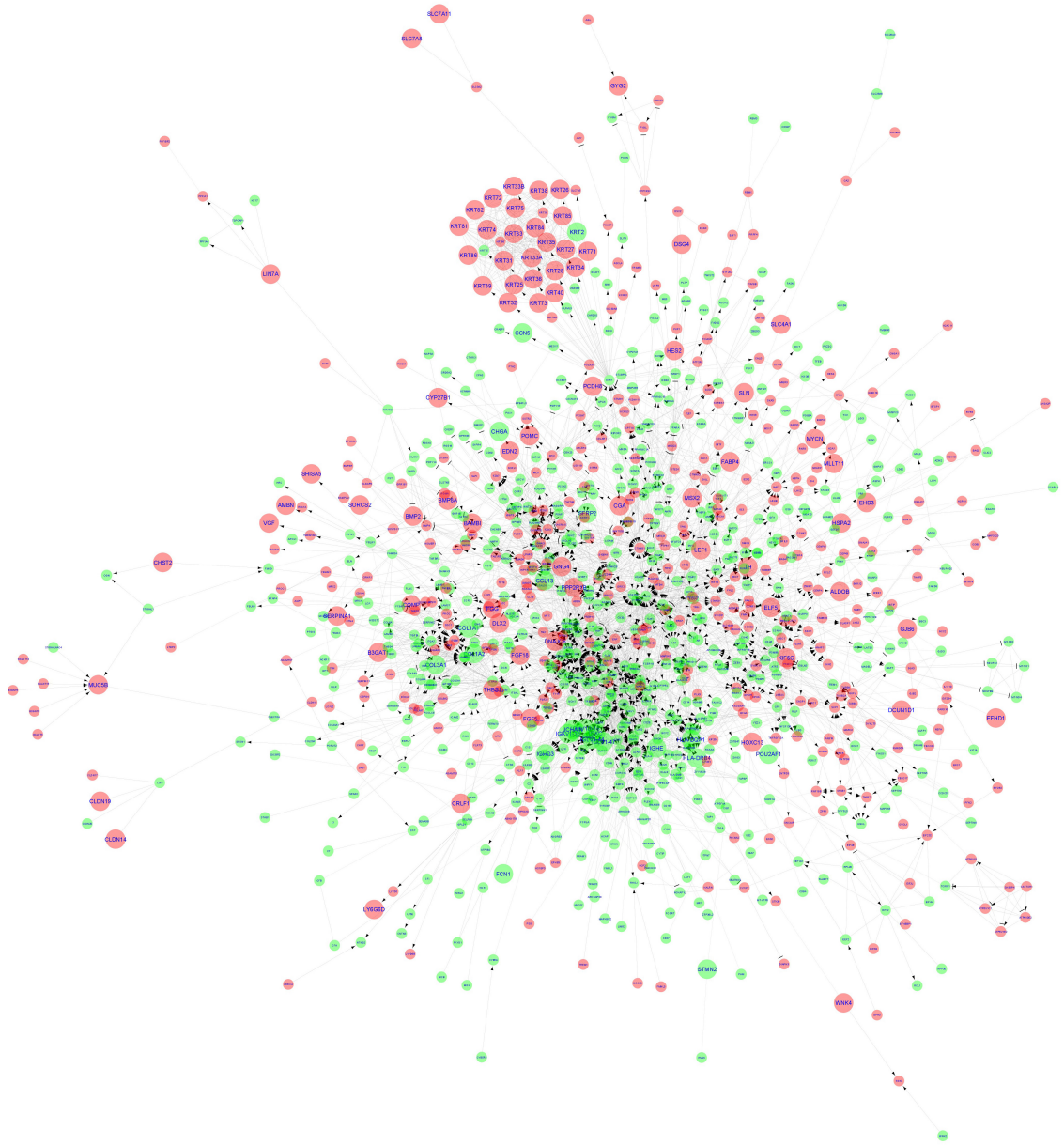


FIGURE 5
Functional interaction network generated by ReactomeFIViz app. The nodes are represented as circles and edges as lines. Red and green nodes indicate downregulated and upregulated genes, respectively. The larger nodes represent DEGs that comply with a $\log_2 FC > |1|$ value.

enrichment, and pathway enrichment analyses. Our results revealed that the down-regulated genes belonged to gene families such as keratins, keratin-associated proteins, frizzled receptors, Bone morphogenetic proteins, Wnt, and metallothioneins (Figure 2). The GO enrichment analysis indicated that these down-regulated genes play vital roles in the structural constituents of the skin epidermis, hair follicle development and hair cycle (Table 1). The pathway enrichment analysis showed that these down-regulated genes participate in the keratinization pathway (Table 2). On the other hand, the up-regulated genes were enriched for CD molecules, Immunoglobulin-like domains, Rho GTPase-activating proteins, receptor tyrosine kinases, minor histocompatibility antigens, and selenoproteins as the top gene families (Figure 2). The GO enrichment analysis also demonstrated that these up-regulated genes were involved in MHC protein complex binding, leukocyte activation, regulation of the immune response, and T-cell activation (Table 1). The pathway enrichment analysis found that the up-regulated genes participated in the innate and adaptive immune systems, cytokine signaling, and interferon signaling pathways (Table 2).

The identification of genetic variants associated with AGA is critical for understanding its etiology. In this study, we annotated the coordinates of AGA-associated genomic loci with our DEGs to identify the potential candidate genes contributing to AGA pathology. Our analysis identified several DEGs located within or near reported AGA risk loci such as MEMO1, SRD5A2, FOXL2NB, FGF5, DKK2, EBF1, IRF4, CENPW, and PAGE2. These findings support the existing knowledge of the association between these

TABLE 5 Top 20 hub proteins identified by different topological algorithms and centralities utilizing cytohubba plugin in the reactome FI network.

Topological algorithms				Centralities	
MCC	MNC	DMNC	Degree	Betweenness	Closeness
LCP2	STAT3	LAT2	STAT3	ESR1	STAT3
HLA-DRB1	LYN	HLA-DOA	CTNNB1	CTNNB1	CTNNB1
HLA-DPA1	CTNNB1	CD74	SPI1	STAT3	PIK3CD
HLA-DRB4	LCK	SGO1	ESR1	SPI1	PTPN6
HLA-DQA1	PIK3CD	SKA2	PIK3CD	PIK3CD	SPI1
HLA-DRB3	SPI1	ZWINT	LYN	GNAI3	LYN
HLA-DPB1	PTPN6	C1R	LCK	PTPN6	ESR1
LCK	NRAS	C1S	NRAS	NRAS	STAT5A
CD3E	GNAI3	LCP1	GNAI3	HIF1A	LCK
ZAP70	VAV1	CIITA	PTPN6	EGR1	NRAS
IGKC	HCK	KIF26A	RAC2	ITGB7	EGF
IGKV1-16	EGF	C2	HCK	GATA2	GNAI3
IGLV1-44	PDGFRB	CELSR1	VAV1	AR	HCK
IGLV1-47	LCP2	PCDHB4	EGF	LYN	PDGFRB
IGKV1D-16	RAC2	PCDH7	STAT5A	RAC2	HIF1A
LYN	STAT5A	PCDH8	PDGFRB	TGFBR2	EGR1
VAV1	BTK	DCHS1	ITGB7	ITGAM	AR
BTK	ITGB7	KIF4A	LCP2	TNF	VAV1
HLA-DMB	CD3E	CDH23	ITGAM	STAT5A	CRKL
HLA-DOA	HLA-DRB1	PCDH11Y	BTK	LCK	BTK

The highlighted genes are present in more than two columns, as indicated by the color code: Red denotes presence in 5 columns, violet denotes presence in 4 columns, blue denotes presence in 3 columns, and green denotes presence in 2 columns.

genes and AGA pathology. Furthermore, our analysis (Table 3) mapped several DEGs including HOXD9, LHPP, CRHR1, STH, and PAGE2B, which are of unknown significance in hair growth, with AGA risk loci in GWAS studies. These genes warrant further investigation. Moreover, the enrichment of many DEGs identified in our analysis within the 500 kb window of AGA risk loci revealed that the genes which have not yet been identified as AGA risk loci could play a critical role in AGA pathology.

In our analysis to identify specific sequence motifs or patterns in the promoter regions of the DEGs, we found several enriched motifs for the down-regulated genes, including those involved in the Wnt/β-catenin signaling pathway (LEF1), TGF-β signaling (SMAD2, SMAD3, and SMAD4), nervous system development (NeuroD1 and NeuroG2), development (HOXB13, HOXD10, HOXA13, HOXA11, HOXD11, and HOXD13), Jun/FOS family (JunB, Jun-AP1, AP-2 gamma, Fosl2, and AP-1), and FOX family (FOXK1, and Fox:Ebox). Among the down-regulated DEGs, we observed the presence of LEF1, SMAD6, SAMD7, HOXA3, HOXC13, FOXN1, FOXE1, and FOXI2. In contrast, the transcription factors such as NEUROD2, HOXD1, HOXD9, FOXL2, and FOXL2NB were up-regulated, confirming that the down-regulation of hair-related genes in AGA may be primarily due to the Wnt/β-catenin signaling component LEF1 (1).

Furthermore, our motif analysis revealed that the top motifs enriched for the up-regulated genes were those for immune system-related transcription factors, such as IRF1, IRF2, IRF3, IRF8, PRDM1, SPI-1 (PU.1), and SF1. Among the up-regulated DEGs, we observed the presence of several IRF family of transcription factors, including IRF1, IRF1-AS1, IRF4, IRF8, IRF9, and SPI. Specifically, IRF1 is critical for apoptosis and the target genes of IRF1 are responsible for apoptotic responses. IRF4 and IRF8 regulate myeloid cell development, while IRF9 mediates STAT1/STAT2 function in downstream signaling of type I IFN receptor signaling and is also involved in autoantibody production (35, 55). These findings suggest that these immune transcription factors may play a role in the up-regulation of immune response genes, implying a heightened immune system activity and immune response against hair growth cycle in the scalp in AGA. Taken together, our results provide further evidence that the genes for hair follicle development and hair cycle are down-regulated, while genes for immune response are up-regulated in the balding scalps of AGA.

The occurrence of inflammatory phenomena in AGA pathogenesis has been reported earlier, but the cause of the inflammation was unknown. Consequently, the role of inflammation in AGA was not heavily emphasized in the past (52, 57). Despite the general belief that scalp inflammation results in folliculitis, perifollicular fibrosis, and destructive scarring alopecia, studies have linked inflammation to male pattern baldness (55–57). Jaworsky et al. (58) discovered the presence of activated T-cell infiltrate in hair follicles and found that these infiltrates were associated with class II antigens. In 2001, Young et al. (56) discovered granular immunoglobulin M and C3 at the basement

TABLE 6 Hub genes identified common in the STRING PPI and reactome FI network.

Sr. no.	Gene symbol	Gene name	Gene family	Gene function summary from Uniprot	Expression direction
1	CTNNB1	Catenin beta 1	Armadillo family of proteins	Important for Wnt signaling and cell adhesion	DOWN
2	EGF	Epidermal growth factor	Epidermal growth factor family	Important for cell growth and differentiation	DOWN
3	GNAI3	Guanine nucleotide-binding protein G(I) Subunit Alpha-3	G protein alpha inhibitory subunit family	Regulates diverse signaling pathways	DOWN
4	NRAS	NRAS proto-oncogene, Gtpase	Ras family of small GTPases	Involves in regulating cell growth and differentiation	DOWN
5	BTK	Bruton tyrosine kinase	Tec family of non-receptor tyrosine kinases	Critical for B cell development and activation	UP
6	ESR1	Estrogen receptor 1	Nuclear receptor family	Acts as a transcription factor for estrogen signaling	UP
7	HCK	HCK proto-oncogene, Src family tyrosine kinase	Src family of non-receptor tyrosine kinases	Has roles in immune cell signaling and activation	UP
8	ITGB7	Integrin subunit beta 7	Integrin family of cell adhesion molecules	Important for immune cell trafficking and activation	UP
9	LCK	LCK proto-oncogene, Src family tyrosine kinase	Src family of non-receptor tyrosine kinases	Plays a critical role in T cell development and activation.	UP
10	LCP2	Lymphocyte cytosolic protein 2	SLP-76 family of adapter proteins	Essential for T cell receptor signaling and activation	UP
11	LYN	LYN proto-oncogene, Src family tyrosine kinase	Src family of non-receptor tyrosine kinases	Functions in B cell signaling and immune responses.	UP
12	PDGFRB	Platelet derived growth factor receptor beta	Rho family of small GTPases	Involves in actin cytoskeleton organization and cell migration	UP
13	PIK3CD	Phosphatidylinositol-4,5-bisphosphate 3-kinase Catalytic Subunit Delta	Phosphoinositide 3-kinase catalytic subunit family	Plays a role in various signaling pathways	UP
14	PTPN6	Protein tyrosine phosphatase non-receptor type 6	Protein tyrosine phosphatase family	Regulates immune cell signaling and homeostasis	UP
15	RAC2	Rac family small Gtpase 2	Rho family of small GTPases	Involves in actin cytoskeleton organization and cell migration	UP
16	SPI1	Spi-1 proto-oncogene	ETS family of transcription factors	Essential for hematopoietic development and differentiation	UP
17	STAT3	Signal transducer and activator Of transcription 3	STAT family of transcription factors	Involves in cytokine signaling and immune responses	UP
18	STAT5A	Signal transducer and activator Of transcription 5A	STAT family of transcription factors	Important for immune cell development and activation	UP
19	VAV1	Vav guanine nucleotide exchange factor 1	Vav family of guanine nucleotide exchange factors	Regulates signaling pathways downstream of receptors	UP
20	PSMB8	Proteasome 20s subunit beta 8	Proteasome beta subunit family	Involves in protein degradation and antigen presentation	UP
21	HLA-A	Major histocompatibility complex, Class I, A	Human leukocyte antigen (HLA) family	Involves in antigen presentation and immune responses	UP
22	HLA-F	Major histocompatibility complex, class I, F	Human leukocyte antigen (HLA) family	Involves in immune tolerance and immune responses	UP
23	HLA-E	Major histocompatibility complex, class I, E	Human leukocyte antigen (HLA) family	Involves in antigen presentation and immune regulation	UP
24	IRF4	Interferon regulatory factor 4	Interferon regulatory factor family	Involves in immune cell differentiation and function	UP
25	ITGAM	Integrin subunit alpha M	Integrin family of cell adhesion molecules	Important for leukocyte function and immune responses	UP

10.3389/fmed.2023.1108358

TABLE 7 Result of GO biological process and reactome pathway enrichment analysis of hub genes from ToppGene Suite (FDR < 0.05).

GO ID	Biological process term	Gene count	Up-regulated genes	Down-regulated genes
GO:0002682	Regulation of immune system process	20	IRF4, PTPN6, LCK, SPI1, HLA-A, LCP2, LYN, ITGAM, PIK3CD, HCK, VAV1, ESR1, STAT3, BTK, STAT5A, RAC2, HLA-E, HLA-F	CTNNB1, NRAS
GO:0042110	T cell activation	15	IRF4, PTPN6, LCK, SPI1, HLA-A, LYN, ITGAM, PIK3CD, VAV1, STAT3, STAT5A, RAC2, HLA-E, HLA-F	CTNNB1
GO:0050778	Positive regulation of immune response	15	PTPN6, LCK, SPI1, HLA-A, LCP2, LYN, ITGAM, PIK3CD, HCK, VAV1, BTK, STAT5A, HLA-E, HLA-F	NRAS
GO:0043299	Leukocyte degranulation	10	SPI1, HLA-A, LYN, ITGAM, PIK3CD, HCK, BTK, RAC2, HLA-E, HLA-F	
GO:0002764	Immune response-regulating signaling pathway	14	IRF4, PTPN6, LCK, HLA-A, LCP2, LYN, PIK3CD, HCK, VAV1, ESR1, BTK, HLA-E, HLA-F	NRAS
GO:1903131	Mononuclear cell differentiation	14	IRF4, PTPN6, LCK, SPI1, LYN, PIK3CD, VAV1, STAT3, BTK, STAT5A, RAC2, HLA-E, HLA-F	CTNNB1
GO:0045321	Leukocyte activation	17	IRF4, PTPN6, LCK, SPI1, HLA-A, LCP2, LYN, ITGAM, PIK3CD, VAV1, STAT3, BTK, STAT5A, RAC2, HLA-E, HLA-F	CTNNB1
GO:0002521	Leukocyte differentiation	15	IRF4, PTPN6, LCK, SPI1, LYN, ITGAM, PIK3CD, VAV1, STAT3, BTK, STAT5A, RAC2, HLA-E, HLA-F	CTNNB1
GO:0046649	Lymphocyte activation	16	IRF4, PTPN6, LCK, SPI1, HLA-A, LYN, ITGAM, PIK3CD, VAV1, STAT3, BTK, STAT5A, RAC2, HLA-E, HLA-F	CTNNB1
GO:0002768	Immune response-regulating cell surface receptor signaling pathway	12	PTPN6, LCK, HLA-A, LCP2, LYN, PIK3CD, HCK, VAV1, BTK, HLA-E, HLA-F	NRAS
Biosystems ID	**Reactome pathway name**	**Gene count**	**Up-regulated genes**	**Down-regulated genes**
1269310	Cytokine signaling in immune system	17	PSMB8, IRF4, PTPN6, LCK, HLA-A, LYN, ITGAM, PDGFRB, PIK3CD, HCK, VAV1, STAT3, STAT5A, HLA-E, HLA-F	NRAS, EGF
1269318	Signaling by interleukins	14	PSMB8, IRF4, PTPN6, LCK, LYN, ITGAM, PDGFRB, PIK3CD, HCK, VAV1, STAT3, STAT5A	NRAS, EGF
1269171	Adaptive immune system	15	PSMB8, PTPN6, LCK, HLA-A, LCP2, LYN, PDGFRB, PIK3CD, ITGB7, VAV1, BTK, HLA-E, HLA-F	NRAS, EGF
1269357	GPVI-mediated activation cascade	7	PTPN6, LCK, LCP2, LYN, PIK3CD, VAV1, RAC2	
1269183	Signaling by the B cell receptor (BCR)	10	PSMB8, PTPN6, LCK, LYN, PDGFRB, PIK3CD, VAV1, BTK	NRAS, EGF
1269487	Signaling by SCF-KIT	11	PSMB8, PTPN6, LCK, LYN, PDGFRB, PIK3CD, VAV1, STAT3, STAT5A	NRAS, EGF
1269323	Interleukin-3, 5 and GM-CSF signaling	10	PSMB8, PTPN6, LYN, PDGFRB, PIK3CD, HCK, VAV1, STAT5A	NRAS, EGF
1269203	Innate immune system	16	PSMB8, PTPN6, LCK, HLA-A, LCP2, LYN, ITGAM, PDGFRB, PIK3CD, HCK, VAV1, BTK, HLA-E	NRAS, EGF
1269284	DAP12 signaling	10	PSMB8, LCK, LCP2, PDGFRB, PIK3CD, VAV1, BTK, HLA-E	NRAS, EGF
1268855	Diseases of signal transduction	10	PSMB8, LCK, PDGFRB, PIK3CD, VAV1, STAT3, STAT5A	CTNNB1, NRAS, EGF

10.3389/fmed.2023.1108358

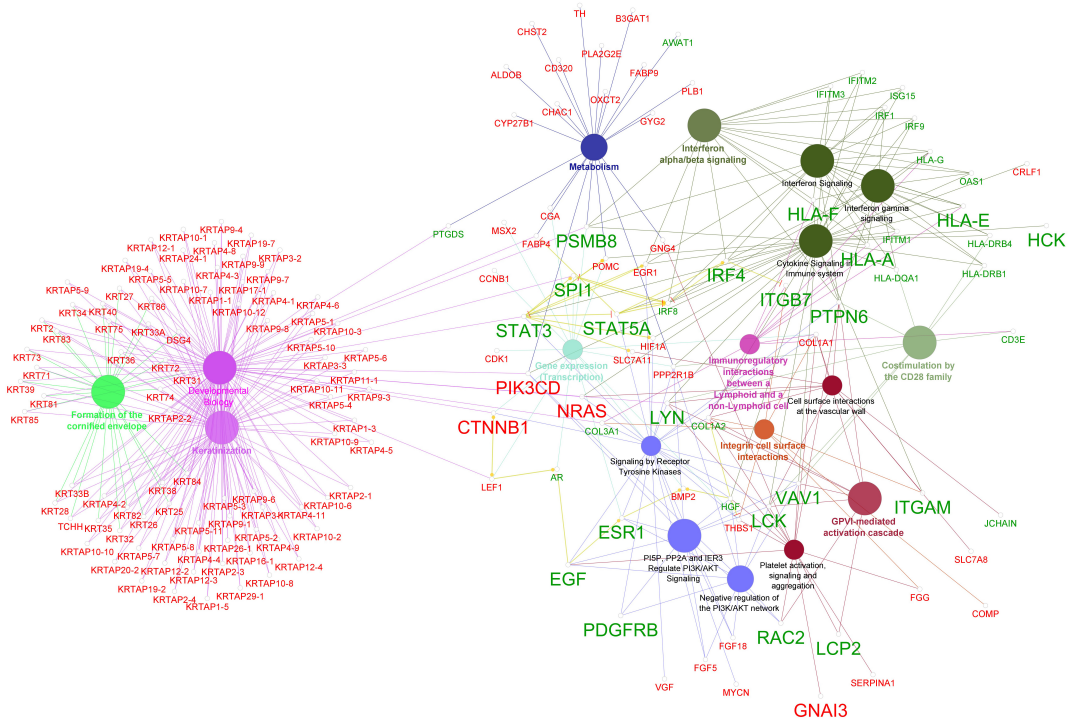


FIGURE 6
Enrichment of reactome pathway terms for the hub genes and DEGs that comply $\log_2 FC > |1|$ using Cytohubba plugin ClueGO. The network shows the connectivity between pathway terms based on shared functional nodes and edges among the DEGs with a kappa score of 0.4. The color of the nodes and edges represents their specific functional classes, and only significant values with $p < 0.05$ are shown in the enrichment. Nodes labeled in red represent down-regulated genes, while nodes labeled in green indicate up-regulated genes. The larger labeled nodes are the hub genes.

membrane, as well as porphyrins in the pilosebaceous canal in biopsy specimens from the bald scalps of AGA patients. They suggested that the local microbiologic flora and environmental factors like UV light could be responsible for the inflammatory reactions (56). Mahe et al. (59) proposed in a 2001 review that the inflammatory process associated with AGA be referred to as microinflammation in contrast to classical inflammatory process. Furthermore, the presence of peripilar signs around the hair follicle ostium, which reflect perifollicular inflammation, has established the presence of follicular microinflammation in AGA (60, 61). Despite these findings, the underlying biological reason, pathways, and genes involved in the inflammatory process of AGA have not yet been elucidated.

In order to deepen our understanding of the inflammatory mechanisms in AGA, we constructed gene interaction networks using the DEGs identified in our study. The Cytoscape plugins StringApp and ReactomeFIplugin were utilized to construct the PPI and FI networks, respectively. The DEGS in the PPI network were connected based on their protein-protein interactions obtained from the STRING database, while the DEGs in the FI network were linked based on their involvement in signaling pathways from the Reactome database. The integrated tools within the Cytoscape StringApp and ReactomeFI plugins were utilized to perform GO and pathway enrichment analyses for both networks. The results were consistent with our previous GO, pathway, and motif enrichment analyses. In addition, a Cytohubba analysis was conducted to identify the hub genes of the biological networks. The hub genes were sorted based on their occurrence in more than one algorithm used in the analysis. As a result, 15 genes (LYN, HLA-A, STAT3, NRAS, CTNNB1, PSMB8, HLA-F, HLA-E, IRF4, LCK, PTPN6, STAT5A, VAV1, EGF, and ITGAM) were identified as key hub genes in the PPI network. Similarly, 19 genes (LCK, LYN, BTK, CTNNB1, GNAI3, NRAS, PIK3CD, PTPN6, SPI1, STAT3, STAT5A, VAV1, EGF, ESR1, HCK, ITGB7, LCP2, PDGFRB, and RAC2) were recognized as key hub genes in the FI network as they were consistently identified across multiple algorithms.

To explore the connections between hub genes and DEGs exhibiting $\log_2 FC > |1|$, we performed reactome pathway enrichment analysis using the ClueGo plugin in Cytoscape (53). The analysis revealed that the hub genes were strongly associated with several important pathways related to immune system functions including interferon signaling, cytokine signaling, GPVI-mediated activation cascade, PI3K/AKT signaling, and signaling by receptor tyrosine kinases (**Figure 6**). Interestingly, recent research has linked the activation of the PI3K/Akt pathway with the apoptosis of hair follicle stem cell (HFSC) mediated by 5α-DHT in AGA (62, 63). Furthermore, the ClueGo network (**Figure 6**) showed that several genes including the hub genes PSMB8, SPI1, STAT3, PIK3CD, NRAS, CTNNB1, and LEF1 connect the keratinization process with inflammatory process terms suggesting that AGA is driven by a complex interplay between various molecular pathways involving immune system dysregulation and abnormal keratinization.

A significant number of the up-regulated hub genes identified in our study such as PTPN6, LCK, LCP2, LYN, HCK, VAV1, STAT3, BTK, and STAT5A belong to the Src Homology 2 (SH2) domain

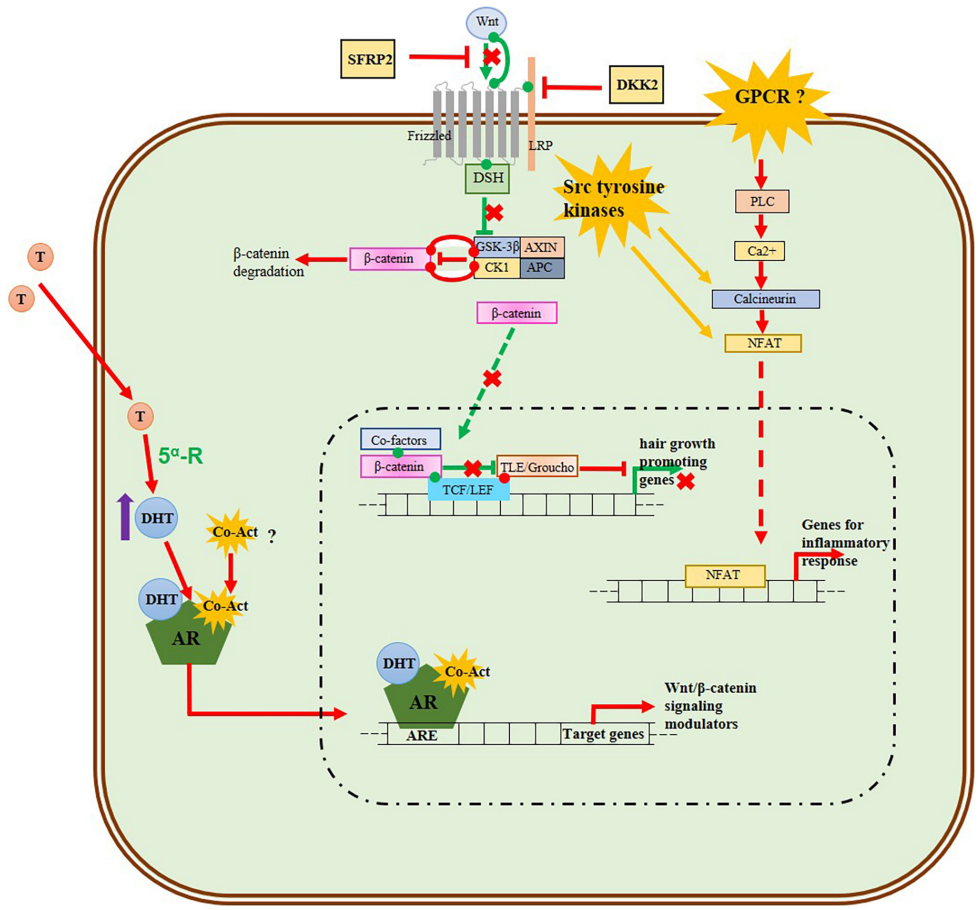


FIGURE 7
Schematic model of 5α-DHT mediated AGA in DPCs including Wnt/β-catenin signaling pathway and up-regulated inflammatory process. The green lines and arrows represent the normal activated Wnt/β-catenin signaling in the DPCs of normal-haired scalp, while the red lines and arrows denote the behavior of signaling pathways in the DPCs of balding scalp. The androgen 5α-DHT-AR complex inhibits the Wnt signaling by transcribing Wnt/β-catenin inhibitors. In our analysis genes for frizzled receptor, Wnt ligands (Wnt2b, Wnt3, Wnt5a, Wnt10b, and Wnt11), beta-catenin, LEF, and TCF are down-regulated, while the Wnt inhibitors DKK2 and SFRP2 are upregulated implying the downregulation of the normal Wnt/β-signaling pathway in AGA. The genes for phospholipase, calcineurin, and NFAT which function downstream of the non-canonical Wnt/Calcium pathway are upregulated, while the Wnt ligand for this pathway Wnt5a and frizzled receptor in the upstream are down-regulated. We propose that Src tyrosine kinase, known to interact with phospholipase and calcineurin, may activate this Wnt/Calcium pathway pathway and this needs further investigation. In addition, other Wnt ligands such as Wnt3a, Wnt4, and Wnt16 are up-regulated in our analysis and these ligands or some GPCR receptors may play a role in the activation of Wnt/calcium signaling pathway and mediate the up-regulation of NFAT which transcribes inflammatory process genes.

gene family. This group of genes encodes proteins containing SH2 domains, which can recognize and bind to phosphorylated tyrosine residues in other proteins. SH2 domain-containing proteins participate in signal transduction pathways serving as adapter molecules linking tyrosine phosphorylation events to downstream signaling pathways (64). Further of the four non-receptor tyrosine kinase hub genes (BTK, HCK, LCK and LYN), three genes namely HCK, LCK, and LYN belong to the Src family of protein tyrosine kinases (65). Recent studies have highlighted the potential role of Src tyrosine kinase in hair growth. One study found that Src inhibition promotes melanogenesis, leading to the production of hair color pigment melanin (66). In another study the flavonoid quercitrin was shown to stimulate hair growth in cultured DPCs by activating several signal transduction elements, including receptor tyrosine kinases and non-receptor tyrosine kinases. Specifically, Src family proteins such as CSK, FRK, HCK, and SRMS, which were not differentially expressed in our analysis, were found to be activated by quercitrin while promoting the hair growth (67). Additionally, recent researches have shown that Src tyrosine kinase can cross-talk with Wnt signaling (65) and with androgen receptor (AR) signaling (66) suggesting a potential interplay between Src tyrosine kinase and androgen-DHT and Wnt/β-catenin signaling in the balding scalps of AGA. Therefore, we suggest that further investigation into the potential interactions between Src tyrosine kinase family genes, AR-5α-DHT, Wnt/β-catenin signaling, and the inflammatory response is needed to gain a more comprehensive understanding of AGA pathogenesis.

The Hair follicle is a fascinating mini-organ that continuously undergoes cycles of growth (anagen), regression (catagen), resting (telogen), and shedding (exogen). This process is regulated by a number of signaling cascades, including Wnt/β-catenin, Sonic Hedgehog (SHH), bone morphogenetic protein (BMP), notch, transforming growth factor β (TGF-β), NF-κB, and fibroblast growth factors (FGFs), which coordinate communication between the epithelial and mesenchymal cells in the hair follicle (68). Although it is well-known that androgen 5α-DHT modulates

the Wnt/β-catenin signaling pathway in DPCs and inhibits the transcription of hair growth genes in AGA, less is known about the behavior of other hair growth signaling pathways in AGA (1). In this study, we identified several DEGs involved in Wnt/β-catenin, NF-κB, TGF-β, BMP, and Vitamin D metabolism signaling pathways more than the original analysis by Michel et al. (5) (**Supplementary I-3**). Our network analysis also identified core genes that could further elucidate the pathogenesis of AGA, with a focus on the upregulated inflammatory response.

Conclusively, to gain a better understanding of the pathogenesis of AGA a schematic model of 5α-DHT mediated AGA in DPCs including the Wnt/β-catenin signaling pathway and the up-regulated inflammatory process is proposed in **Figure 7**. The nuclear factor associated with T cells (NFAT) family of transcription factors controls the expression of proinflammatory genes. The Calcineurin-NFAT signaling pathway regulates the immune system and inflammatory response (69–71). The NFAT and Wnt pathways are shown to reciprocally regulate each other constituting a non-canonical Wnt/Ca2 + /NFAT pathway in certain cells and tissues for coordinating their effects on cell growth and differentiation (71, 72). In addition, Src tyrosine kinase gene LCK are shown to interact with calcineurin and NFAT promoting NFAT activity (70, 73–75). Also the Src tyrosine kinase genes such as LCK and LYN promotes cytosolic accumulation of Ca2+ which activates calcineurin (76). In the non-canonical Wnt/Ca2+ signaling pathway calcineurin and NFAT acts downstream, but our analysis shown that the genes coding them are upregulated and the genes in the up-stream of the pathway are down-regulated. Given that the Src tyrosine kinases cross-talk with Wnt signaling and that increased activity of Src is seen during aberrant Wnt signaling in many diseases (77), we suggest that Src-tyrosine kinases may cross-talk with the androgen 5α-DHT modulated Wnt Signaling pathway and promote inflammatory response. Therefore, further investigation into the potential interactions between Src tyrosine kinase family genes, AR-5α-DHT, Wnt/β-catenin signaling, and the inflammatory response is needed to gain a more comprehensive understanding of AGA pathogenesis.

Conclusion

Differential gene expression analysis is a powerful technique for identifying genes associated with specific conditions such as AGA. In this study, we analyzed the gene expression data from the scalps of individuals with premature AGA and normal volunteers to identify core genes associated with AGA. We identified 2,439 DEGs using a stringent criterion of $\log_2 FC \geq \pm 0.3$ with a strict FDR value and performed gene family enrichment, GO enrichment, pathway enrichment, and motif analysis for the DEGs. Our findings indicate that down-regulated genes in AGA play significant roles in the structural makeup of the skin epidermis, hair follicle development, and hair cycle, while up-regulated genes are implicated in the innate and adaptive immune systems, cytokine signaling, and interferon signaling pathways. Moreover, we identified potential candidate genes that may contribute to AGA pathology and require further investigation. Our study also highlights the critical role of Src family tyrosine kinases in AGA pathology. Overall, this study enhances our understanding of the underlying molecular mechanisms of AGA and may lead to the development of new therapeutic strategies for treating this condition.

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

AP: conceptualization, design of study, analysis and interpretation of data, and writting the manuscript. BR: conceiving and supervising the study and reviewing the manuscript. Both authors contributed to the article and approved the submitted version.

Funding

AP acknowledges the financial support provided through the Senior Research Fellowship by the Council of Scientific and Industrial Research, New Delhi, India [Award no. 09/844(0045)2017-EMR-I].

Acknowledgments

The authors acknowledge the Vellore Institute of Technology, Vellore, India for providing the computational facilities.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1108358/full#supplementary-material

References

1. Premanand A, Reena Rajkumari B. Androgen modulation of Wnt/β-catenin signaling in androgenetic alopecia. *Arch Dermatol Res.* (2018) 310:391–9. doi: 10.1007/s00403-018-1826-8

2. Heilmann-Heimbach S, Hochfeld L, Paus R, Nöthen M. Hunting the genes in male-pattern alopecia: how important are they, how close are we and what will they tell us? *Exp Dermatol.* (2016) 25:251–7. doi: 10.1111/exd.12965

3. Premanand A, Rajkumari B. In silico analysis of gene expression data from bald frontal and haired occipital scalp to identify candidate genes in male androgenetic alopecia. *Arch Dermatol Res.* (2019) 311:815–24. doi: 10.1007/s00403-019-01973-2

4. Jain R, De-Eknamkul W. Potential targets in the discovery of new hair growth promoters for androgenic alopecia. *Expert Opin Ther Targets.* (2014) 18:787–806. doi: 10.1517/14728222.2014.922956

5. Michel L, Reygagne P, Benech P, Jean-Louis F, Scalvino S, Ly Ka So S, et al. Study of gene expression alteration in male androgenetic alopecia: evidence of predominant molecular signalling pathways. *Br J Dermatol.* (2017) 177:1322–36. doi: 10.1111/bjd.15577

6. Barrett T, Wilhite S, Ledoux P, Evangelista C, Kim I, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* (2013) 41:D991–5. doi: 10.1093/nar/gks1193

7. Ritchie M, Phipson B, Wu D, Hu Y, Law C, 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

8. 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.

9. Chen J, Bardes E, Aronow B, Jegga A. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* (2009) 37:W305–11. doi: 10.1093/nar/gkp427

10. Xu Q, Fu R, Yin G, Liu X, Liu Y, Xiang M. Microarray-based gene expression profiling reveals genes and pathways involved in the oncogenic function of REG3A on pancreatic cancer cells. *Gene.* (2016) 578:263–73. doi: 10.1016/j.gene.2015.12.039

11. Richards J, Yuan X, Geller F, Waterworth D, Bataille V, Glass D, et al. Male-pattern baldness susceptibility locus at 20p11. *Nat Genet.* (2008) 40:1282–4. doi: 10.1038/ng.255

12. Hillmer A, Brockschmidt F, Hanneken S, Eigelshoven S, Steffens M, Flaquer A, et al. Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nat Genet.* (2008) 40:1279–81. doi: 10.1038/ng.228

13. Li R, Brockschmidt F, Kiefer A, Stefansson H, Nyholt D, Song K, et al. Six novel susceptibility Loci for early-onset androgenetic alopecia and their unexpected association with common diseases. *PLoS Genet.* (2012) 8:e1002746. doi: 10.1371/journal.pgen.1002746

14. Brockschmidt F, Heilmann S, Ellis J, Eigelshoven S, Hanneken S, Herold C, et al. Susceptibility variants on chromosome 7p21.1 suggest HDAC9 as a new candidate gene for male-pattern baldness. *Br J Dermatol.* (2011) 165:1293–302. doi: 10.1111/j.1365-2133.2011.10708.x

15. Pirastu N, Joshi P, de Vries P, Cornelis M, McKeigue P, Keum N, et al. GWAS for male-pattern baldness identifies 71 susceptibility loci explaining 38% of the risk. *Nat Commun.* (2017) 8:1584. doi: 10.1038/s41467-017-01490-8

16. Jiang L, Zheng Z, Fang H, Yang J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet.* (2021) 53:1616–21. doi: 10.1038/s41588-021-00954-4

17. Pickrell J, Berisa T, Liu J, Ségurel L, Tung J, Hinds D. Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet.* (2016) 48:709–17. doi: 10.1038/ng.3570

18. Quinlan AR. BEDTools: the swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics.* (2014) 47:1–34. doi: 10.1002/0471250953.bi1112s47

19. Chew E, Tan J, Bahta A, Ho B, Liu X, Lim T, et al. Differential expression between human dermal papilla cells from balding and non-balding scalps reveals new candidate genes for androgenetic alopecia. *J Invest Dermatol.* (2016) 136:1559–67. doi: 10.1016/j.jid.2016.03.032

20. Heinz S, Benner C, Spann N, Bertolino E, Lin Y, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* (2010) 38:576–89. doi: 10.1016/j.molcel.2010.05.004

21. Szklarczyk D, Morris J, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* (2017) 45:D362–8. doi: 10.1093/nar/gkw937

22. 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

23. Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Res.* (2014) 3:146. doi: 10.12688/f1000research.4431.2

24. Wu G, Haw R. Functional interaction network construction and analysis for disease discovery. *Methods Mol Biol.* (2017) 1558:235–53. doi: 10.1007/978-1-4939-6783-4_11

25. Croft D, Mundo A, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res.* (2014) 42:D472–7. doi: 10.1093/nar/gkt1102

26. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. *Nucleic Acids Res.* (2016) 44:D481–7. doi: 10.1093/nar/gkv1351

27. Assenov Y, Ramírez F, Schelhorn S, Lengauer T, Albrecht M. Computing topological parameters of biological networks. *Bioinformatics.* (2008) 24:282–4. doi: 10.1093/bioinformatics/btm554

28. Chin C, Chen S, Wu H, Ho C, Ko M, Lin C. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol.* (2014) 8 (Suppl 4):S11. doi: 10.1186/1752-0509-8-S4-S11

29. Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics.* (2003) 4:2. doi: 10.1186/1471-2105-4-2

30. 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 (Camb).* (2021) 2:100141. doi: 10.1016/j.xinn.2021.100141

31. Zhang Y, Yu J, Shi C, Huang Y, Wang Y, Yang T, et al. Lef1 contributes to the differentiation of bulge stem cells by nuclear translocation and cross-talk with the Notch signaling pathway. *Int J Med Sci.* (2013) 10:738–46. doi: 10.7150/ijms.5693

32. Kömüves L, Ma X, Stelnicki E, Rozenfeld S, Oda Y, Largman C. HOXB13 homeodomain protein is cytoplasmic throughout fetal skin development. *Dev Dyn.* (2003) 227:192–202. doi: 10.1002/dvdy.10290

33. Jefferies C. Regulating IRFs in IFN driven disease. *Front Immunol.* (2019) 10:325. doi: 10.3389/fimmu.2019.00325

34. Pruitt K, Tatusova T, Klimke W, Maglott D. NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.* (2009) 37:D32–6. doi: 10.1093/nar/gkn721

35. Nusse R, Clevers H. Wnt/β-Catenin signaling, disease, and emerging therapeutic modalities. *Cell.* (2017) 169:985–99. doi: 10.1016/j.cell.2017.05.016

36. Ulmert I, Henriques-Oliveira L, Pereira C, Lahl K. Mononuclear phagocyte regulation by the transcription factor Blimp-1 in health and disease. *Immunology.* (2020) 161:303–13. doi: 10.1111/imm.13249

37. Gupta P, Gurudutta G, Saluja D, Tripathi R. PU.1 and partners: regulation of haematopoietic stem cell fate in normal and malignant haematopoiesis. *J Cell Mol Med.* (2009) 13:4349–63. doi: 10.1111/j.1582-4934.2009.00757.x

38. Soundararajan M, Willard F, Kimple A, Turnbull A, Ball L, Schoch G, et al. Structural diversity in the RGS domain and its interaction with heterotrimeric G protein alpha-subunits. *Proc Natl Acad Sci U.S.A.* (2008) 105:6457–62. doi: 10.1073/pnas.0801508105

39. Miranda M, Avila I, Esparza J, Shwartz Y, Hsu Y, Berdeaux R, et al. Defining a role for G-protein coupled receptor/cAMP/CRE-binding protein signaling in hair follicle stem cell activation. *J Invest Dermatol.* (2022) 142:53–64.e3. doi: 10.1016/j.jid.2021.05.031

40. Pedro, M, Lund K, Iglesias-Bartolome R. The landscape of GPCR signaling in the regulation of epidermal stem cell fate and skin homeostasis. *Stem Cells.* (2020) 38:1520–31. doi: 10.1002/stem.3273

41. Pasternack S, von Kügelgen I, Al Aboud K, Lee Y, Rüschendorf F, Voss K, et al. G protein-coupled receptor P2Y5 and its ligand LPA are involved in maintenance of human hair growth. *Nat Genet.* (2008) 40:329–34. doi: 10.1038/ng.84

42. O'Leary N, Wright M, Brister J, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* (2016) 44:D733–45. doi: 10.1093/nar/gkv1189

43. Mak K, Chan S. Epidermal growth factor as a biologic switch in hair growth cycle. *J Biol Chem.* (2003) 278:26120–6. doi: 10.1074/jbc.M212082200

44. Zhang H, Nan W, Wang S, Zhang T, Si H, Yang F, et al. Epidermal growth factor promotes proliferation and migration of follicular outer root sheath cells via Wnt/β-Catenin Signaling. *Cell Physiol Biochem.* (2016) 39:360–70. doi: 10.1159/000445630

45. Ramírez-Marín H, Tosti A. Evaluating the therapeutic potential of ritlecitinib for the treatment of alopecia areata. *Drug Des Devel Ther.* (2022) 16:363–74. doi: 10.2147/DDDT.S334727

46. Pruitt K, Tatusova T, Brown G, Maglott D. NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* (2012) 40:D130–5. doi: 10.1093/nar/gkr1079

47. Nesterovitch A, Gyorfy Z, Hoffman M, Moore E, Elbuluk N, Tryniszewska B, et al. Alteration in the gene encoding protein tyrosine phosphatase nonreceptor type 6 (PTPN6/SHP1) may contribute to neutrophilic dermatoses. *Am J Pathol.* (2011) 178:1434–41. doi: 10.1016/j.ajpath.2010.12.035

48. Miyauchi K, Ki S, Ukai M, Suzuki Y, Inoue K, Suda W, et al. Essential role of STAT3 signaling in hair follicle homeostasis. *Front Immunol.* (2021) 12:663177. doi: 10.3389/fimmu.2021.663177

49. Kitamura A, Maekawa Y, Uehara H, Izumi K, Kawachi I, Nishizawa M, et al. A mutation in the immunoproteasome subunit PSMB8 causes autoinflammation and lipodystrophy in humans. *J Clin Invest.* (2011) 121:4150–60. doi: 10.1172/JCI58414

50. Redler S, Tazi-Ahnini R, Drichel D, Birch M, Brockschmidt F, Dobson K, et al. Selected variants of the steroid-5-alpha-reductase isoforms SRD5A1 and SRD5A2 and the sex steroid hormone receptors ESR1, ESR2 and PGR: no association with female pattern hair loss identified. *Exp Dermatol.* (2012) 21:390–3. doi: 10.1111/j.1600-0625.2012.01469.x

51. Choi N, Kim W, Oh S, Sung J. HB-EGF improves the hair regenerative potential of adipose-derived stem cells via ROS generation and Hck phosphorylation. *Int J Mol Sci.* (2019) 21:122. doi: 10.3390/ijms21010122

52. Lattanand A, Johnson W. Male pattern alopecia a histopathologic and histochemical study. *J Cutan Pathol.* (1975) 2:58–70. doi: 10.1111/j.1600-0560.1975.tb00209.x

53. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*. (2009) 25:1091–3. doi: 10.1093/bioinformatics/btp101 [Epub ahead of print].

54. Udhaya Kumar S, Thirumal Kumar D, Siva R, George Priya Doss C, Younes S, Younes N, et al. Dysregulation of signaling pathways due to differentially expressed genes from the B-cell transcriptomes of systemic lupus erythematosus patients – A bioinformatics approach. *Front Bioeng Biotechnol.* (2020) 8:276. doi: 10.3389/fbioe.2020.00276

55. Abell E. Pathology of male pattern alopecia. *Arch Dermatol.* (1984) 120:1607–8.

56. Young J, Conte E, Leavitt M, Nafz M, Schroeter A. Cutaneous immunopathology of androgenetic alopecia. *J Am Osteopath Assoc.* (1991) 91:765–71.

57. Kligman A. The comparative histopathology of male-pattern baldness and senescent baldness. *Clin Dermatol.* (1988) 6:108–18. doi: 10.1016/0738-081x(88)90074-0

58. Jaworsky C, Kligman A, Murphy G. Characterization of inflammatory infiltrates in male pattern alopecia: implications for pathogenesis. *Br J Dermatol.* (1992) 127:239–46. doi: 10.1111/j.1365-2133.1992.tb00121.x

59. Mahé Y, Michelet J, Billoni N, Jarrousse F, Buan B, Commo S, et al. Androgenetic alopecia and microinflammation. *Int J Dermatol.* (2000) 39:576–84. doi: 10.1046/j.1365-4362.2000.00612.x

60. Deloche C, de Lacharrière O, Misciali C, Piraccini B, Vincenzi C, Bastien P, et al. Histological features of peripilar signs associated with androgenetic alopecia. *Arch Dermatol Res.* (2004) 295:422–8. doi: 10.1007/s00403-003-0447-y

61. Jain N, Doshi B, Khopkar U. Trichoscopy in alopecias: diagnosis simplified. *Int J Trichol.* (2013) 5:170–8. doi: 10.4103/0974-7753.130385

62. Zhang X, Zhou D, Ma T, Liu Q. Vascular endothelial growth factor protects CD200-rich and CD34-positive hair follicle stem cells against androgen-induced apoptosis through the phosphoinositide 3-Kinase/Akt pathway in patients with androgenic alopecia. *Dermatol Surg.* (2020) 46:358–68. doi: 10.1097/DSS.0000000000002091

63. Teng Y, Fan Y, Ma J, Lu W, Liu N, Chen Y, et al. The PI3K/Akt pathway: emerging roles in skin homeostasis and a group of non-malignant skin disorders. *Cells.* (2021) 10:1219. doi: 10.3390/cells10051219

64. Sudol M. From Src homology domains to other signaling modules: proposal of the 'protein recognition code'. *Oncogene.* (1998) 17:1469–74. doi: 10.1038/sj.onc.1202182

65. Filippakopoulos P, Müller S, Knapp S. SH2 domains: modulators of nonreceptor tyrosine kinase activity. *Curr Opin Struct Biol.* (2009) 19:643–9. doi: 10.1016/j.sbi.2009.10.001

66. Ku K, Choi N, Oh S, Kim W, Suh W, Sung J. Src inhibition induces melanogenesis in human G361 cells. *Mol Med Rep.* (2019) 19:3061–70. doi: 10.3892/mmr.2019.9958

67. Kim J, Kim S, Choi Y, Shin J, Kim C, Kang N, et al. Quercitrin stimulates hair growth with enhanced expression of growth factors via activation of MAPK/CREB signaling pathway. *Molecules.* (2020) 25:4004. doi: 10.3390/molecules25174004

68. Rishikaysh P, Dev K, Diaz D, Qureshi W, Filip S, Mokry J. Signaling involved in hair follicle morphogenesis and development. *Int J Mol Sci.* (2014) 15:1647–70. doi: 10.3390/ijms15011647

69. Minematsu H, Shin M, Celil Aydemir A, Kim K, Nizami S, Chung G, et al. Nuclear presence of nuclear factor of activated T cells (NFAT) c3 and c4 is required for Toll-like receptor-activated innate inflammatory response of monocytes/macrophages. *Cell Signal.* (2011) 23:1785–93. doi: 10.1016/j.cellsig.2011.06.013

70. Otsuka S, Melis N, Gaida M, Dutta D, Weigert R, Ashwell J. Calcineurin inhibitors suppress acute graft-versus-host disease via NFAT-independent inhibition of T cell receptor signaling. *J Clin Invest.* (2021) 131:e147683. doi: 10.1172/JCI147683

71. Pan M, Xiong Y, Chen F. NFAT gene family in inflammation and cancer. *Curr Mol Med.* (2013) 13:543–54. doi: 10.2174/1566524011313040007

72. De A. Wnt/Ca2+ signaling pathway: a brief overview. *Acta Biochim Biophys Sin (Shanghai).* (2011) 43:745–56. doi: 10.1093/abbs/gmr079

73. Li Q, Sun X, Wu J, Lin Z, Luo Y. PKD2 interacts with Lck and regulates NFAT activity in T cells. *BMB Rep.* (2009) 42:35–40. doi: 10.5483/bmbrep.2009.42.1.035

74. Carter N, Pomerantz J. Calcineurin inhibitors target Lck activation in graft-versus-host disease. *J Clin Invest.* (2021) 131:e149934. doi: 10.1172/JCI149934

75. Baer A, Colon-Moran W, Xiang J, Stapleton J, Bhattarai N. Src-family kinases negatively regulate NFAT signaling in resting human T cells. *PLoS One.* (2017) 12:e0187123. doi: 10.1371/journal.pone.0187123

76. Anguita E, Villalobo A. Src-family tyrosine kinases and the Ca2+ signal. *Biochim Biophys Acta Mol Cell Res.* (2017) 1864:915–32. doi: 10.1016/j.bbamcr.2016.10.022

77. Min J, Park H, Lee Y, Kim J, Kim J, Park J. Cross-talk between Wnt signaling and src tyrosine kinase. *Biomedicines.* (2022) 10:1112. doi: 10.3390/biomedicines10051112

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 15 June 2023
DOI 10.3389/fmed.2023.1107128

OPEN ACCESS

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

REVIEWED BY
Jin Wu,
University at Buffalo, United States
Ronald Lubet,
National Cancer Institute (NIH), United States

*CORRESPONDENCE
Sivakumar Arumugam
✉ siva_kumar.a@vit.ac.in

RECEIVED 24 November 2022
ACCEPTED 11 April 2023
PUBLISHED 15 June 2023

CITATION
Ramalingam PS, Priyadharshini A, Emerson IA and Arumugam S (2023) Potential biomarkers uncovered by bioinformatics analysis in sotorasib resistant-pancreatic ductal adenocarcinoma. *Front. Med.* 10:1107128. doi: 10.3389/fmed.2023.1107128

Potential biomarkers uncovered by bioinformatics analysis in sotorasib resistant-pancreatic ductal adenocarcinoma

Prasanna Srinivasan Ramalingam[1], Annadurai Priyadharshini[2], Isaac Arnold Emerson[2] and Sivakumar Arumugam[1]*

[1]Protein Engineering Lab, School of Biosciences and Technology, Vellore Institute of Technology, Vellore, India, [2]Bioinformatics Programming Laboratory, Department of Biotechnology, School of Biosciences and Technology, Vellore Institute of Technology, Vellore, Tamil Nadu, India

Background: Mutant KRAS-induced tumorigenesis is prevalent in lung, colon, and pancreatic ductal adenocarcinomas. For the past 3 decades, KRAS mutants seem undruggable due to their high-affinity GTP-binding pocket and smooth surface. Structure-based drug design helped in the design and development of first-in-class KRAS G12C inhibitor sotorasib (AMG 510) which was then approved by the FDA. Recent reports state that AMG 510 is becoming resistant in non-small-cell lung cancer (NSCLC), pancreatic ductal adenocarcinoma (PDAC), and lung adenocarcinoma patients, and the crucial drivers involved in this resistance mechanism are unknown.

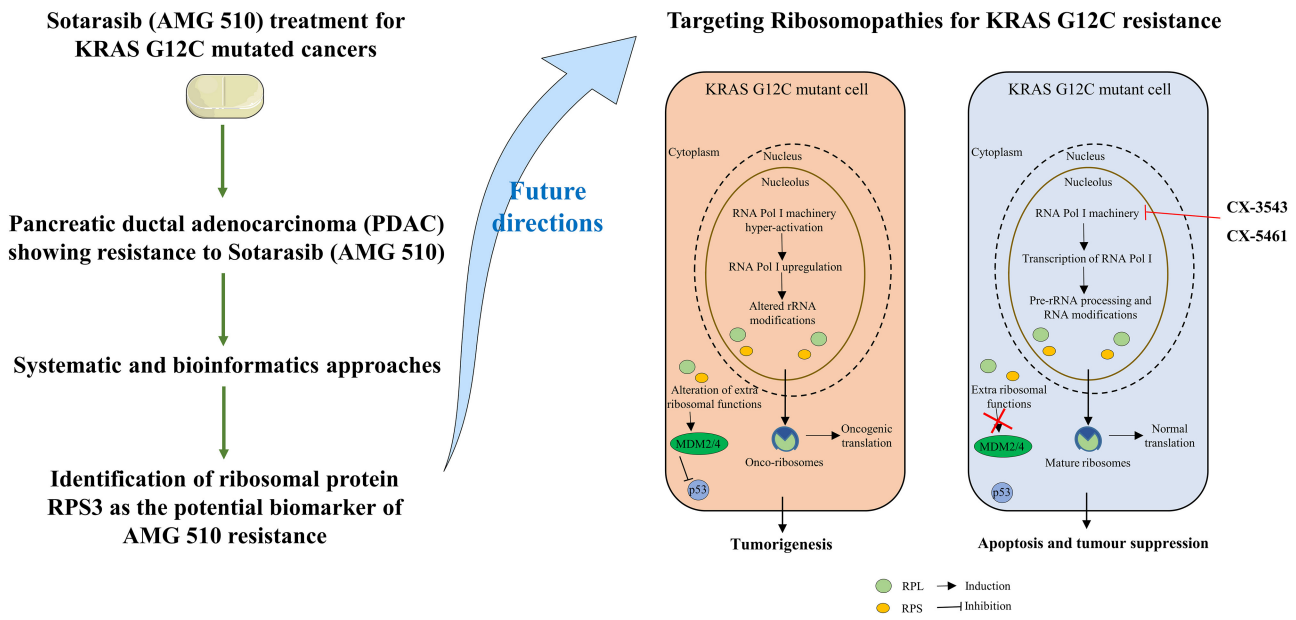
Methods: In recent years, RNA-sequencing (RNA-seq) data analysis has become a functional tool for profiling gene expression. The present study was designed to find the crucial biomarkers involved in the sotorasib (AMG 510) resistance in KRAS G12C-mutant MIA-PaCa2 cell pancreatic ductal adenocarcinoma cells. Initially, the GSE dataset was retrieved from NCBI GEO, pre-processed, and then subjected to differentially expressed gene (DEG) analysis using the limma package. Then the identified DEGs were subjected to protein–protein interaction (PPI) using the STRING database, followed by cluster analysis and hub gene analysis, which resulted in the identification of probable markers.

Results: Furthermore, the enrichment and survival analysis revealed that the small unit ribosomal protein (RP) RPS3 is the crucial biomarker of the AMG 510 resistance in KRAS G12C-mutant MIA-PaCa2 cell pancreatic ductal adenocarcinoma cells.

Conclusion: Finally, we conclude that RPS3 is a crucial biomarker in sotorasib resistance which evades apoptosis by MDM2/4 interaction. We also suggest that the combinatorial treatment of sotorasib and RNA polymerase I machinery inhibitors could be a possible strategy to overcome resistance and should be studied in *in vitro* and *in vivo* settings in near future.

KEYWORDS

sotorasib, KRAS G12C inhibitor, resistance, pancreatic ductal adenocarcinoma, ribosomal proteins, precision medicine



GRAPHICAL ABSTRACT
Study outcomes and proposed future directions to combat sotorasib resistance in KRAS G12C mutant cells.

1. Introduction

Mutant RAS-harboring cancers are predominant in many cancers including pancreatic, breast, colon, and lung, which corresponds to nearly 30% of all cancers (1, 2). Unlike NRAS and HRAS isoforms of RAS, the KRAS isoform has high mutation frequencies at mutational hotspots G12 (89%), G13 (9%), and Q61 (1%) residues (3–5). Overall, the G12th residue is the most mutated position of KRAS with G12D as the most prevalent mutation with 36%, followed by the G12V and G12C mutations with 23 and 14%, respectively (6). KRAS is a small GTPase that acts as a molecular switch by GTP-bound (active form) and GDP-bound (inactive form) states and triggers the downstream signal transduction pathways (7, 8). The GDP to GTP conversion is mediated by the guanine nucleotide exchange factors (GEFs), and the GTP to GDP hydrolysis is mediated by GTPase-activating proteins (GAPs) (9, 10). The mutant KRAS maintains the GTP-bound active state and overcomes the GTPase activity and initiates nearly 80 different downstream effector signaling pathways including MAPK and PI3K-mTOR signaling which further activates JUN and MYC transcription factors and promotes the cancer cell survival and proliferation (11–15).

Several strategies have been carried out to inhibit the mutant KRAS signaling such as targeting the upstream effectors (EGFR inhibitors, FGFR1 inhibitors, and IGF1R inhibitors); targeting the inhibitors of KRAS regulators (SOS1 inhibitors and SHP2 inhibitors); direct targeting of KRAS (KRAS on state and off-state inhibitors); downstream effector inhibitors (PI3K inhibitors, mTOR inhibitors, and MEK inhibitors); and cell cycle arrest (CDK4/6 inhibitors) (16–19). Moreover, targeting the other mediators and effectors in the MAPK pathway result in the signaling crosstalk such as MEK-PI3K, RAF-AKT, RAS-SKF, RAS-YAP, and SHP2-dependent MAPK reactivation and SHP2-independent PI3K reactivation (20–22). All the strategies have shown significant outcomes, but the complete inhibition of KRAS was promising in the direct targeting strategy. In general, the intracellular levels of GTP are in micromolar (μM) ranges, and its binds with picomolar (pM) affinity to the GTP-binding pocket of the KRAS, which challenges it as undruggable to the medicinal chemistry and drug discovery researchers to design and develop a potent KRAS mutant small molecule inhibitors (23–25). Finally, the undruggable became druggable by the successful discovery and FDA approval of KRAS G12C inhibitor sotorasib (AMG 510) for the treatment of non-small-cell lung cancer (NSCLC) and other solid tumors (26–28). The sotorasib specifically targets the cryptic pocket of the KRAS G12C (H95/Y96/Q99) and forms the covalent bond with the reactive cysteine at the 12th position, which also limits its ability to target other KRAS mutants such as G12D and G12V that lacks reactive cysteine (29). Recently, in December 2022, FDA granted the accelerated approval for adagrasib (MRTX849) for the treatment of KRAS G12C-mutated NSCLC (30).

Accumulating pieces of evidence report that sotorasib is becoming resistant among NSCLC, pancreatic ductal adenocarcinoma, and colorectal adenocarcinoma patients bearing KRAS G12C mutation and even resulting in hepatotoxicity (31, 32). The understanding of this resistance mechanism is challenging due to the intracellular heterogeneity and variability of KRAS G12C-mutated cancer cells (33). Hence, to identify the crucial biomarkers involved in the sotorasib resistance, we have retrieved the RNA-seq data from the NCBI GEO database of AMG 510 treated (resistant) and untreated in KRAS G12C-mutant MIA-PaCa2 pancreatic ductal adenocarcinoma cells. The differentially expressed genes (DEGs) were identified by the linear model, and then, the DEGs were subjected to protein–

protein interaction (PPI), cluster analysis, and hub gene analysis. In addition to this, the resulting probable biomarkers were also subjected to gene ontology (GO), pathway enrichment, and survival analyses to find the crucial biomarker in the sotorasib resistance.

2. Materials and methods

2.1. Data collection and pre-processing

The RNA-seq dataset retrieved for this study was accessed through NCBI Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/). The keywords used for filtering the dataset include "KRAS mutated Pancreatic cancer" and "*Homo sapiens*" (organism). The datasets were screened, and "GSE178479" was retrieved for this study in which the sotorasib (AMG 510) resistance in the KRAS G12C-mutant MIA-PaCa2 pancreatic ductal adenocarcinoma cells was reported (34). The sequencing platform and the platform ID of the sample were "Illumina HiSeq 4000" and "GPL20301," respectively. The number of samples used in this study was two, which includes RNA-seq profiles of AMG 510 treated (rep1 and rep2) and AMG 510 untreated (rep1 and rep2) MIA-PaCa2 cells. The present study was carried out to predict the crucial biomarkers involved in the AMG 510 resistance in pancreatic ductal adenocarcinoma cells.

The count matrix of the samples was prepared based on the matrix file information provided in the GEO database (35). The lowly expressed genes were filtered based on their counts using the counts per million (CPM) function in the *edgeR* package with the threshold of 0.5. Box plots were used to check the distribution of the read counts on the log2 scale (36). The CPM function provided the log2 counts per million which are then corrected for different library sizes. The CPM function also adds a small offset to avoid taking a log of zero. The trimmed mean of M-value (TMM) normalization was performed to eliminate composition biases between the libraries (37). This generates a set of normalization factors, where the product of these factors and the library sizes define the effective library size. The *calcNormFactors* function calculated the normalization factors between libraries.

2.2. Differential gene expression analysis

The *limma* package (38, 39) with the voom function was used, which transforms the read counts into logCPMs while taking account of the mean–variance relationship in the given data (40, 41). After vooming, we applied a linear model to the voom transformed data to test for differentially expressed genes (DEGs) using standard *limma* commands.

The voom transformed data have been used in *limma* to test for differential gene expression. The linear model fit was designed for each gene using the lmFit function in *limma* which estimates the groups and gene-wise variances. The contrast between the groups was then analyzed based on the makeContrasts function. Then the contrasts matrix was fitted to the object to get the statistics and estimated parameters. Here, we called the contrasts.fit function in *limma*. Furthermore, we called the eBayes function to perform the empirical Bayes shrinkage on the variances and estimated the logFC of 0.05 and their associated *p*-values. Finally, to increase the significance and reduce the false discovery rates, we used the TREAT function to predict specific genes (42–44).

2.3. Network analysis

The differentially expressed genes (DEGs) filtered through the TREAT function were then subjected to the STRING database (https://string-db.org/) to predict the protein–protein interactions (PPIs) with a confidence level of 0.004 and higher, and the first shell of 10 interactions was used as a filter (45). The MCODE and CytoHubba were used to analyze the probable marker genes among the DEGs (46).

2.4. Enrichment and survival analysis

The hub genes resulting from the network analysis were then subjected to gene ontology using the enrichGO function in the clusterProfiler package (47). The enriched biological process (BP), cellular components (CC), and molecular functions (MF) were analyzed using the enrichGO function. The KEGG pathway analysis was also carried out using the enrichKEGG function to analyze the enriched terms.

The Kaplan–Meier (KM) survival analysis was carried out based on the Spearman correlation using the Kaplan–Meier plotter online tool employing the median patient splitting mode (48, 49). Hazard is the defined slope for the survival curve which measures the incidence of death, and the hazard ratio (HR) compares the two treatment groups. If HR is 2.0, then the rate of death in one treatment group is twice the other group (50). A statistical hypothesis test was calculated based on a log-rank test. The schematic representation of the workflow of the study is shown in Figure 1.

3. Results

3.1. Identification of differently expressed genes

Through *limma* analysis, we have tested the difference between the sotorasib (AMG 510) treated and untreated samples to analyze the genes responsible for the AMG 510 resistance in the treated group. The voom transformation of adjusting the library size with the normalization factors was analyzed through a mean–variance trend. The comparative boxplot analysis of unnormalized logCPM with the voom transformed logCPM is shown in Figure 2 which represents the precision of normalization. The CPM plot of count data after filtering the lowly expressed genes is provided in Supplementary Figure 1. The mean–variance relationship helps to analyze whether the low counts are filtered adequately and variation in the data by estimating the relationship of the log counts, which generates a precision weight for each observation and enters these into the limma empirical Bayes analysis. The voom mean–variance trend curve is shown in Supplementary Figure 2.

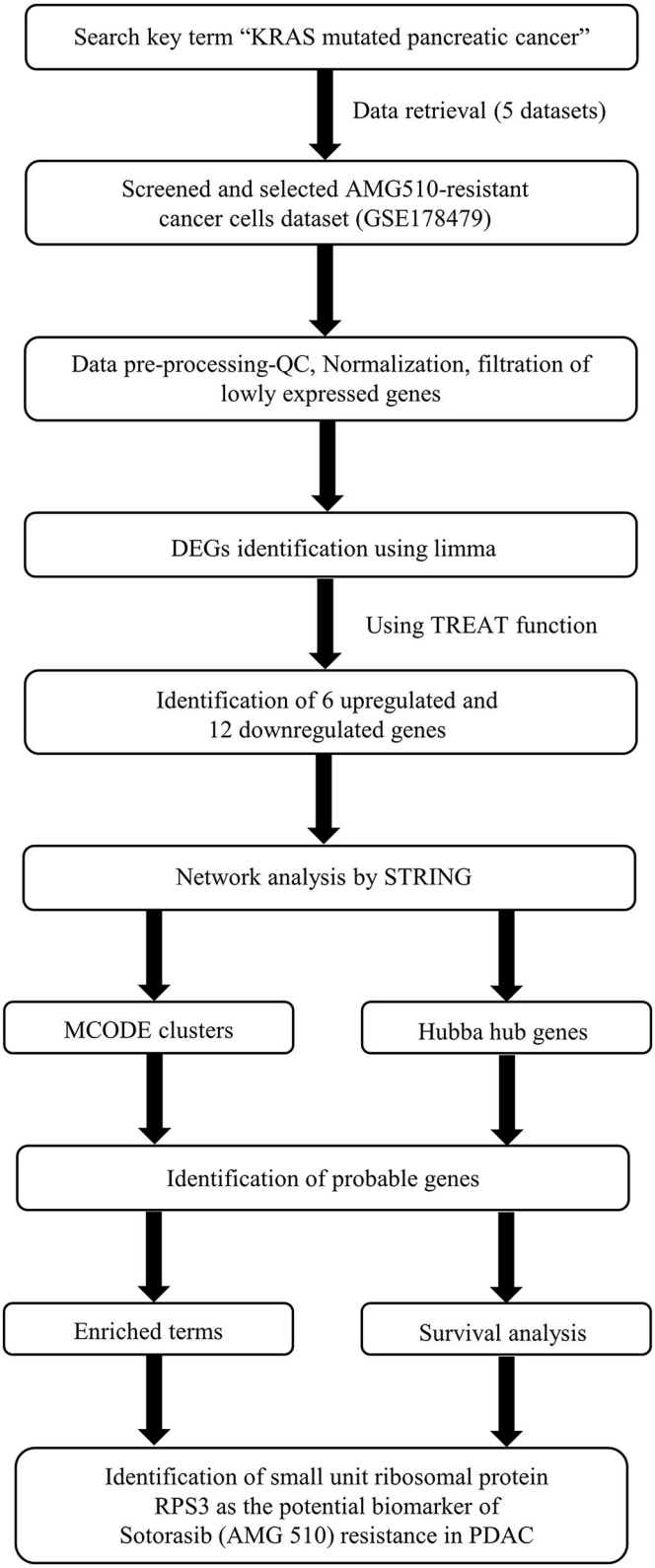


FIGURE 1
Schematic representation of the workflow of the study.

The empirical Bayes function was used to analyze the DEGs with the linear model fit. The linear model fit resulted in the identification of upregulated and downregulated genes from the DEGs. In this study, it resulted in the differentially expressed genes among the AMG 510 treated (resistant) and untreated groups, which are repressed through the MA plot as shown in Figure 3 and

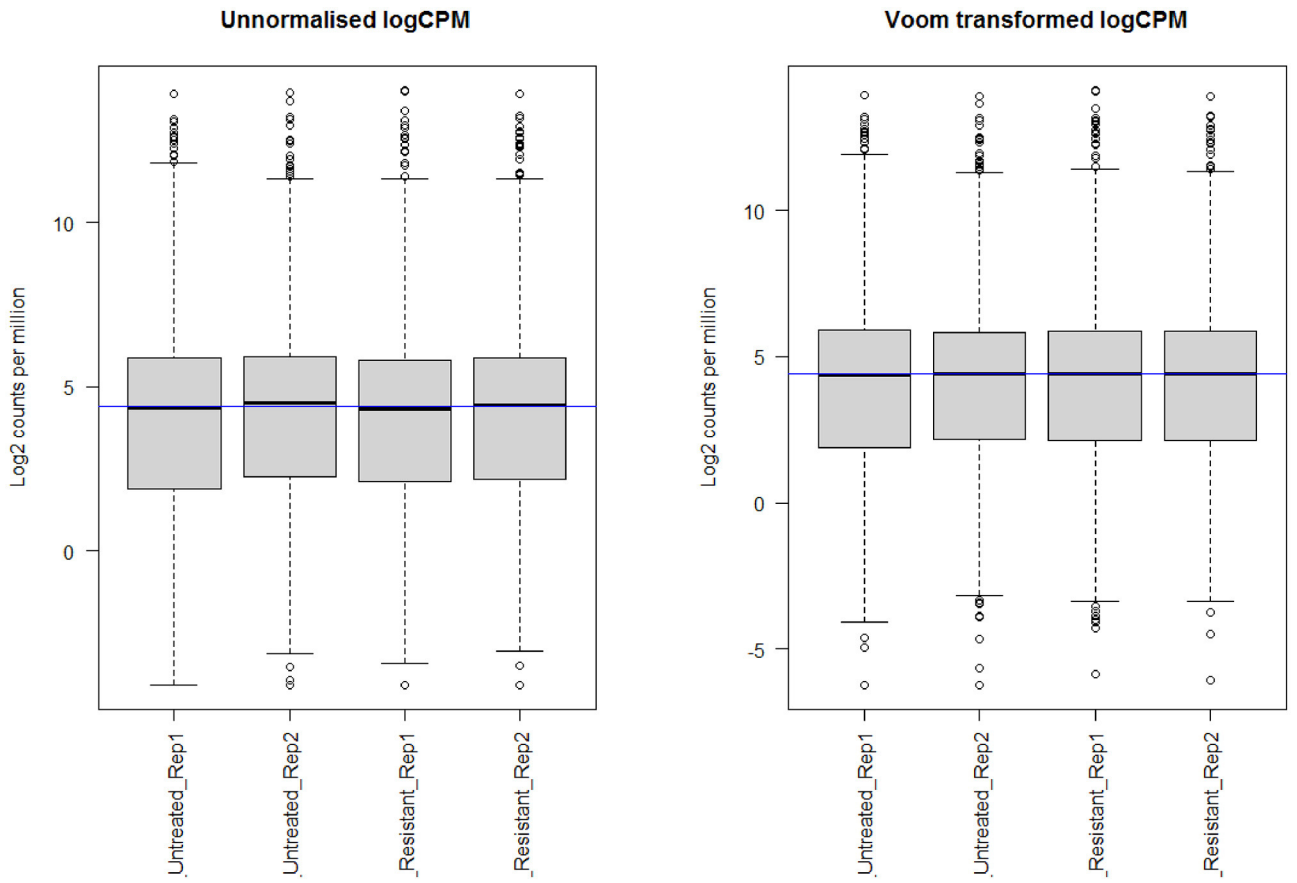


FIGURE 2
Boxplot analysis of unnormalized logCPM with the voom transformed logCPM.

the volcano plot as shown in Figure 4. Initially, the raw RNA-seq data were retrieved, pre-processed, and the differentially expressed genes (DEGs) were predicted using a cutoff on the log fold change threshold of 0.5. The *p*-value threshold of 0.05 resulted in the identification of 330 upregulated genes and 499 downregulatory genes as shown in Figure 4, and the complete list of DEGs is provided in Supplementary Table 1. To reduce false discovery rates, we further applied TREAT (*t*-tests relative to a threshold) function in the limma package, which resulted in the identification of six upregulated DEGs and 12 downregulated DEGs.

3.2. Network analysis

The interaction network was visualized using Cytoscape using molecular complex detection (MCODE) to find the significant clusters between each node representing a gene while edges represent the interaction of the molecules. The default parameters were set including the degree cutoff of 2, node score cutoff of $\geq$0.2, K-score of $\geq$2, and max depth from seed of 100. Finally, the MCODE resulted in six clusters with the highest nodal score of 22 as shown in Figure 5.

The probable marker genes have been identified based on the highly connected nodes using CytoHubba in Cytoscape. It uses 12 scoring methods to identify the markers, namely, betweenness, bottleneck, closeness, clustering coefficient (CC), degree, the density of maximum neighborhood component (DMNC), eccentricity (EcC), edge percolated component (EPC), maximal clique centrality (MCC), maximum neighborhood component (MNC), radiality, and stress. The top 10 genes from each scoring method were isolated. Genes that are common in more than five scoring methods and also have an impact on MCODE were considered hub genes.

3.3. Enrichment analysis

The enrichment analysis was performed with the GO terms: biological process (BP), cellular components (CC), and molecular functions (MF). The biological process includes cytoplasmic translation, ribosomal small subunit assembly, ribosome assembly, ribosomal small subunit biogenesis, non-membrane-bounded organelle assembly, negative regulation of protein ubiquitination, and negative regulation of protein modification by small protein conjugation or removal. Cellular components include cytosolic ribosome, ribosomal subunit, ribosome, cytosolic small ribosomal subunit, cytosolic large ribosomal subunit, small ribosomal subunit, large ribosomal

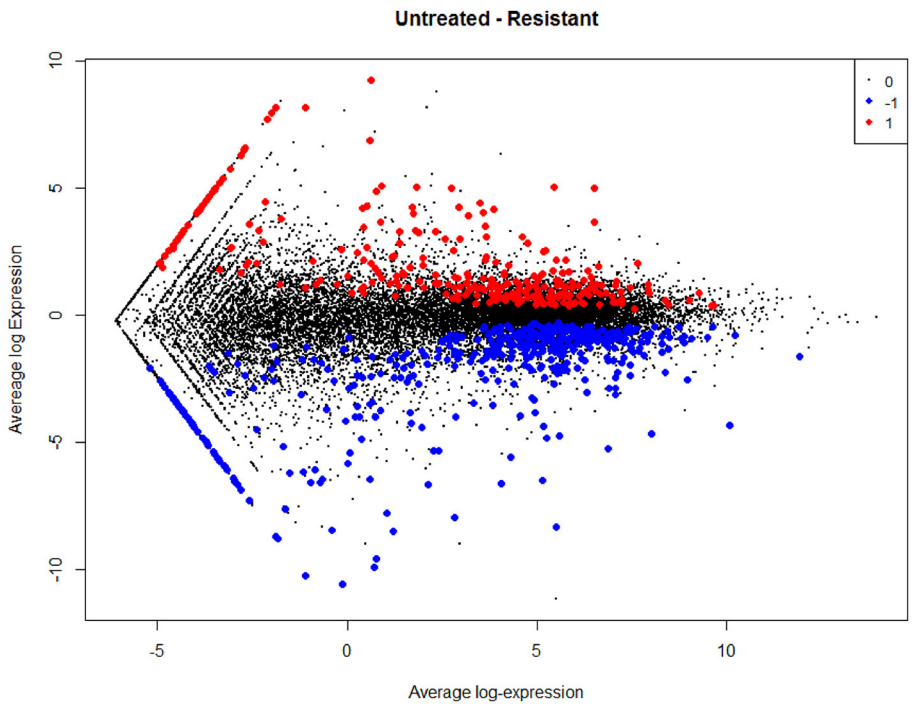


FIGURE 3
MA plot used to represent log fold change vs. mean expression between the two groups (AMG 510 treated and untreated). A scatter plot depicts the normalized mean expression on the x-axis and base-2 log fold change on the y-axis. The red dots represent the upregulated genes, the blue dots represent the downregulated genes, and the black dots represent the non-significant genes.

subunit, focal adhesion, cell–substrate junction, polysome, polysomal ribosome, rough endoplasmic reticulum, cytoplasmic side of endoplasmic reticulum membrane, rough endoplasmic reticulum membrane, and euchromatin. Molecular functions are structural constituents of the ribosome and rRNA binding. The enriched GO terms of biological process (BP), cellular components (CC), and molecular functions (MF) are shown in Figure 6 and Table 1. Then the KEGG pathway analysis was also carried out, and the enriched term was observed as "hsa03010:Ribosome."

3.4. Survival analysis

The Kaplan–Meier (KM) survival analysis plot was created based on Spearman's correlation, using the hazard ratio (HR) and log-rank test of the genes. In general, HR > 1 represents that the low-expression group has a higher chance of survival than the high-expression group, and HR < 1 represents that high-expression groups have a higher chance of survival than the low-expression group. The survival analysis of probable genes showed that the low expression of RPL4, RPL32, RPLP1, and RPS3 would have a higher probability for survival, and the high expression of RPS28, RPS15, RPS9, RPL15, and JUN would have a higher probability for survival. Based on the log-rank test, the significance level was set to 0.05, and if the calculated p-value is >0.05, the null hypothesis is retained. Based on these criteria, the ribosomal protein RPS3 was identified as a probable biomarker that showed high survival rates and $p < 0.05$ as shown in Figure 7. In addition, the HR of RPS3 is almost near two which indicates that it has twice the rate of death when compared to the others. The KM survival plots of RPL15, RPS15, RPS28, RPL4, RPL32, RPLP1, RPS9, and JUN are shown in Supplementary Figure 3.

4. Discussion

KRAS mutations are prevalent in many cancers including pancreatic, breast, colon, and lung with mutational hotspots at G12 (89%), G13 (9%), and Q61 (1%) residues (1, 2). The G12D, G12C, and G12V are frequent mutations with 36, 23, and 14% expressions, respectively (6). Of note, the KRAS G12C mutation is relatively high in lung adenocarcinoma than in pancreatic adenocarcinoma patients. The direct inhibition of the mutant KRAS is very prominent over other strategies but challenges the small molecule inhibitor development due to their high-affinity GTP-binding pocket and smooth surface (16, 51). Structure-based drug design guided the development and FDA approval of first-in-class potential KRAS G12C inhibitor sotorasib (AMG 510) that has changed the scenario in which the mutant KRAS became undruggable (26). Recently, in December 2022, FDA granted the accelerated approval for Adagrasib (MRTX849) for the treatment of KRAS G12C-mutated NSCLC (30). In addition to this, several pharma industries have initiated to design and develop novel KRAS mutant inhibitors (mutant specific/pan-KRAS). Several KRAS G12C (GDP-bound off state) inhibitors, such as sotorasib

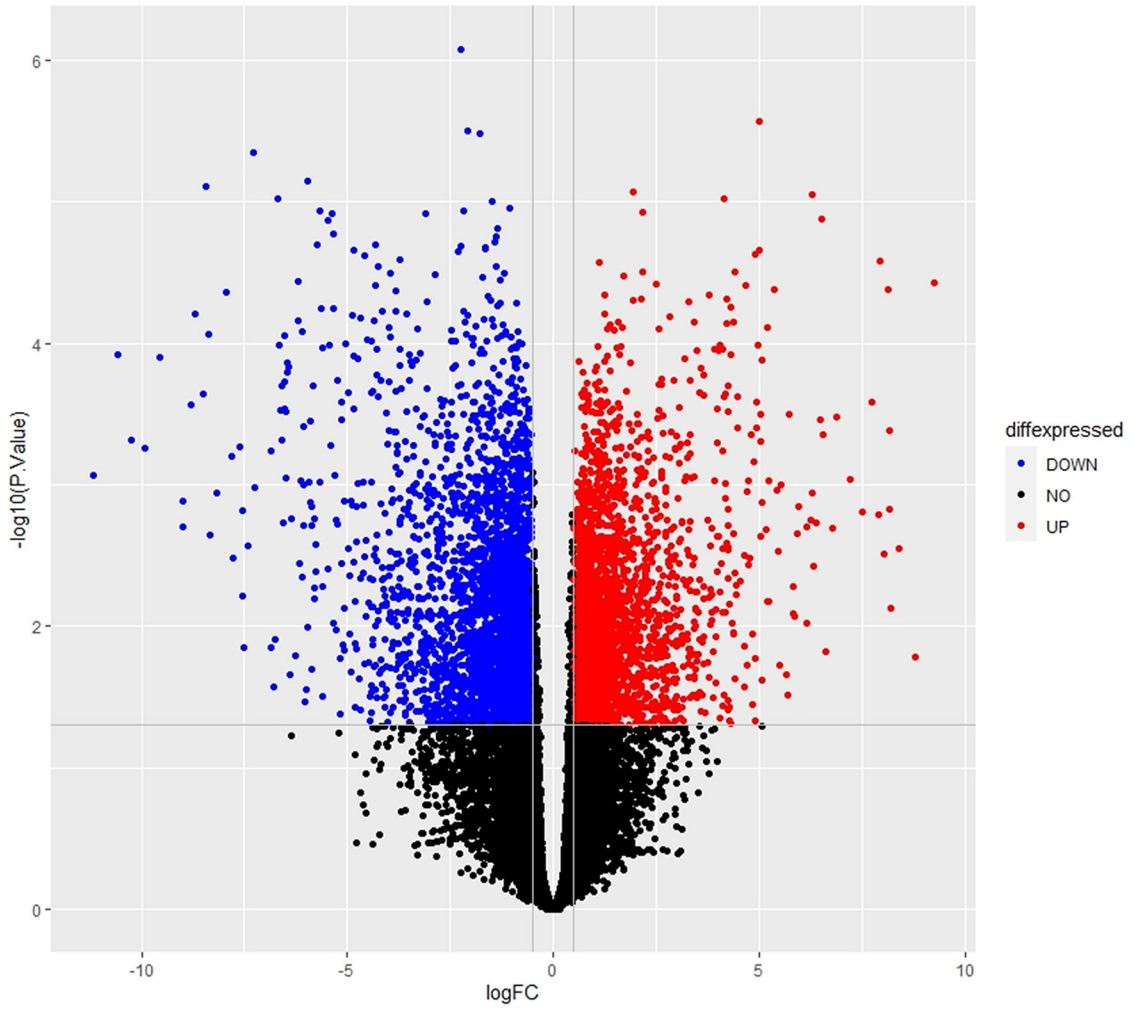


FIGURE 4
Volcano plot of the DEGs depicts the logFC on the *x*-axis and −log10 (*p*-value) on the *y*-axis. The red dots represent the upregulated genes, the blue dots represent the downregulated genes, and the black dots represent the non-significant genes.

(AMG 510), adagrasib (MRTX849), GDC-6036, JNJ-74699157, D-1553, JDQ443, LY3537982, LY3499446, ARS1620, and KRAS G12C (GDP-bound off state) inhibitors such as RMC-6291, RMC-6236, and RM-018, and Pan KRAS Switch I/II inhibitors such as BI-2852, are being studied in preclinical and clinical studies (18, 52–55). Recent pieces of evidence report the resistance to AMG 510 among KRAS G12C-mutant cancer patients (31, 33). Moreover, Adagrasib (MRTX849) and ARS1620 were reported to have acquired resistance in KRAS G12C-mutant cells (33, 56). Amplification of the mesenchymal epithelial transition factor receptor (MET); activating mutations of downstream effectors, such as BRAF, and dual specificity mitogen-activated protein kinase kinase 1 (MEK1); oncogenic fusion with fibroblast growth factor receptor 3 (FGFR3) and CCDC6-RET; and loss-of-function mutations of phosphatase and tensin homolog (PTEN) and neurofibromin 1 (NF1) were reported to be the key elements involved in the resistance mechanisms to KRAS mutant inhibitors in lung adenocarcinoma and colorectal adenocarcinoma (56, 57). Unlike the abovementioned resistance mechanisms, our results revealed a significant correlation between the sotorasib resistance in KRAS G12C-mutant cells and ribosomopathies.

Recently Chan et al. (34) reported an interesting study on the identification of sotorasib (AMG 510) resistance in the KRAS G12C-mutant MIA-PaCa2 pancreatic ductal adenocarcinoma cells when treated with increasing dosage (0.1–5 μM) for 60 days and found that MIA-PaCa2 showed resistance at 5 μM treatment of AMG 510 (34). This interested us to identify the crucial biomarkers involved in the AMG 510 resistance in the KRAS G12C-mutant MIA-PaCa2 pancreatic ductal adenocarcinoma cells. In addition to MIA-PaCa2 cells, they have also tested the AMG 510 resistance in SW1463 human Caucasian rectum adenocarcinoma, LU99 lung giant cell carcinoma, and LU65 lung carcinoma cell lines which have KRAS G12C mutations.

The main aim of the present study was to identify the key biomarker genes involved in the AMG 510 resistance. Initially, the raw RNA-seq data were retrieved, pre-processed, and the differentially expressed genes (DEGs) were predicted

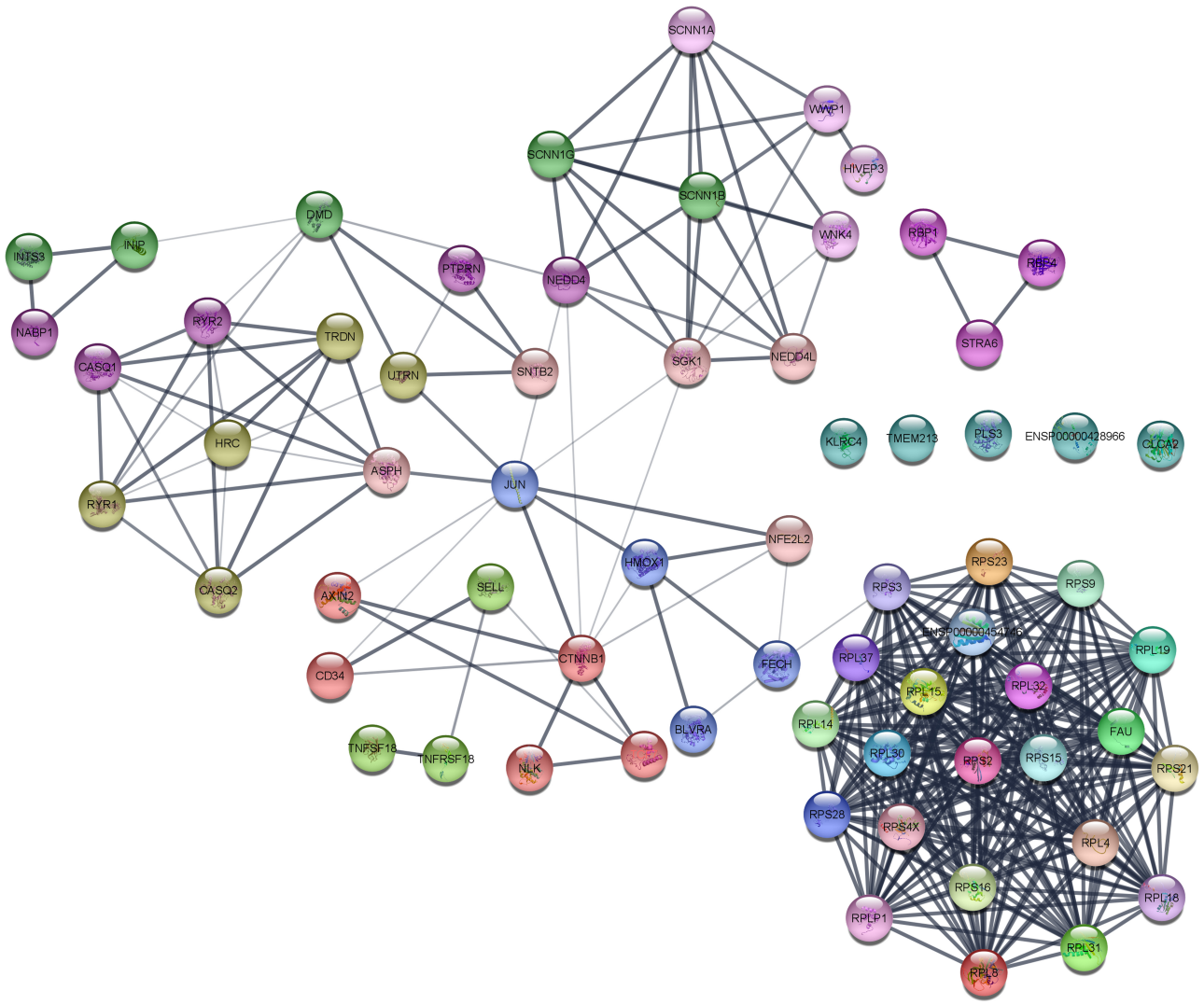


FIGURE 5
Protein–protein interaction (PPI) network of DEGs obtained from the STRING database.

which resulted in the identification of 330 upregulated genes and 499 downregulatory genes as shown in Figure 4 and Supplementary Table 1. The *t*-tests relative to a threshold (TREAT) function reduced the false discovery rates of DEGs (42), which further resulted in the identification of six upregulated and 12 downregulated genes. These filtered DEGs were studied for the protein–protein interaction network using STRING which resulted in four MCODE clusters, and the MCODE cluster 1 showed the highest nodal density among the other clusters as shown in Figure 5. In addition, cluster analysis and hub gene analysis were carried out which resulted in probable biomarkers as shown in Figure 6, and the enriched GO terms of biological process (BP), cellular components (CC), and molecular functions (MF) are shown in Table 1. In general, HR > 1 represents that the low-expression group has a high chance of survival than the high-expression group, and HR < 1 represents that the high-expression group has a high chance of survival than the low-expression group (58). Finally, the survival analysis based on the hazard ratio and log-rank test resulted in the identification of RPS3 as the probable biomarker with high survival rates and $p < 0.05$ as shown in Figure 7. Based on the log-rank test, the significance level was set to 0.05, and if the calculated p-value is >0.05, the null hypothesis is retained. Moreover, the HR of RPS3 is nearly 2 which indicates that it has twice the rate of death when compared to the others. The KM survival plots of RPL15, RPS15, RPS28, RPL4, RPL32, RPLP1, RPS9, and JUN are shown in Supplementary Figure 3. In addition, the GO of all the 330 upregulated genes and 499 downregulatory genes shown in Supplementary Table 1 reveals that the myc transcriptional targets, such as E2F transcription factor 6 (ENSG00000169016), are upregulated and the CDK10 (ENSG00000185324) is downregulated. Generally, the E2F6 regulates the gene expression of proteins involved in cell proliferation and the CDK10 acts as a tumor suppressor.

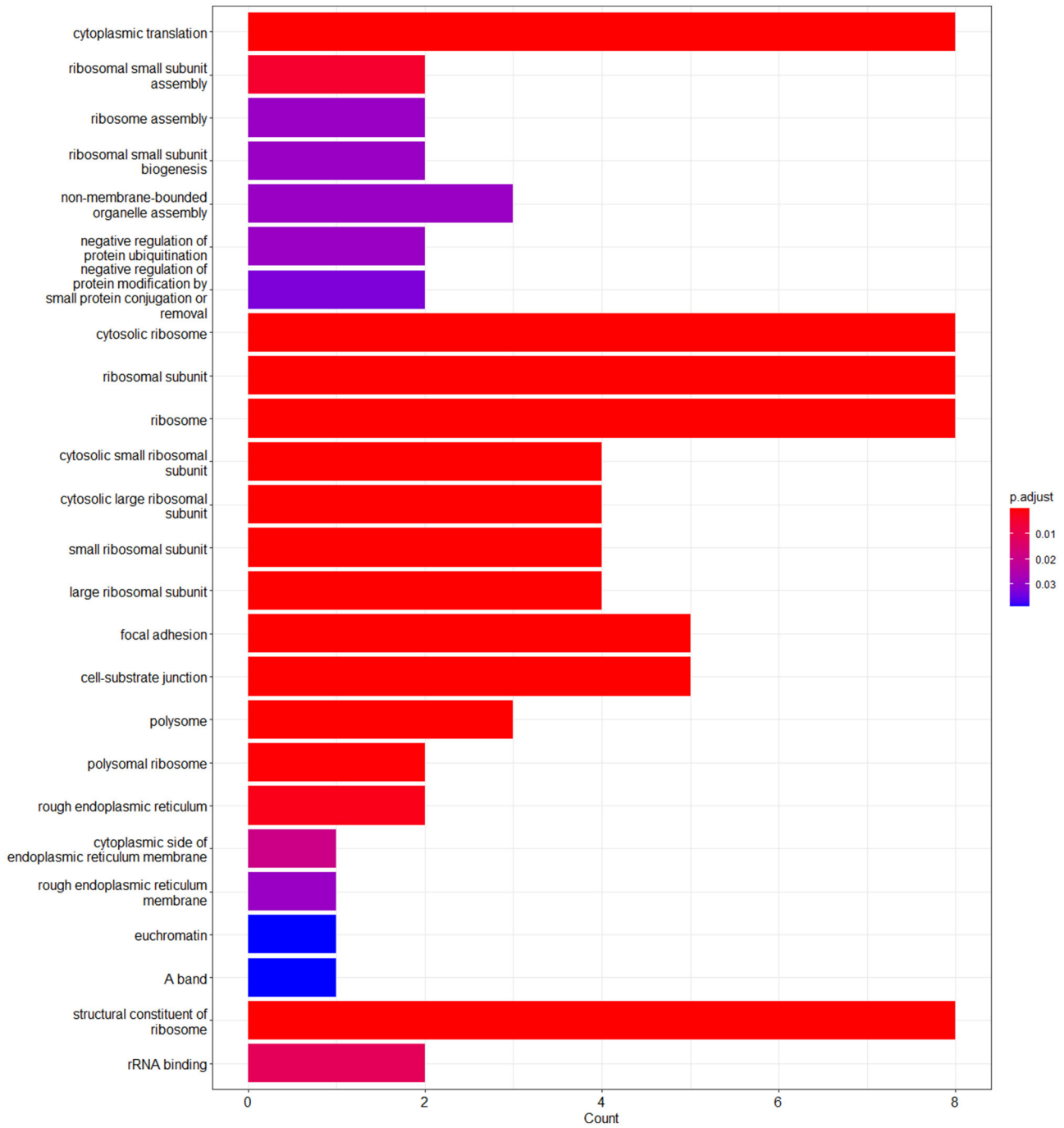


FIGURE 6
Bar plot of the enriched GO terms analyzed using enrichGO function using cluster profiler.

Furthermore, the CDC25B (ENSG00000101224) expression has a p53-dependent tumor suppressive effect, which is downregulated. The anti-apoptotic BCL-6 (ENSG00000113916) is downregulated. The abovementioned targets are also involved in the RAS signaling pathway. These data suggest that the resistance could be a result of RNA pol I machinery hyperactivation and apoptosis evasion. The present study revealed that the small unit ribosomal protein RPS3 is known to be only expressed in the AMG 510 resistant MIA-PaCa2 cells and identified as a significant biomarker involved in the resistance of AMG 510. These novel identifications resulted from the emergence and accumulation of RNA-Seq data of drug-resistant cancer cells.

Ribosome biogenesis starts from the nucleolus and ends in the cytoplasm with the formation of the mature ribosome from rRNA and ribosomal proteins (59). In normal cells, the RNA pol I initiates the Pol I transcription followed by the pre-rRNA processing and modification and then assembled

TABLE 1 Gene ontology analysis of the enriched terms.

GO term and GO ID	DEGs	*p*-value	Adjusted *p*-value	Genes
Cytoplasmic translation (GO:0002181)	BP	1.13E-16	2.91E-14	RPL4/RPLP1/RPS28/RPS9/RPL32/RPL15/RPS15/RPS3
Ribosomal small subunit assembly (GO:0000028)	BP	3.50E-05	0.004512	RPS28/RPS15
Ribosome assembly (GO:0042255)	BP	0.00037	0.029458	RPS28/RPS15
Ribosomal small subunit biogenesis (GO:0042274)	BP	0.00053	0.029458	RPS28/RPS15
Non-membrane-bounded organelle assembly (GO:0140694)	BP	0.000575	0.029458	RPS28/RPS15/RPS3
Negative regulation of protein ubiquitination (GO:0031397)	BP	0.000685	0.029458	RPS15/RPS3
Negative regulation of protein modification by small protein conjugation or removal (GO:1903321)	BP	0.000896	0.033031	RPS15/RPS3
Cytosolic ribosome (GO:0022626)	CC	3.72E-18	1.34E-16	RPL4/RPLP1/RPS28/RPS9/RPL32/RPL15/RPS15/RPS3
Ribosomal subunit (GO:0044391)	CC	3.95E-16	7.10E-15	RPL4/RPLP1/RPS28/RPS9/RPL32/RPL15/RPS15/RPS3
Ribosome (GO:0005840)	CC	4.09E-15	4.91E-14	RPL4/RPLP1/RPS28/RPS9/RPL32/RPL15/RPS15/RPS3
Cytosolic small ribosomal subunit (GO:0022627)	CC	2.30E-09	2.07E-08	RPS28/RPS9/RPS15/RPS3
Cytosolic large ribosomal subunit (GO:0022625)	CC	8.69E-09	6.26E-08	RPL4/RPLP1/RPL32/RPL15
Small ribosomal subunit (GO:0015935)	CC	1.98E-08	1.19E-07	RPS28/RPS9/RPS15/RPS3
Large ribosomal subunit (GO:0015934)	CC	1.30E-07	6.71E-07	RPL4/RPLP1/RPL32/RPL15
Focal adhesion (GO:0005925)	CC	5.12E-07	2.23E-06	RPL4/RPLP1/RPS9/RPS15/RPS3
Cell-substrate junction (GO:0030055)	CC	5.56E-07	2.23E-06	RPL4/RPLP1/RPS9/RPS15/RPS3
Polysome (GO:0005844)	CC	3.04E-06	1.10E-05	RPS28/RPL32/RPS3
Polysomal ribosome (GO:0042788)	CC	9.28E-05	0.000304	RPS28/RPL32
Rough endoplasmic reticulum (GO:0005791)	CC	0.000614	0.001842	RPL4/RPS28
Cytoplasmic side of endoplasmic reticulum membrane (GO:0098554)	CC	0.006886	0.019069	RPS28
Rough endoplasmic reticulum membrane (GO:0030867)	CC	0.011453	0.029451	RPS28
Euchromatin (GO:0000791)	CC	0.017363	0.039066	JUN
A band (GO:0031672)	CC	0.017363	0.039066	RPL15
Structural constituent of ribosome (GO:0003735)	MF	7.11E-16	3.34E-14	RPL4/RPLP1/RPS28/RPS9/RPL32/RPL15/RPS15/RPS3
rRNA binding (GO:0019843)	MF	0.000478	0.011236	RPS9/RPS3

with ribosomal proteins (RPs) to form mature 60s and 40s subunits and ultimately takes part in protein synthesis. Unlike normal cells, the RNA pol I is hyperactivated leading to the altered rRNA modifications and altered RPs extraribosomal functions, thus forming the onco-ribosomes and translating the oncogenic mRNAs and ultimately ending with ribosomopathies (59). Some large subunit ribosomal proteins, such as RPL5, RPL9, RPL10, RPL11, RPL15, RPL21, RPL22, RPL23A, RPL27, RPL31 RPL34, RPL35, RPL36, and large subunit ribosomal proteins, such as RPS7, RPS15, RPS15A, RPS17, RPS19, RPS20, RPS24, RPS27, and RPSA, are reported to have significant roles in the progression of various types of cancers including lung, colon, breast, and pancreatic cancers (60–62). Generally, the ribosomal proteins (RPs) directly/indirectly interact with the Mdm2/Mdm4 E3 ubiquitin-protein ligases, which in turn regulate the degradation of p53 tumor suppressor protein resulting in the tumor progression (62, 63). An interesting study reports that the WD repeat-containing protein 74 (WDR74) alters the RPL5 levels and promotes metastasis by degrading p53 via the RPS15-Mdm2 axis in

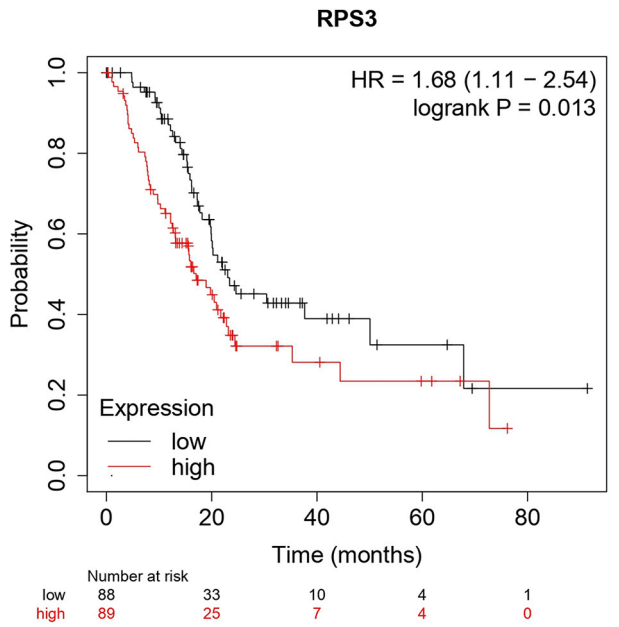


FIGURE 7
The Kaplan–Meier plot for survival analysis of key biomarkers RPS3. The *x*-axis represents the time in months, while the *y*-axis represents the probability of survival. The red and black colors represent the high expression and low expression of the biomarkers, respectively.

lung carcinoma (64). The ribosomal proteins were upregulated in KRAS mutant Panc-1 cells, and their inhibition results in cell cycle arrest, apoptosis induction, and antiproliferation (65, 66).

RPS3 knockdown in Caco-2 colon cancer cells showed decreased cancer progression and increased apoptosis via p53 upregulation and reduced activity of lactate dehydrogenase (LDH) (67). RPS3 was also reported to induce apoptosis by disrupting its interaction with E2F1 and also upregulates the expression of pro-survival genes in NSCLC (68). On this note, the mutations in the ribosomal proteins are also highly involved in tumorigenesis. The RPs were reported to interact with MDM2/4 and inhibit p53, and overexpression was observed as a result of the hyperactivation of RNA polymerase I machinery. The inhibition of RNA polymerase I machinery by inhibitors, such as CX-3543 and CX-5461, promotes p-53-dependent apoptosis in several cancers (69, 70). The clinical trials of RNA polymerase I machinery by inhibitors CX-5461 (NCT02719977) and CX-3543 (NCT00955786) resulted in the identification of safety, tolerable dosage, and effective dosage regimes and also resulted in less toxicity in patients (71). The potential of individual RNA polymerase I machinery inhibitors was studied, and combination strategies have to be studied in near future from the successful interventions from preclinical studies. Chan et al. (34) reported that the sotorasib resistance was offered by the PAK/PI3K pathway in KRAS G12C-mutant MIA-PaCa2 cells, and our bioinformatics analysis showed that RPS3 was the crucial biomarker. Recent reports show that RPS3 mediates the PI3K-Akt signaling axis in cancer cells, which correlates with our findings from the study (72, 73).

From the above understandings, we observe and conclude that the small unit ribosomal protein RPS3 is the crucial biomarker of the AMG 510 resistance in KRAS G12C-mutant MIA-PaCa2 cell pancreatic ductal adenocarcinoma cells. The study outcomes and the possible future directions to combat the Sotorasib resistance in KRAS G12C mutant cells were shown in the Graphical Abstract. Co-targeting of ribosomal proteins along with the target-specific inhibitors (here KRAS G12C-mutant inhibitor) will pave way for the development of precision treatment, such as using CRISPR-Cas and T-cell immunotherapy, in cancer.

5. Conclusion

The current study was performed to evaluate the crucial biomarkers involved in the KRAS G12C inhibitor, sotorasib (AMG 510). From the analysis, we finally conclude that the ribosomal protein RPS3 is the crucial biomarker involved in the AMG 510 resistance in the KRAS G12C-mutant MIA-PaCa2 cell pancreatic ductal adenocarcinoma cells. From the study results and previous literature, we also report that resistance could result from the degradation of p53 via the RPs-MDM2/MDM4-p53 axis. Thus, the combinatorial treatment strategy of (i) KRAS G12C-mutant inhibitors and (ii) RNA polymerase I machinery inhibitors, such as CX-3543 and CX-5461, could be a possible strategy to tackle resistance and has to be studied in *in vitro* and *in vivo* settings, which promotes the increased therapeutic treatment of KRAS G12C-mutated cancers in the era of precision medicine.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: GSE178479.

Author contributions

PSR conceptualized and designed the study. PSR and AP retrieved the data, carried out all the analyses, and wrote the manuscript. All the results were validated and the manuscript was corrected by IE and SA. All authors proofread the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The authors thank Vellore Institute of Technology, Vellore for providing 'VIT SEED Grant-RGEMS Fund (SG20220095)' for carrying out this research work.

Acknowledgments

The authors would like to thank Vellore Institute of Technology (VIT), Vellore, India, for providing the necessary facilities to carry out this study.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1107128/full#supplementary-material

SUPPLEMENTARY FIGURE 1
CPM plot of count data after filtering the poorly expressed genes.

SUPPLEMENTARY FIGURE 2
Voom mean–variance trend curve. It depicts that the lowly expressed genes are filtered properly. t. Counts nearly 0 (plot *x*-axis value −1) have low standard deviations. This rises immediately for low counts and then gradually decreases.

SUPPLEMENTARY FIGURE 3
Kaplan–Meier plot for survival analysis of RPL4 **(A)**, RPL32 **(B)**, RPLP1 **(C)**, RPS9 **(D)**, JUN **(E)**, RPL15 **(F)**, RPS15 **(G)**, and RPS28 **(H)**. The *x*-axis represents the time in months, while the *y*-axis represents the probability of survival. The red and black colors represent the high expression and low expression of the biomarkers, respectively.

References

1. Prior IA, Lewis PD, Mattos C. A comprehensive survey of Ras mutations in cancer. *Cancer Res.* (2012) 72:2457–67. doi: 10.1158/0008-5472.CAN-11-2612

2. Fernández-Medarde A, Santos E. Ras in cancer and developmental diseases. *Genes Cancer.* (2011) 2:344–58. doi: 10.1177/1947601911411084

3. Malumbres M, Barbacid M, RAS. oncogenes: the first 30 years. *Nat Rev Cancer.* (2003) 3:459–65. doi: 10.1038/nrc1097

4. Harvey JJ. An unidentified virus which causes the rapid production of tumours in mice. *Nature.* (1964) 204:1104–5. doi: 10.1038/2041104b0

5. Kirsten WH, Schauf V, McCoy J. Properties of a murine sarcoma virus. *Bibl Haematol.* (1970) 36:246–9. doi: 10.1159/000391714

6. Hobbs GA, Der CJ, Rossman KL. RASisoforms and mutations in cancer at a glance. *J Cell Sci.* (2016) 129:1287–92. doi: 10.1242/jcs.182873

7. Yin G, Kistler S, George SD, Kuhlmann N, Garvey L, Huynh M, et al. GTPase K104Q mutant retains downstream signaling by offsetting defects in regulation. *J Biol Chem.* (2017) 292:4446–56. doi: 10.1074/jbc.M116.762435

8. Colicelli J. Human RAS superfamily proteins and related GTPases. *Sci STKE.* (2004) 2004:RE13. doi: 10.1126/stke.2502004re13

9. Yorimitsu T, Sato K, Takeuchi M. Molecular mechanisms of Sar/Arf GTPases in vesicular trafficking in yeast and plants. *Front Plant Sci.* (2014) 5:411. doi: 10.3389/fpls.2014.00411

10. Cherfils J, Zeghouf M. Regulation of small GTPases by GEFs, GAPs, and GDIs. *Physiol Rev.* (2013) 93:269–309. doi: 10.1152/physrev.00003.2012

11. Takács T, Kudlik G, Kurilla A, Szeder B, Buday L, Vas V. The effects of mutant Ras proteins on the cell signalome. *Cancer Metastasis Rev.* (2020) 39:1051–65. doi: 10.1007/s10555-020-09912-8

12. Han CW, Jeong MS, Jang SB. Understand KRAS and the quest for anti-cancer drugs. *Cells.* (2021) 10:cells10040842. doi: 10.3390/cells10040842

13. Healy FM, Prior IA, MacEwan DJ. The importance of Ras in drug resistance in cancer. *Br J Pharmacol.* (2022) 179:2844–67. doi: 10.1111/bph.15420

14. Merz V, Gaule M, Zecchetto C, Cavaliere A, Casalino S, Pesoni C, et al. Targeting KRAS: the elephant in the room of epithelial cancers. *Front Oncol.* (2021) 11:638360. doi: 10.3389/fonc.2021.638360

15. Ferreira A, Pereira F, Reis C, Oliveira MJ, Sousa MJ, Preto A. Crucial role of oncogenic KRAS mutations in apoptosis and autophagy regulation: therapeutic implications. *Cells.* (2022) 11:cells11142183. doi: 10.3390/cells11142183

16. Désage A-L, Léonce C, Swalduz A, Ortiz-Cuaran S. Targeting KRAS mutant in non-small cell lung cancer: novel insights into therapeutic strategies. *Front Oncol.* (2022) 12:796832. doi: 10.3389/fonc.2022.796832

17. Huang L, Guo Z, Wang F, Fu L. KRAS mutation: from undruggable to druggable in cancer. *Signal Transduct Target Ther.* (2021) 6:386. doi: 10.1038/s41392-021-00780-4

18. Lindsay CR, Garassino MC, Nadal E, Öhrling K, Scheffler M, Mazières J. On target: rational approaches to KRAS inhibition for treatment of non-small cell lung carcinoma. *Lung Cancer.* (2021) 160:152–65. doi: 10.1016/j.lungcan.2021.07.005

19. Zhu C, Guan X, Zhang X, Luan X, Song Z, Cheng X, et al. Targeting KRAS mutant cancers: from druggable therapy to drug resistance. *Mol Cancer.* (2022) 21:159. doi: 10.1186/s12943-022-01629-2

20. Rozengurt E, Eibl G. Crosstalk between KRAS, SRC and YAP signaling in pancreatic cancer: interactions leading to aggressive disease and drug resistance. *Cancers.* (2021) 13:5126. doi: 10.3390/cancers13205126

21. Adachi Y, Kimura R, Hirade K, Ebi H. Escaping KRAS: gaining autonomy and resistance to KRAS inhibition in KRAS mutant cancers. *Cancers.* (2021) 13:cancers13205081. doi: 10.3390/cancers13205081

22. Sun C, Hobor S, Bertotti A, Zecchin D, Huang S, Galimi F, et al. Intrinsic resistance to MEK inhibition in KRAS mutant lung and colon cancer through transcriptional induction of ERBB3. *Cell Rep.* (2014) 7:86–93. doi: 10.1016/j.celrep.2014.02.045

23. Nagasaka M, Li Y, Sukari A, Ou S-HI, Al-Hallak MN, Azmi AS, et al. G12C game of thrones, which direct KRAS inhibitor will claim the iron throne? *Cancer Treat Rev.* (2020) 84:101974. doi: 10.1016/j.ctrv.2020.101974

24. Mustachio LM, Chelariu-Raicu A, Szekvolgyi L, Roszik J. Targeting KRAS in cancer: promising therapeutic strategies. *Cancers.* (2021) 13:1204. doi: 10.3390/cancers13061204

25. Conroy M, Cowzer D, Kolch W, Duffy AG. Emerging RAS-directed therapies for cancer. *Cancer Drug Resist.* (2021) 4:543–58. doi: 10.20517/cdr.2021.07

26. Skoulidis F, Li BT, Dy GK, Price TJ, Falchook GS, Wolf J, et al. Sotorasib for lung cancers with KRAS pG12C mutation. *N Engl J Med.* (2021) 384:2371–81. doi: 10.1056/NEJMoa2103695

27. Hyun S, Shin D. Small-molecule inhibitors and degraders targeting KRAS-driven cancers. *Int J Mol Sci.* (2021) 22:12142. doi: 10.3390/ijms222212142

28. Strickler JH, Satake H, George TJ, Yaeger R, Hollebecque A, Garrido-Laguna I, et al. Sotorasib in KRAS pG12C-mutated advanced pancreatic cancer. *N Engl J Med.* (2023) 388:33–43. doi: 10.1056/NEJMoa2208470

29. Lanman BA, Allen JR, Allen JG, Amegadzie AK, Ashton KS, Booker SK, et al. Discovery of a covalent inhibitor of KRAS(G12C) (AMG 510) for the treatment of solid tumors. *J Med Chem.* (2020) 63:52–65. doi: 10.1021/acs.jmedchem.9b01180

30. Jänne PA, Riely GJ, Gadgeel SM, Heist RS, Ou S-HI, Pacheco JM, et al. Adagrasib in non-small-cell lung cancer harboring a KRAS(G12C) mutation. *N Engl J Med.* (2022) 387:120–31. doi: 10.1056/NEJMoa2204619

31. Tsai YS, Woodcock MG, Azam SH, Thorne LB, Kanchi KL, Parker JS, et al. Rapid idiosyncratic mechanisms of clinical resistance to KRAS G12C inhibition. *J Clin Invest.* (2022) 132:e155523. doi: 10.1172/JCI155523

32. Begum P, Goldin RD, Possamai LA, Popat S. Severe immune checkpoint inhibitor hepatitis in KRAS G12C-mutant NSCLC potentially triggered by sotorasib: case report. *JTO Clin Res Rep.* (2021) 2:100213. doi: 10.1016/j.jtocrr.2021.100213

33. Liu J, Kang R, Tang D. The KRAS-G12C inhibitor: activity and resistance. *Cancer Gene Ther.* (2022) 29:875–8. doi: 10.1038/s41417-021-00383-9

34. Chan C-H, Chiou L-W, Lee T-Y, Liu Y-R, Hsieh T-H, Yang C-Y, et al. and PI3K pathway activation confers resistance to KRAS(G12C) inhibitor sotorasib. *Br J Cancer.* (2023) 128:148–59. doi: 10.1038/s41416-022-02032-w

35. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics.* (2014) 30:923–30. doi: 10.1093/bioinformatics/btt656

36. 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

37. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* (2010) 11:R25. doi: 10.1186/gb-2010-11-3-r25

38. Lun ATL, Chen Y, Smyth GK. It's DE-licious: a recipe for differential expression analyses of RNA-seq experiments using quasi-likelihood methods in edgeR. *Methods Mol Biol.* (2016) 1418:391–416. doi: 10.1007/978-1-4939-3578-9_19

39. 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

40. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* (2014) 15:R29. doi: 10.1186/gb-2014-15-2-r29

41. Law CW, Alhamdoosh M, Su S, Dong X, Tian L, Smyth GK, Ritchie ME. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research.* (2016) 5. doi: 10.12688/f1000research.9005.1

42. McCarthy DJ, Smyth GK. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics.* (2009) 25:765–71. doi: 10.1093/bioinformatics/btp053

43. Tumminello M, Bertolazzi G, Sottile G, Sciaraffa N, Arancio W, Coronnello C, et al. Multivariate statistical test for differential expression analysis. *Sci Rep.* (2022) 12:8265. doi: 10.1038/s41598-022-12246-w

44. Vaes E, Khan M, Mombaerts P. Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes. *BMC Bioinformat.* (2014) 15:39. doi: 10.1186/1471-2105-15-39

45. 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/gkaa1074

46. 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

47. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. et al. clusterProfiler 40: A universal enrichment tool for interpreting omics data. *Innovation.* (2021) 2:100141. doi: 10.1016/j.xinn.2021.100141

48. Nagy Á, Munkácsy G, Gyorffy B. Pancancer survival analysis of cancer hallmark genes. *Sci Rep.* (2021) 11:6047. doi: 10.1038/s41598-021-84787-5

49. Yu L, Kim HT, Kasar S, Benien P, Du W, Hoang K, et al. Survival of Del17p CLL depends on genomic complexity and somatic mutation. *Clin cancer Res.* (2017) 23:735–45. doi: 10.1158/1078-0432.CCR-16-0594

50. Lánczky A, Gyorffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. *J Med Internet Res.* (2021) 23:e27633. doi: 10.2196/27633

51. Vasta JD, Peacock DM, Zheng Q, Walker JA, Zhang Z, Zimprich CA, et al. KRAS is vulnerable to reversible switch-II pocket engagement in cells. *Nat Chem Biol.* (2022) 18:596–604. doi: 10.1038/s41589-022-00985-w

52. Parikh K, Banna G, Liu S V, Friedlaender A, Desai A, Subbiah V, et al. Drugging KRAS: current perspectives and state-of-art review. *J Hematol Oncol.* (2022) 15:152. doi: 10.1186/s13045-022-01375-4

53. Indini A, Rijavec E, Ghidini M, Cortellini A, Grossi F. Targeting KRAS in solid tumors: current challenges and future opportunities of novel KRAS inhibitors. *Pharmaceutics.* (2021) 13:653. doi: 10.3390/pharmaceutics13050653

54. Yang A, Li M, Fang M. The research progress of direct KRAS G12C mutation inhibitors. *Pathol Oncol Res.* (2021) 27:631095. doi: 10.3389/pore.2021.631095

55. Kwan AK, Piazza GA, Keeton AB, Leite CA. The path to the clinic: a comprehensive review on direct KRAS(G12C) inhibitors. *J Exp Clin Cancer Res.* (2022) 41:27. doi: 10.1186/s13046-021-02225-w

56. Tanaka N, Lin JJ Li C, Ryan MB, Zhang J, Kiedrowski LA, Michel AG, et al. Clinical acquired resistance to KRAS(G12C) inhibition through a novel KRAS switch-II pocket mutation and polyclonal alterations converging on RAS-MAPK reactivation. *Cancer Discov.* (2021) 11:1913–22. doi: 10.1158/2159-8290.CD-21-0365

57. Awad MM, Liu S, Rybkin II, Arbour KC, Dilly J, Zhu VW, et al. Acquired resistance to KRAS(G12C) inhibition in cancer. *N Engl J Med.* (2021) 384:2382–93. doi: 10.1056/NEJMoa2105281

58. Maharjan M, Tanvir RB, Chowdhury K, Duan W, Mondal AM. Computational identification of biomarker genes for lung cancer considering treatment and non-treatment studies. *BMC Bioinformat.* (2020) 21:218. doi: 10.1186/s12859-020-3524-8

59. Elhamamsy AR, Metge BJ, Alsheikh HA, Shevde LA, Samant RS. Ribosome biogenesis: a central player in cancer metastasis and therapeutic resistance. *Cancer Res.* (2022) 82:2344–53. doi: 10.1158/0008-5472.CAN-21-4087

60. Kang J, Brajanovski N, Chan KT, Xuan J, Pearson RB, Sanij E. Ribosomal proteins and human diseases: molecular mechanisms and targeted therapy. *Signal Transduct Target Ther.* (2021) 6:323. doi: 10.1038/s41392-021-00728-8

61. Yan T-T, Fu X-L, Li J, Bian Y-N, Liu DJ, Hua R, et al. Downregulation of RPL15 may predict poor survival and associate with tumor progression in pancreatic ductal adenocarcinoma. *Oncotarget.* (2015) 6:37028–42. doi: 10.18632/oncotarget.5939

62. Daftuar L, Zhu Y, Jacq X, Prives C. Ribosomal proteins RPL37, RPS15 and RPS20 regulate the Mdm2-p53-MdmX network. *PLoS ONE.* (2013) 8:e68667. doi: 10.1371/journal.pone.0068667

63. Yadavilli S, Mayo LD, Higgins M, Lain S, Hegde V, Deutsch WA. Ribosomal protein S3: a multi-functional protein that interacts with both p53 and MDM2 through its KH domain. *DNA Repair.* (2009) 8:1215–24. doi: 10.1016/j.dnarep.2009.07.003

64. Li Y, Zhou Y, Li B, Chen F, Shen W, Lu Y, et al. WDR74 modulates melanoma tumorigenesis and metastasis through the RPL5-MDM2-p53 pathway. *Oncogene.* (2020) 39:2741–55. doi: 10.1038/s41388-020-1179-6

65. Li C, Ge M, Yin Y, Luo M, Chen D. Silencing expression of ribosomal protein L26 and L29 by RNA interfering inhibits proliferation of human pancreatic cancer PANC-1 cells. *Mol Cell Biochem.* (2012) 370:127–39. doi: 10.1007/s11010-012-1404-x

66. El Khoury W, Nasr Z. Deregulation of ribosomal proteins in human cancers. *Biosci Rep.* (2021) 41. doi: 10.1042/BSR20211577

67. Alam E, Maaliki L, Nasr Z. Ribosomal protein S3 selectively affects colon cancer growth by modulating the levels of p53 and lactate dehydrogenase. *Mol Biol Rep.* (2020) 47:6083–90. doi: 10.1007/s11033-020-05683-1

68. Yang HJ, Youn H, Seong KM, Jin Y-W, Kim J, Youn B. Phosphorylation of ribosomal protein S3 and antiapoptotic TRAF2 protein mediates radioresistance in non-small cell lung cancer cells. *J Biol Chem.* (2013) 288:2965–75. doi: 10.1074/jbc.M112.385989

69. Yao Y-X, Xu B-H, Zhang Y. CX-3543 promotes cell apoptosis through downregulation of CCAT1 in colon cancer cells. *Biomed Res Int.* (2018) 2018:9701957. doi: 10.1155/2018/9701957

70. Makhale A, Nanayakkara D, Raninga P, Khanna KK, Kalimutho M. CX-5461 enhances the efficacy of APR-246 via induction of DNA damage and replication stress in triple-negative breast cancer. *Int J Mol Sci.* (2021) 22:5782. doi: 10.3390/ijms22115782

71. Hilton J, Gelmon K, Bedard PL, Tu D, Xu H, Tinker A V, et al. Results of the phase I CCTG IND.231 trial of CX-5461 in patients with advanced solid tumors enriched for DNA-repair deficiencies. *Nat Commun.* (2022) 13:3607. doi: 10.1038/s41467-022-31199-2

72. Sun M-Y, Xu B, Wu Q-X, Chen W-L, Cai S, Zhang H, et al. Cisplatin-Resistant gastric cancer cells promote the chemoresistance of cisplatin-sensitive cells via the exosomal RPS3-mediated PI3K-Akt-Cofilin-1 signaling axis. *Front cell Dev Biol.* (2021) 9:618899. doi: 10.3389/fcell.2021.618899

73. Wang T, Jin C, Yang P, Chen Z, Ji J, Sun Q, et al. UBE2J1 inhibits colorectal cancer progression by promoting ubiquitination and degradation of RPS3. *Oncogene.* (2023) 42:651–64. doi: 10.1038/s41388-022-02581-7

frontiers | Frontiers in Medicine

TYPE Brief Research Report
PUBLISHED 13 July 2023
DOI 10.3389/fmed.2023.1162808

OPEN ACCESS

EDITED BY
Stefano Cacciatore,
International Centre for Genetic Engineering and Biotechnology (ICGEB), South Africa

REVIEWED BY
Wimal Pathmasiri,
University of North Carolina at Chapel Hill, United States
Cheng-Rong Yu,
National Eye Institute (NIH), United States
Yue Victor Zhang,
Shenzhen Futian Hospital for Rheumatic Diseases, China

*CORRESPONDENCE
K. H. Mok
✉ mok1@tcd.ie

RECEIVED 10 February 2023
ACCEPTED 16 June 2023
PUBLISHED 13 July 2023

CITATION
Sherlock L, Martin BR, Behsangar S and Mok KH (2023) Application of novel AI-based algorithms to biobank data: uncovering of new features and linear relationships.
Front. Med. 10:1162808.
doi: 10.3389/fmed.2023.1162808

Application of novel AI-based algorithms to biobank data: uncovering of new features and linear relationships

Lee Sherlock[1,2], Brendan R. Martin[1], Sinah Behsangar[1] and K. H. Mok[2]*

[1]Meta-Flux Ltd., Dublin, Ireland, [2]Trinity Biomedical Sciences Institute (TBSI), School of Biochemistry and Immunology, Trinity College Dublin, The University of Dublin, Dublin, Ireland

We independently analyzed two large public domain datasets that contain ^{1}H-NMR spectral data from lung cancer and sex studies. The biobanks were sourced from the Karlsruhe Metabolomics and Nutrition (KarMeN) study and Bayesian Automated Metabolite Analyzer for NMR data (BATMAN) study. Our approach of applying novel artificial intelligence (AI)-based algorithms to NMR is an attempt to globalize metabolomics and demonstrate its clinical applications. The intention of this study was to analyze the resulting spectra in the biobanks via AI application to demonstrate its clinical applications. This technique enables metabolite mapping in areas of localized enrichment as a measure of true activity while also allowing for the accurate categorization of phenotypes.

KEYWORDS

metabolomics, NMR, KarMeN, BATMAN, AI-based algorithm, lung cancer

1. Introduction

The field of metabolomics is the most recent addition to the "-Omics" discipline. The core objective of this emerging field is to record all metabolites within a biological sample. Metabolites are understood to be by-products of cellular metabolism with a weight of ∼2 kDa or less (1, 2). Water-soluble metabolites have the ability to communicate with the environment and the microbiome due to the mobility around the open biological system (3). Consequently, metabolomics is essential for "systems biology" due to its particular scope analogous to fields such as genomics and proteomics (4). "Hence, genomics and proteomics identify what could happen, metabolomics identifies what is currently happening in a system" (5). The metabolomics framework is capable of examining endogenous metabolites and signal molecules that are by-products or participate in gene regulation, protein function, and enzymatic activity. Based on these, we identify 'true activity' as a representation of what is currently happening in a biological system (5). Additionally, metabolomics is often a consequence of "exposomics", which is a series of factors that include diet, lifestyle, pollutants, medication, and the microbiome itself (Figure 1A) (7). It is particularly valuable as it is capable of capturing the thousands of small molecule interactions within a given organism (8). Therefore, a significant portion of research has been invested in the potential of tracking the downregulation and upregulation patterns of metabolites or biomarkers in order to interpret fluctuations in biological function (9, 10).

Broadly speaking, there are two metabolomics methodologies: The first is targeted metabolomics, which establishes associations between defined metabolites and known phenotypic states (1). This approach remains to be desired as it requires a deep understanding of that pre-defined state and access to bioinformatic databases to cross-validate. Alternatively, untargeted metabolomics is the widening of the search for metabolites without prior knowledge of the state in question. This unbiased and semi-quantitative approach measures thousands of small molecules simultaneously with the core objective being the development of statistical and analytical methods that allow the tracking of entire metabolic pathways and fluctuation patterns (11–13).

A potential workhorse instrumentation for untargeted metabolomics integration is nuclear magnetic resonance (NMR) due to its holistic detection capability combined with high sensitivity (though not as high as mass spectrometry) for low molecular weight biomarkers. It is typical to use NMR and mass spectrometry (MS) in tandem with multivariate analysis (14). NMR spectroscopy is a technique that exploits atomic nuclei with non-zero magnetic moments to act as tiny probes for the detection of the local structure, dynamics, reaction state, and chemical environment within molecules. NMR spectra are unique, well-resolved, analytically tractable, and often highly predictable for small molecules. NMR analysis is, therefore, used for confirming the identity of a substance. Different functional groups are easily distinguishable, and identical functional groups with differing neighbors still give distinguishable signals. Following NMR's discovery in the 1940s, a plethora of new applications have emerged, and the technique has undergone major technological developments. NMR has now become an essential tool in the fields of chemistry, physics, biology, and medicine. Potential applications of this technology exist in multiple areas including structural biology, metabolomics, food science, toxicology, natural products research, pharmaceutical reaction and process monitoring, and organic chemistry (15–17). As NMR is inherently quantitative, its ability to determine metabolite concentrations in a reproducible manner allows it to serve as an additional variable of analysis for multiple phenotypes from a variety of biofluids.

In the case of NMR, the standardized workflow generates thousands of signals which include true signals from metabolites, adducts, and fragments, as well as noise signals from contaminants and artifacts (11, 12). Due to the sheer quantity of signals generated from a single NMR workflow, it is essential to develop tools that are capable of noise reduction, aiding in the analysis of "true signals," allowing for more impactful outputs from downstream analysis. At present, there are issues regarding the scalability of technologies that are required to mainstream global metabolomics. Currently, there are software tools developed such as MVAPack, NMRProcFlow, and WorkFlow4Metabolomics. However, there are problems regarding the high-throughput applications of such software tools allowing for the development of artificial intelligence (AI) integration.

There is an abundance of applications that have demonstrated that AI is not a one size fits all; therefore, one must borrow and hybridize concepts from genome-wide association studies (GWASs) and Mummichog in an attempt to map all possible metabolite matches to a pathway via mass spectroscopy, solely focusing on regions of localized enrichment as they are assumed to be a reflection of "true activity" (18). Other methods include the Bayesian Automated Metabolite Analyzer for NMR (BATMAN) data approach, which performs spectral deconvolution using prior information on the spectral signatures of metabolites (19). When handling large metabolomic datasets, it is common to attempt to find meaning through multivariate analysis (MVA) methods such as principal component analysis (PCA) and partial least squares projection to latent structures (PLSs), all of which are attempts to segregate features that contribute to variation that are separated for further analysis, not too dissimilar from the mummichog approach (20). The recent integrations of AI into this space have seen the use of the least absolute shrinkage and selection operator (LASSO), PCA, self-organization maps (SOMs), and partial least square-discriminant analysis (PLS-DA) (8). AI is capable of identifying phenotypic variation via dimensional reduction, which indicates the biological pathway that differs among phenotypes and demonstrates the value and power these approaches have as they lend themselves to precision health (21).

Our approach involves harnessing global metabolomics in addition to multivariate analysis in tandem with NMR to investigate metabolites and their correlation with sex and lung cancer. In this study, we use the data provided by two large biobank databases. All data relating to sex were curated and analyzed by Rist et al. (22) and Bub et al. (23), while the lung cancer data were curated and analyzed by Padayachee et al. (19). The objective was to examine open-source datasets and apply our analytical techniques to observe variations and establish relationships in regions of localized enrichment. Regions of enrichment are then separated and probed for further correlations. Further probing defines the change in functional parameters induced via disease or aging. Upon examining the blood and urine, it became apparent that it was possible to identify patterns and classify participants in accordance to sex and lung cancer, with >90% accuracy.

2. Materials and methods

2.1. Data collection

For this investigation, we obtained open-source datasets from the health study by Rist et al. (22) and the lung cancer study by Padayachee et al. (19). In this study, we focused solely on the previously analyzed ^{1}H-NMR spectra of blood plasma and urine samples obtained from lung cancer patients ($n_{cases} = 69$, $n_{control} = 74$) (19) and healthy men and women ($n = 301$) (23). Procedural steps differed per study; these include fasting periods, preparation, and storage of NMR sampling.

The KarMeN study (22, 23) recruited healthy men and women (+18 years old). In addition to blood and urine sampling (tested by NMR, GC-MS, and LC-MS), a variety of anthropomorphic measurements were taken but not utilized during our analysis. The

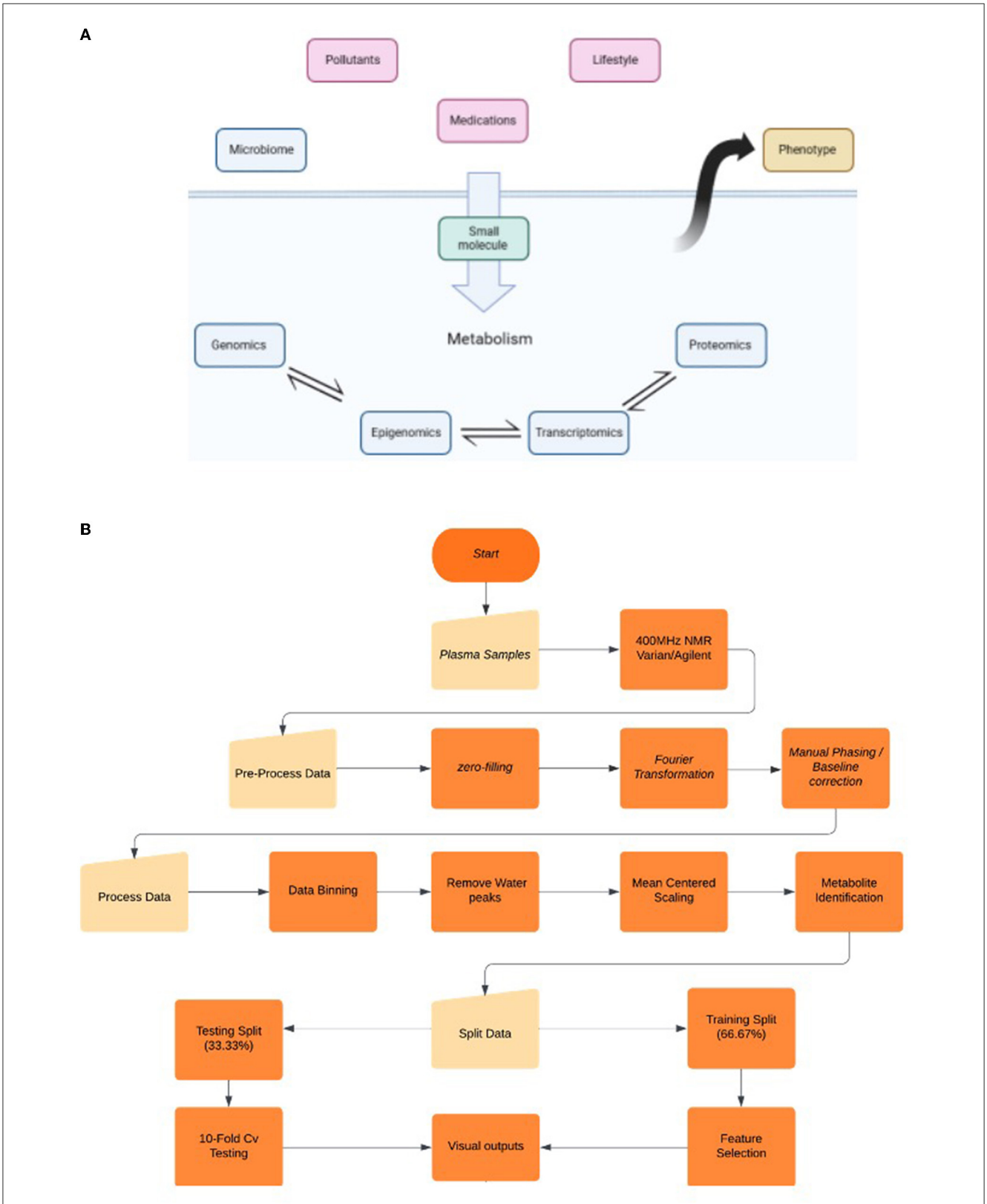


FIGURE 1
(A) Biological "omics" cascade and the factors that govern them. Targeted metabolomics focuses on the measurements of endogenous small molecules as a by-product of a metabolic pathway, while global metabolomics focuses on the fluctuation patterns and attributes said pattern to a pathway. Fluctuations in the "omics" cascade (blue layer) can be due to the influence of exogenous non-genetic factors (red) and can lead to alterations in phenotypes. Global metabolomics analysis can aid in the enhanced understanding of biomarkers/pathways and their correlation with etiology and diagnosis (6). **(B)** Workflow diagram highlighting the important milestones of the NMR and AI processes.

sole features used for this study were the ^{1}H NMR blood and urine analyses performed following a post-fasting period of 6 h, which meant that we were availing of only approximately 35% of the entire dataset provided by the study (22).

Padayachee et al. (19) collected previously analyzed data from lung cancer patients ($n_{case} = 69$) from the Limburg Positron Emission Tomography Center (Hasselt, Belgium), while the control data ($n_{control} = 74$) were from Ziekenhuis Oost-Limburg (Genk, Belgium). Additional parameters of this study included: a 6-h fastening period, a glucose level of $\geq$ 200 mg/dl, and morning medication intake.

The strict inclusion/exclusion parameters and the handling of samples in both studies gave us confidence in the integrity and excellence of both datasets, thus enabling us to perform our own analysis. The inputs we availed of were solely that of ^{1}H-NMR datasets.

2.2. Data processing

In one-dimensional ^{1}H-NMR spectroscopy, the signals are represented as the frequency domain resulting from the Fourier transform of a time-domain signal. These are given in units of parts per million (ppm), which is pre-determined at 0.0 ppm based on the chemical shift reference. Data processing was performed prior to any analysis to ensure the integrity and reliability of the results.

For the Padayachee et al. (19) data, several pre-processing steps were conducted on the 400-MHz spectra using the Varian/Agilent software. These steps involved zero-filling and multiplication by an exponential apodization function of 0.7 Hz before Fourier transformation. Additionally, the spectra underwent manual phasing, automatic baseline correction using polynomials or splines, and referencing to trimethylsilyl-2,2,3,3-tetradeuteropropionic acid (TSP) at 0.015 ppm. The final pre-processing step involved normalizing the spectra by the total area under the curve, without accounting for the water and TSP signals.

Regarding the Rist et al. (22) data, both plasma and urine samples were subjected to untargeted NMR analysis using ^{1}D ^{1}H NMR spectroscopy. Plasma samples were measured at 310 K on an AVANCE II 600 MHz NMR spectrometer equipped with a 1H-BBI probehead and a BACS sample changer, while urine samples were analyzed at 300 K on a Bruker 600 MHz spectrometer equipped with either an AVANCE III with a 1H,13C,15N-TCI inversely detected cryoprobe or an AVANCE II with a 1H-BBI room temperature probe. The plasma spectra were referenced to the ethylenediaminetetraacetic (EDTA) acid signal at 2.5809 ppm and bucketed graphically, ensuring that each bucket contained only one signal or group of signals and no peaks were split between buckets. The urine spectra were resampled for a uniform frequency axis and aligned using "correlation optimized warping." Subsequently, bucketing was performed using an in-house developed software based on Python, aiming to assign signals or groups of signals to individual buckets without splitting peaks between them. Finally, the resulting bucket tables were used for statistical analyses and machine learning algorithms.

Furthermore, the resulting pre-processing steps from the studies by Rist et al. (22) and Padayachee et al. (19) were subject to further investigation. The investigation of the above outputs was performed using Chenomx NMR Suite 8.1 (Chenomx, Edmonton, Canada) and Human Metabolome Database (HMDB) for the identification of metabolites. In addition, there were a variety of unknowns that could not be identified by harnessing either methodology. Therefore, the results section and corresponding graphs contain these unknown variables that can be identified as "Unknown – PPM".

The data obtained from the study by Padayachee et al. (19) required further processing steps in an attempt to reduce the background noise and increase the overall resolution of the data. This was conducted by binning the data into further sub-intervals of 0.01 ppm. Conversely, the same approach could not be conducted on the data obtained from the study by Rist et al. (22) as the binning was conducted in-house and correlated with pre-defined metabolites. The difference in binning processes and MHz may be factors that allowed for variation in the results.

As per common practice in NMR, we removed water and its corresponding ppm as this often accounts for the majority of peak intensity and can mask minor variations in the NMR spectra. Due to the difference in obtained data, standardization was required, whereby the negative values within the dataset were set to zero and mean-centered scaling was applied to the Rist et al. (22) data. Feature values were transformed to follow a uniform or normal distribution for the Padayachee et al. (19) data. This helped to stabilize the variance and minimize the effects of outliers, resulting in improved performance of the predictive model. Scaling is important as it facilitates a fair comparison between different features.

Finally, the dataset was divided into two sets: a test set comprising 33% of the data and a training set with 66% of the data. This partitioning ensures an unbiased evaluation of the algorithm's performance. To determine the significance of different features in the dataset, the widely adopted statistical test known as the ANOVA F-test was employed for feature selection. In order to comprehensively evaluate the algorithm, a 10-fold cross-validation technique was applied. This method is commonly employed in machine learning to assess the algorithm's performance across multiple subsets of the dataset. By dividing the data into 10 equal parts, the algorithm was trained and evaluated 10 times, each time using a different combination of nine parts for training and one part for testing. This approach provides a more robust assessment of the algorithm's generalization capability and overall performance.

3. Results

The data were generated by obtaining open-source datasets from the Rist et al. (22) and Padayachee et al. (19) lung cancer studies. In this study, we focused solely on the previously analyzed ^{1}H-NMR spectra of blood plasma and urine samples obtained from lung cancer patients ($n_{cases} = 69$, $n_{control} = 74$) (19) and healthy men and women ($n = 301$) (23). The data were structured and analyzed using our own in-house artificial intelligence (AI) and machine learning (ML) combined with classic statistical approaches to isolate features of interest and hone in

on localized regions of enrichment for further analysis and to correlate said features with individual metabolites and extrapolate for metabolites that are predictive of phenotypes of interest. The analysis in this section was performed via global metabolomics, which demonstrates simultaneous analyses of multiple features to categorize a phenotype of interest. The figures below show heatmaps, minimum spanning trees, boxplots, volcano plots, and PLS to demonstrate the phenotypic categorization, which lends itself to clinical capabilities.

We tested the integrity of our outputs by comparing them to the published analyses of the original datasets (19, 22). The mean specificity - which describes the amount of correctly predicted positives or "regions of enrichment" - we obtained was 0.97 for the KarMeN study (22) and 0.93 when distinguishing lung cancer of Padayachee et al. (19). Additionally, the precision of the model, which describes the portion of true positives among actual positives, was measured to be 0.96 in KarMeN and 0.93 in the Padayachee et al. study. The above statistics can be represented on a scale of 0–1, where 0 represents poor performance and 1 perfect performance.

3.1. Lung cancer case study

Our analysis of the data provided from the Bayesian Automated Metabolite Analyzer lung cancer study (19) yielded an overall 0.92 accuracy, with a mean specificity of 0.90 and a mean sensitivity of 0.93. The healthy precision value was 0.93, with a recall of 0.91 and an f1-score of 0.92. For the disease precision, it was 0.90, with a recall of 0.93 and an f1-score of 0.91. The area under the receiver operating characteristic curve (AUC-ROC) is calculated by plotting the true positive rate against false positive, where 1 represents perfect and 0.5 worst. The Padayachee et al. (19) data had an AUC-ROC of 0.92 (Figures 2–6).

Figure 2A is a heatmap of leading features in lung cancer cohorts. The leading 20 metabolites contained in this heatmap are essential for characterizing phenotypic states. Of these 20, we have found asparagine, creatine, glycerol, threonine, glucose, citrate, and lactate. Moreover, we have identified tartaric acid, which was not on the list of key metabolites in the Padayachee et al. (19) study. Interestingly, tartaric acid is known as a lung cancer biomarker and can be found in HMDB (24).

Our *in silico* analysis provided the following: Figures 3A and B are graphical outputs to visualize metabolomic relationships distilled down from a total of approximately 2 million relationships. The distillation of these relationships is further represented in Figures 4A and 5A which highlight the variability in the top-ranking metabolites. In summary, we have funneled down the key metabolites involved in lung cancer.

3.2. KarMeN health analysis among sexes

Our analysis of the data provided from the Karlsruhe Metabolomics and Nutrition study (22, 23) predicted sex solely using ^{1}H-NMR data derived from plasma, yielding an overall accuracy of 0.95, with a mean specificity of 0.97 and a mean sensitivity of 0.92. The male precision value was 0.95, with a recall of 0.97 and an f1-score of 0.96. For the female precision, it was 0.96, with a recall of 0.93 and an f1-score of 0.94. The AUC-ROC was computed to be 0.95 (Figures 2–6).

Figure 2B is a heatmap of leading features in the determination of sex in healthy cohorts. The leading 20 metabolites contained in this heatmap are essential for characterizing phenotypic states. Of these 20, we have found creatinine, creatine, glycerol, glycine, sarcosine, isoleucine, and valine. Moreover, we have identified 2-hydroxy-2-methylbutyric (HMB) acid, which was not in the list of key metabolites in the Rist et al. (22) study.

Figures 6A and B are graphical outputs to visualize metabolomic relationships distilled down from a total of approximately 2 million relationships. The distillation of these relationships is further represented in Figures 4B, 5B, which highlight the variability in the top-ranking metabolites. In summary, we have funneled down the key metabolites involved in distinguishing sex in healthy people.

4. Discussion

The primary objective of this study was to analyze the human metabolome in the plasma by way of globalized metabolomics profiling by harnessing ^{1}H-NMR, to determine the factors that significantly impact the metabolic profile of a healthy cohort compared to a lung cancer cohort, and to distinguish the variables among the sexes. Therefore, we performed our study and established a strict *in silico* experimental standardization, which we applied to data structuring, data treatment, and post-analysis treatments. When collecting open-source data, we ensured that all sample collections were standardized in terms of fasting, collection time points, and general pre-analysis handling. We also searched for healthy datasets with strict exclusion and inclusion criteria that excluded groups that suffered from acute or chronic diseases or were on medication, as we wanted a dataset that represented "true health," thereby decreasing variation. In contrast, the medication and acute/chronic disease exclusion criteria cannot be applied to the lung cancer cohort as they must undergo medical treatment in tandem with the study. Furthermore, this fundamental difference may be one variable that explains the variability when testing the integrity of the algorithm. Through additional analysis, we found that our process is capable of generating high-integrity categorization with minimal variation. The difference among predictive capabilities per dataset could be due to the number of samples; $n = 301$ (22) and $n = 143$ (19). More specifically, Rist et al. (22) binned 138 sex features as pre-determined metabolites, while 1,134 features were binned as 0.01 ppm increments in the data of Padayachee et al. (19).

Furthermore, some AI algorithms may require a relatively small amount of data to achieve satisfactory results, while others, particularly deep learning algorithms, often benefit from large-scale datasets. The size of the dataset required is directly proportional to the type of AI used and its field of application. Even a large dataset may not be useful if it is noisy, incomplete, or biased. A primary issue is the problem of complex, highly specialized, and specific fields focusing on molecular interactions, protein structures, or drug discovery that typically require domain expertise

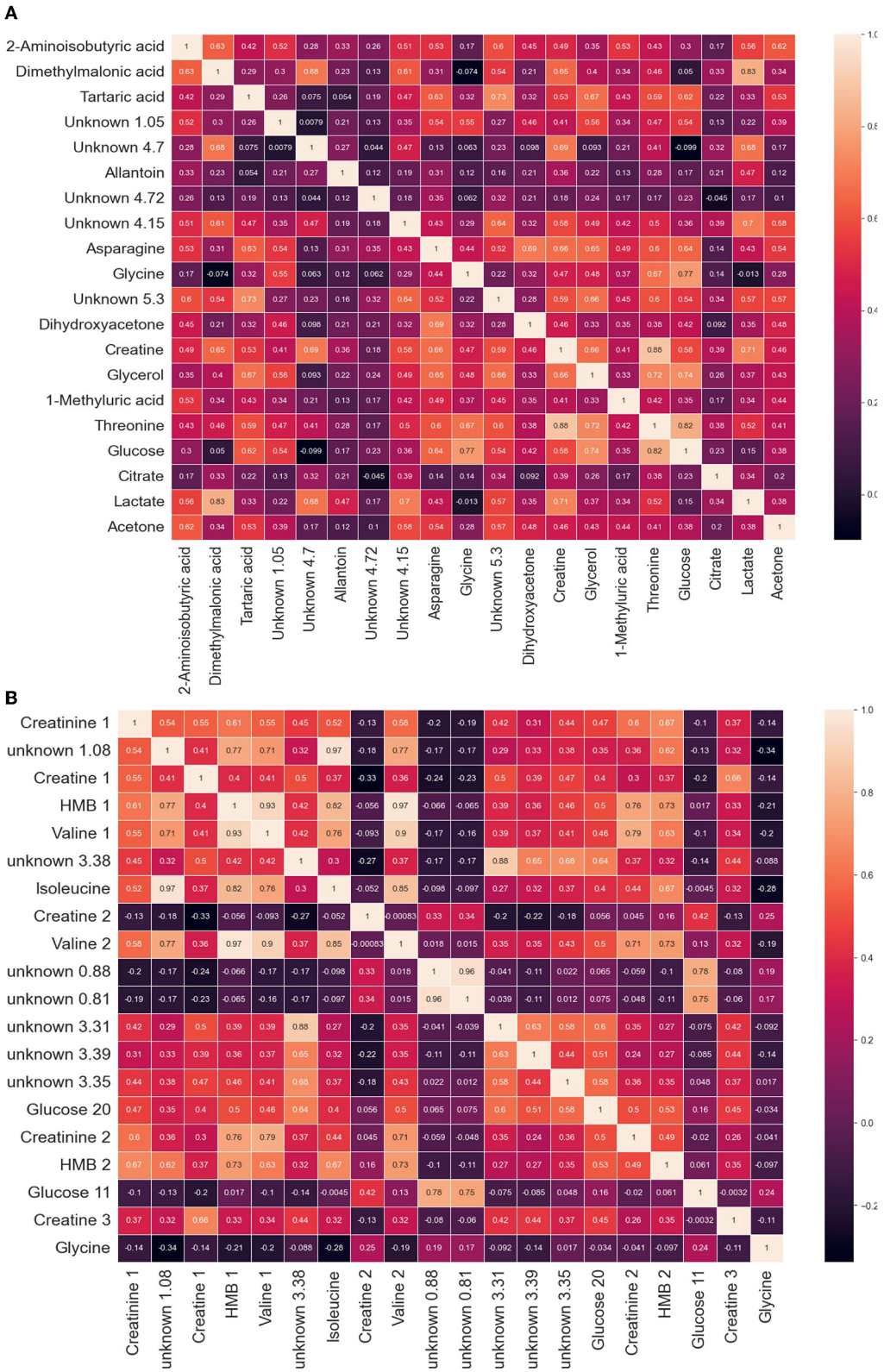


FIGURE 2
Heatmap of leading features in **(A)** lung cancer cohorts and in **(B)** health and sexes. This heatmap is a representation of the top features and the correlations relative to other features. The feature was determined by a singular NMR unit (bin or bucket), measured in units of chemical shift (ppm). The location of the ppm was determined by ANOVA F-values. The features found through NMR analysis of plasma can be used to categorize the **(A)** lung cancer metabolome and **(B)** among sexes and determine the states of health.

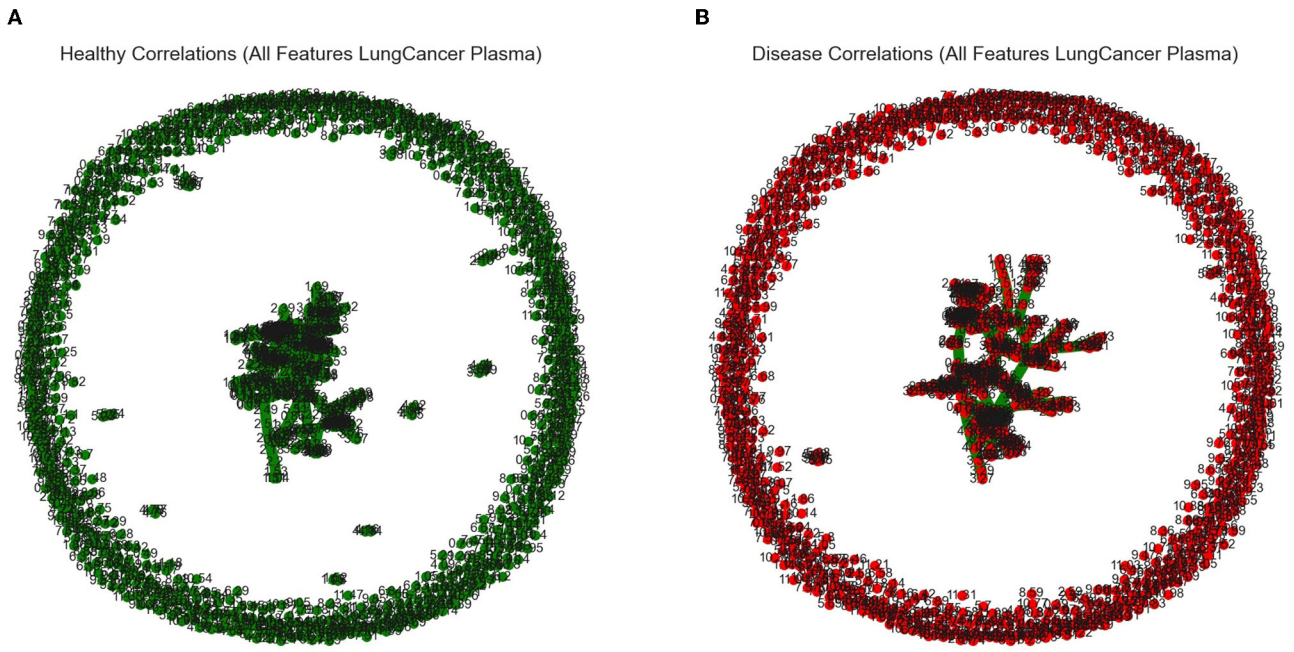


FIGURE 3
Graphical outputs visualizing the linear relationship between ppm. **(A)** Minimum spanning tree (Mst) generated by the Fruchterman–Reingold algorithm used to visualize all ppm in the healthy category with correlations above a 90% threshold. Nodes closer together in the center have a stronger correlation and nodes far apart around the perimeter have little to no correlation. **(B)** Mst used to visualize all ppm in the diseased category with correlations above a 90% threshold.

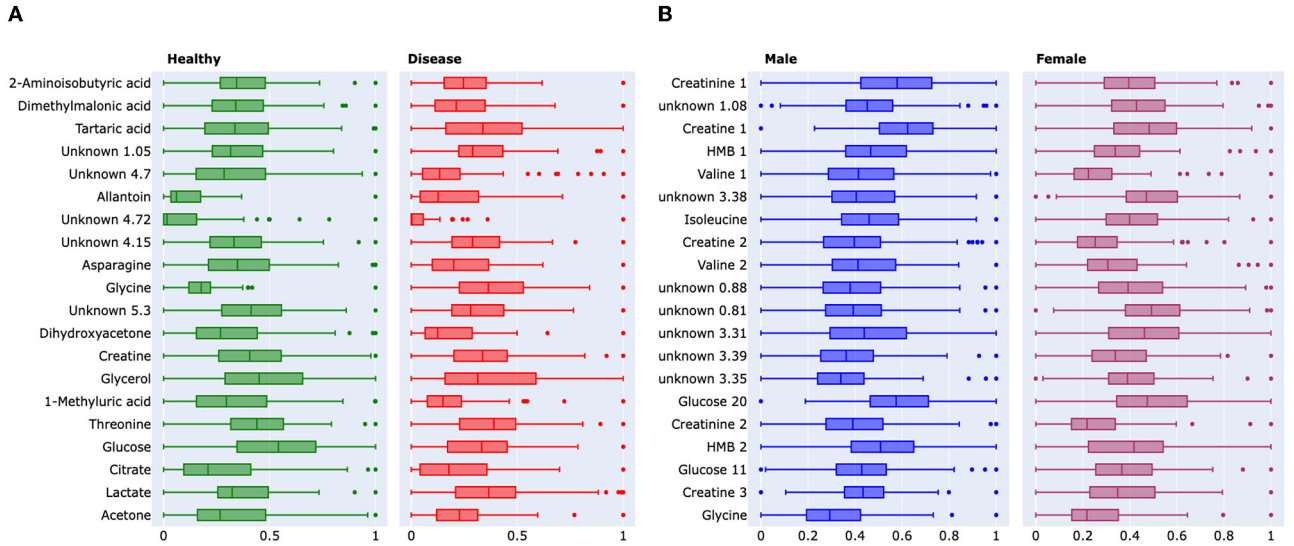


FIGURE 4
(A) Boxplots demonstrating the significance of changes between healthy controls and lung cancer groups and between male–female cohorts. The boxplot demonstrates the absolute difference between the means of each feature. These features were further analyzed and identified to be the following metabolites; 2-aminoisobutyric acid, dimethylmalonic acid, tartaric acid, and glycine. These identified metabolites were among the lead features used to categorize the phenotypes of interest. Green represents the healthy controlled cohort, while red represents the lung cancer cohort. The binned NMR spectral data from the Padayachee et al. (19) study were used to generate these graphs. **(B)** Boxplot demonstrates the absolute difference of the means of each feature. These features were further analyzed and identified to be the following metabolites; creatinine 1, creatine 1, 2-hydroxy-2-methylbutyric (HMB) acid, valine 1, valine 2, isoleucine, and glucose 20. These identified metabolites were among the lead features used to categorize the phenotypes of interest, while other points of interest include U 0.88 ppm and U 1.08 ppm. Blue represents the male cohort, while purple represents the female cohort. The binned Plasma NMR spectral data from the Bub et al. study were used to generate these graphs.

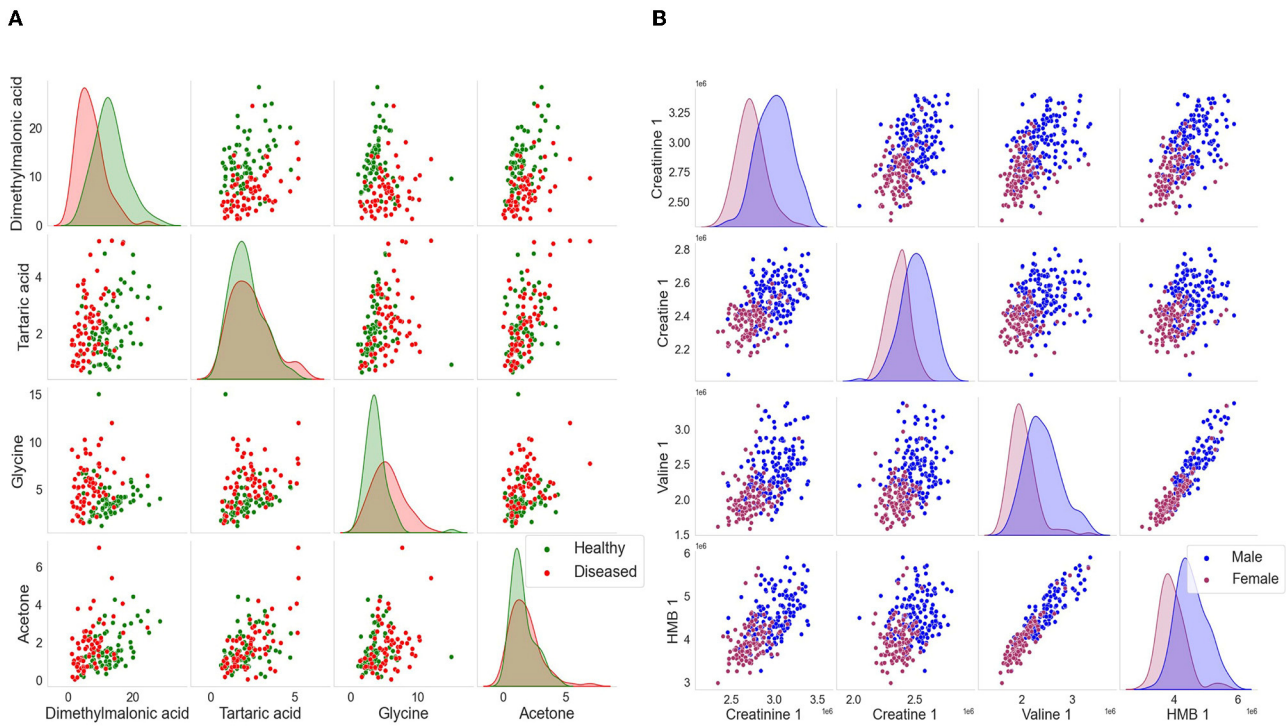


FIGURE 5
Kernel density plot used to visualize the distribution of lung cancer and the distribution of male–female cohorts. The above scatter plots demonstrate a clear separation among the cohorts. **(A)** For the lung cancer cohorts, the features of interest include dimethylmalonic, tartaric acid, glycine, and acetone. **(B)** For the distribution of sexes, the features of interest include creatinine 1, creatine 1, 2-hydroxy-2-methylbutyric (HMB) acid, and valine 1.

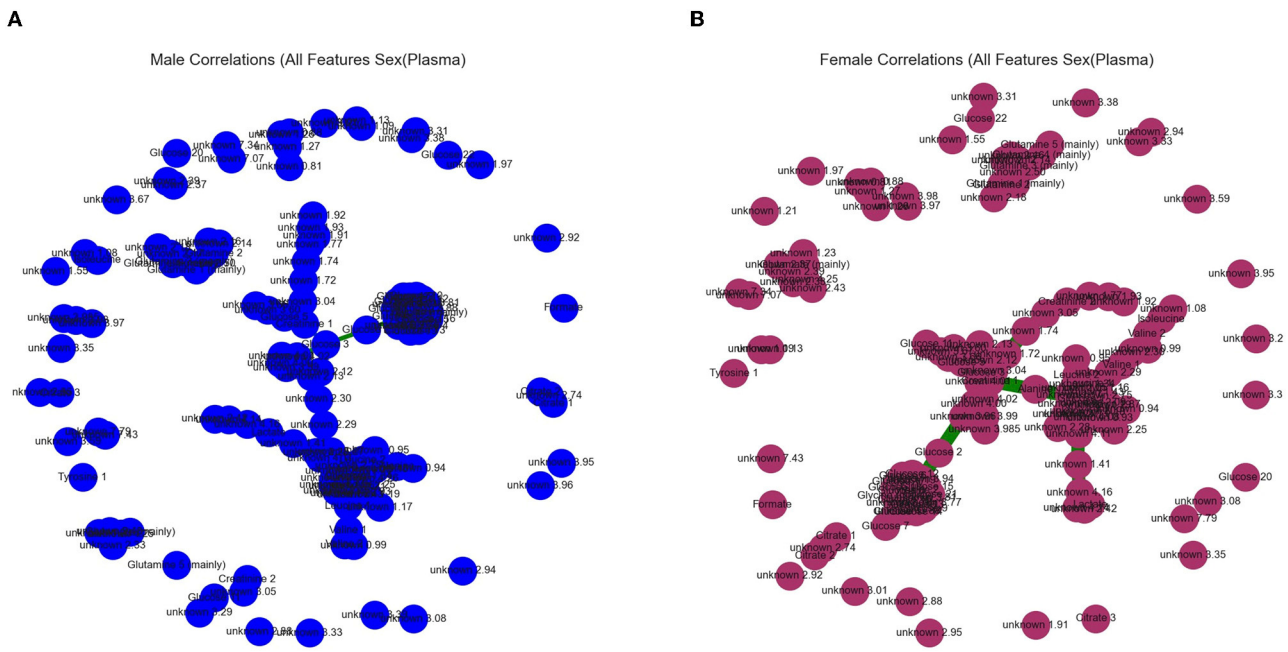


FIGURE 6
Graphical outputs visualizing the linear relationship between ppm. **(A)** Minimum spanning tree (Mst) generated by the Fruchterman–Reingold algorithm used to visualize all ppm in the male (blue) category with correlations above a 90% threshold. Nodes closer together in the center have a stronger correlation and nodes far apart around the perimeter have little to no correlation. **(B)** Mst used to visualize all ppm in the female (purple) category with correlations above a 90% threshold.

and specialized knowledge. As a result, the problem space is more constrained, and the available data may be more targeted and focused. In such cases, a smaller sample size can still provide meaningful insights and accurate predictions.

The impact of our analytical approach can be found in Figure 4. Many of our leading 20 metabolites have significant overlap with the pre-existing analysis (19, 22). Along with these, we have uncovered previously unidentified metabolites, such as tartaric acid and 2-hydroxy-2-methylbutyric acid (HMB), in lung cancer and sex identification, respectively (22, 24). We wish to emphasize that Rist et al. utilized clinical chemistry, liquid chromatography, and mass spectrometry along with NMR spectroscopy to identify the top metabolites. However, our analysis only required one-third of the original dataset, and we only utilized the NMR dataset. Despite this, our analysis has uncovered not only similar metabolites but also those which are unique.

We recognize that there are requirements for additional analysis and broadening of the inclusion criteria. Participants that are obese and/or smoking must be included and recorded for an accurate representation of the healthy population, as studies demonstrate that nicotine does have neuroprotective qualities (25); therefore, we can assume their metabolic profile would be variable. We also need to recognize the influence of "exposomics" and how it can greatly influence the "omics" cascade, especially those that are variable per region, such as carcinogens and diet (Figure 1A) (6).

Owing to the fact that NMR metabolomics provides a quantitative and holistic view of all of the metabolites contained, there is no reason that this technology cannot be applied to other diseases. In this article, we have successfully harnessed AI and metabolomic techniques to broaden the search parameters that aid in a comprehensive understanding of disease and wellbeing. The advancements made here can offer a snapshot of the entire biological system, which allows us to ascertain an accurate understanding of the phenotype in question, paving the way for true precision medicine.

5. Conclusion

From our analyses of NMR spectra from two separate biobanks, we have established that our approach has direct clinical applications. Our approach of harnessing AI and NMR to globalize metabolomics enables us to identify metabolites, to highlight them as regions of localized enrichment as a measure of true activity, while enabling us to accurately categorize phenotypes of interest.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: Padayachee et al. (19) and Rist et al. (22). Subsequently, the data was formatted, converted and processed, and are made available in this publication.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

LS and KHM initially discussed the potential of this research. LS, BM, and SB were involved in the coding and statistical evaluation of the data. LS, BM, and KHM wrote the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The authors thank Enterprise Ireland (EI) for their mentorship and encouragement.

Acknowledgments

The authors thank Alsessor, an AI accelerator program sponsored by the Tangent of Trinity College Dublin, for initial mentoring and encouragement.

Conflict of interest

LS, BM, and SB was employed by the Meta-Flux Ltd.

The remaining author declares 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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1162808/full#supplementary-material

References

1. Johnson C, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol.* (2016) 17:451–9. doi: 10.1038/nrm.2016.25

2. Beger RD, Dunn W, Schmidt MA, Gross SS, Kirwan JA, Cascante M. Metabolomics enables precision medicine: A White Paper, Community perspective. *Metabolomics.* (2016) 12:149–50. doi: 10.1007/s11306-016-1094-6

3. Turnbaug P, Ley R, Hamady M, Fraser-Ligget C,. Knight, R, Gordon, J. The human microbiome project. *Nature.* (2007) 449:804–10. doi: 10.1038/nature06244

4. Riekeberg E, Powers R. New frontiers in metabolomics: from measurement to insight. *F1000Res.* (2017) 6:1148. doi: 10.12688/f1000research.11495.1

5. Sherlock L, Mok KH. "Metabolomics and Its Applications to Personalized Medicine" in *EKC 2019 Conference Proceedings*, Springer, Cham (2021), p 25-42.

6. Zhang A, Sun H, Yan G, Wang P, Wang X. Metabolomics for Biomarker Discovery: Moving to the Clinic. *Biomed Res Int.* (2015) 2015:e354671. doi: 10.1155/2015/354671

7. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: where chemistry meets biology. *Science.* (2020) 367:392–6. doi: 10.1126/science.aay3164

8. Lauren M. Petrick, Noam S. AI/ML-driven advances in untargeted metabolomics and exposomics for biomedical applications. *Cell Rep. Phys. Sci.* (2002) 3:2666–3864. doi: 10.1016/j.xcrp.2022.100978

9. Ter Kuile BH, Westerhoff HV. Transcriptome meets metabolome hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* (2001) 500:169–71. doi: 10.1016/S0014-5793(01)02613-8

10. Guo L, Milburn MV, Ryals JA, Lonergan SC, Mitchell MW, Wulff JE. Plasma metabolomic profiles enhance precision medicine for volunteers of normal health. *Proc Natl Acad Sci USA*. (2015) 112:4901–10. doi: 10.1073/pnas.1508425112

11. Nash WJ, Dunn WB. From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. *TrAC Trends Anal Chem.* (2018) 120:e115324. doi: 10.1016/j.trac.2018.11.022

12. Chong J, Yamamoto M, Xia J. Metabo Analyst R 2.0: From raw spectra to biological insights. *Metabolites.* (2019) 9:57–8. doi: 10.3390/metabo9030057

13. Roberts LD, Souza AL, Gerszten RE. Targeted metabolomics. *Curr Protoc Mol Biol.* (2012) 30:1–24. doi: 10.1002/0471142727.mb3002s98

14. Duarte IF, Diaz SO, Gil AM, NMR. metabolomics of human blood and urine in disease research. *J Pharm Biomed Anal.* (2014) 93:17–26. doi: 10.1016/j.jpba.2013.09.025

15. Louis E, Cantrelle FX, Mesotten L, Reekmans G, Bervoets L, Vanhove K. Metabolic phenotyping of human plasma by 1 H-NMR at high and medium magnetic field strengths: a case study for lung cancer. *Magn Reson Chem.* (2017) 55:706–13. doi: 10.1002/mrc.4577

16. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. *Mass Spec Rev.* (2007) 26:51–78. doi: 10.1002/mas.20108

17. Dumez J-N, Milani J, Vuichoud B, Bornet A, Lalande-Martin J, Tea I. Hyperpolarized NMR of plant and cancer cell extracts at natural abundance. *Analyst.* (2015) 140:5860–3. doi: 10.1039/C5AN01203A

18. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA. Predicting network activity from high throughput metabolomics. *PLOS Comput Biol.* (2013) 9:e1003123. doi: 10.1371/journal.pcbi.1003123

19. Padayachee T, Khamiakova T, Louis E, Adriaensens P, Burzykowski T. The impact of the method of extracting metabolic signal from 1H-NMR data on the classification of samples: A case study of binning and BATMAN in lung cancer. *PLoS One.* (2019) 14:e0211854. doi: 10.1371/journal.pone.0211854

20. Worley B. Powers R. *Multivariate Analysis in Metabolomics*" *Curr Metabol.* (2013) 1:92–107. doi: 10.2174/2213235X11301010092

21. Mazzella M, Sumner SJ, Gao S, Su L, Diao N, Mostofa G. Quantitative methods for metabolomic analyses evaluated in the children's health exposure analysis resource (CHEAR). *J Expo Sci Environ Epidemiol.* (2020) 30:16–27. doi: 10.1038/s41370-019-0162-1

22. Rist MJ, Roth A, Frommherz L, Weinert CH, Kruéger R, Merz B, et al. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLoS ONE.* (2017) 12:e0183228. doi: 10.1371/journal.pone.0183228

23. Bub A, Kriebel A, Dörr C, Bandt S, Rist M, Roth A. The karlsruhe metabolomics and nutrition (KarMeN) study: protocol and methods of a cross-sectional study to characterize the metabolome of healthy men and women. *JMIR Res Protoc.* (2016) 5:e2603148. doi: 10.2196/resprot.5792

24. Stretch C, Eastman T, Mandal R, Eisner R, Wishart DS, Mourtzakis M. Prediction of skeletal muscle and fat mass in patients with advanced cancer using a metabolomic approach. *J Nutr.* (2012) 142:14–21. doi: 10.3945/jn.111.147751

25. Ferrea S, Winterer G. Neuroprotective and neurotoxic effects of nicotine. *Pharmacopsychiatry.* (2009) 42:255–65. doi: 10.1055/s-0029-1224138

frontiers | Frontiers in Medicine

TYPE Original Research
PUBLISHED 31 August 2023
DOI 10.3389/fmed.2023.1149860

OPEN ACCESS

EDITED BY
Balu Kamaraj,
Imam Abdulrahman Bin Faisal University,
Saudi Arabia

REVIEWED BY
Thiyagarajan Ramesh,
Prince Sattam Bin Abdulaziz University,
Saudi Arabia
Achraf El Allali,
Mohammed VI Polytechnic University, Morocco
Yutian Zou,
Sun Yat-sen University Cancer Center
(SYSUCC), China

*CORRESPONDENCE
Nehad M. Alajez
✉ nalajez@hbku.edu.qa

RECEIVED 23 January 2023
ACCEPTED 25 July 2023
PUBLISHED 31 August 2023

CITATION
Vishnubalaji R and Alajez NM (2023) Long non-coding RNA AC099850.4 correlates with advanced disease state and predicts worse prognosis in triple-negative breast cancer. *Front. Med.* 10:1149860.
doi: 10.3389/fmed.2023.1149860

Long non-coding RNA AC099850.4 correlates with advanced disease state and predicts worse prognosis in triple-negative breast cancer

Radhakrishnan Vishnubalaji[1] and Nehad M. Alajez[1,2]*

[1]Translational Cancer and Immunity Center (TCIC), Qatar Biomedical Research Institute (QBRI), Hamad Bin Khalifa University (HBKU), Qatar Foundation (QF), Doha, Qatar, [2]College of Health and Life Sciences, Hamad Bin Khalifa University (HBKU), Qatar Foundation (QF), Doha, Qatar

Our understanding of the function of long non-coding RNAs (lncRNAs) in health and disease states has evolved over the past decades due to the many advances in genome research. In the current study, we characterized the lncRNA transcriptome enriched in triple-negative breast cancer (TNBC, $n = 42$) and estrogen receptor (ER+, $n = 42$) breast cancer compared to normal breast tissue ($n = 56$). Given the aggressive nature of TNBC, our data revealed selective enrichment of 57 lncRNAs in TNBC. Among those, AC099850.4 lncRNA was chosen for further investigation where it exhibited elevated expression, which was further confirmed in a second TNBC cohort ($n = 360$) where its expression correlated with a worse prognosis. Network analysis of AC099850.4^{high} TNBC highlighted enrichment in functional categories indicative of cell cycle activation and mitosis. Ingenuity pathway analysis on the differentially expressed genes in AC099850.4^{high} TNBC revealed the activation of the canonical kinetochore metaphase signaling pathway, pyridoxal 5′-phosphate salvage pathway, and salvage pathways of pyrimidine ribonucleotides. Additionally, upstream regulator analysis predicted the activation of several upstream regulator networks including CKAP2L, FOXM1, RABL6, PCLAF, and MITF, while upstream regulator networks of TP53, NUPR1, TRPS1, and CDKN1A were suppressed. Interestingly, elevated expression of AC099850.4 correlated with worse short-term relapse-free survival (log-rank $p = 0.01$). Taken together, our data are the first to reveal AC099850.4 as an unfavorable prognostic marker in TNBC, associated with more aggressive clinicopathological features, and suggest its potential utilization as a prognostic biomarker and therapeutic target in TNBC.

KEYWORDS

noncoding RNA, lncRNA, AC099850.4, biomarkers, triple negative breast cancer, prognosis

Introduction

Breast cancers represent a diverse group of cancers with different underlying biological features exhibiting differences in their clinical management, responses to treatment, and clinical outcomes (1). Recent advances in genomic research led to the BC classification of defined molecular subtypes, based on hormone receptor (HR), including estrogen receptor (ER) and progesterone receptor (PR), expression, as well as ERBB2 [also known as human epidermal growth factor receptor 2 (HER2)]

amplification, while tumors lacking overexpression of HR and lacking HER2 amplifications are referred to as triple-negative breast cancer (TNBC), comprising ∼10–20% of all breast cancers. TNBC is oftentimes diagnosed at a younger age and has more aggressive clinicopathological features at presentation (larger tumor size, higher grade, and lymph node involvement) compared to other breast cancer subtypes. TNBC is also classified based on mRNA expression into four intrinsic subtypes: basal-like and immune suppressed (BLIS), immunomodulatory subtype (IM), mesenchymal-like subtype (MES), and luminal androgen receptor (LAR) subtype, with BLIS being the most aggressive subtype (2). While most of the research on breast cancer classification has focused on protein-coding mRNAs, the utilization of non-coding RNAs (ncRNAs), including miRNA and long non-coding RNAs (lncRNAs), is currently gaining momentum for breast cancer classification and as diagnostic and prognostic biomarkers (3–5). In our previous analysis, we identified 13 lncRNAs that were able to discriminate TNBC from normal breast tissue (3). A previous study by Huang et al. reported low NEAT1^{low} and MAL2^{high} to predict unfavorable outcomes in TNBC (6). In another study, Song et al. reported low-NEF lncRNA expression to correlate with poor prognosis in TNBC (7), thus corroborating a prognostic value for several lncRNA in TNBC.

lncRNAs represent a major class of ncRNAs with lengths exceeding 200 nucleotides and a lack of functional protein translation. lncRNAs can be divided into six different groups based on their genomic positions, subcellular localizations, and functions: (1) enhancer lncRNAs, (2) intronic lncRNAs, (3) antisense lncRNAs, (4) sense lncRNA, (5) intergenic lncRNA, and (6) bidirectional lncRNAs (8, 9). Increasing evidence has implicated lncRNAs in the onset and progression of various human cancers, through the regulation of key cellular processes, including proliferation, migration, invasion, and apoptosis at the transcriptional and post-transcriptional levels (10). Phase II/III clinical trials highlighted the potential use of RNA-based therapeutics, including antisense oligonucleotides (ASOs) and small interfering RNAs (siRNAs) to treat various human diseases (11).

Compelling data have implicated lncRNAs in regulating various biological processes, which could play oncogenic or tumor suppressor roles in breast cancer (12–15). Our data recently highlighted the prognostic and therapeutic functions of MALAT1 and LINC00511 in TNBC (16, 17).

In the current study, we characterized the differentially expressed lncRNAs in TNBC and ER$^+$ breast cancers compared to normal breast tissues. Given the aggressive nature and lack of targeted therapies for TNBC, we subsequently aimed at identifying unique lncRNA transcripts expressed in TNBC, but not ER$^+$ BC, which could potentially be used as prognostic biomarkers and therapeutic targets. Subsequently, we focused our study on AC099850.4 (alternatively named lnc-SKA2-1, AC099850.3, or ENSG00000265415), revealing AC099850.4 as a novel prognostic biomarker associated with unfavorable disease outcomes in TNBC. Comprehensive bioinformatics and network analysis revealed a plausible role of AC099850.4 in cell cycle regulation.

Results

To provide a global overview of the differentially expressed lncRNAs in different BC subtypes, transcriptomic data from 42 TNBC, 42 ER$^+$HER2$^-$ (referred to as ER$^+$ throughout the article), and 56 normal breast tissues (NT) were pseudo-aligned to the GENCODE release (V33) reference genome using Kallisto. Data presented in Figure 1 revealed a distinct lncRNA expression profile for the indicated breast cancer molecular subtypes compared to NT (Figure 1A, Supplementary Table S1). Concordantly, PCA analysis revealed similar segregation of TNBC from ER$^+$ and NT (Figure 1B). Our analysis revealed 226 lncRNAs that were upregulated in TNBC vs. NT and in ER$^+$ vs. NT (Figure 1C). Interestingly, we identified 57 lncRNAs that were upregulated in TNBC vs. ER$^+$ and in TNBC vs. NT, but not in ER$^+$ vs. NT, suggesting their specific expression in TNBC (Figure 1C).

AC099850.4 expression correlates with advanced tumor grade and worse prognosis

Among the identified TNBC-enriched lncRNAs, AC099850.4 was chosen for further analysis since its expression was enriched in TNBC and has not been implicated in TNBC thus far. The expression AC099850.4 in TNBC, ER$^+$, and NT is shown in Figure 2A. We subsequently confirmed the upregulated expression of AC099850.4 in a larger cohort of TNBC ($n = 360$) compared to normal ($n = 88$) exhibiting 2.2 fc, p(Adj) $= 1.3 \times 10^{-30}$, as shown in Figure 2B. Interestingly, we observed the highest expression of AC099850.4 in TNBC with advanced tumor grade (Figure 2C) and the BLIS TNBC subtype exhibiting the worst prognosis (18) (Figure 2D).

Elevated expression of AC099850.4 correlates with the mitotic cell cycle in TNBC

To better understand the role of AC099850.4 in driving TNBC, the cohort of 360 TNBC was grouped into AC099850.4^{high} ($n = 180$) and AC099850.4^{low} ($n = 180$). We subsequently analyzed the corresponding protein-coding transcriptome of the AC099850.4^{high} vs. AC099850.4^{low} using the GENCODE v33 reference genome. Our data revealed a remarkable difference in mRNA expression between the AC099850.4^{high} vs. AC099850.4^{low}, with majority of functional enrichment being in categories indicative of proliferation and mitosis (Figure 3A). Differentially expressed genes in AC099850.4^{high} are illustrated as volcano

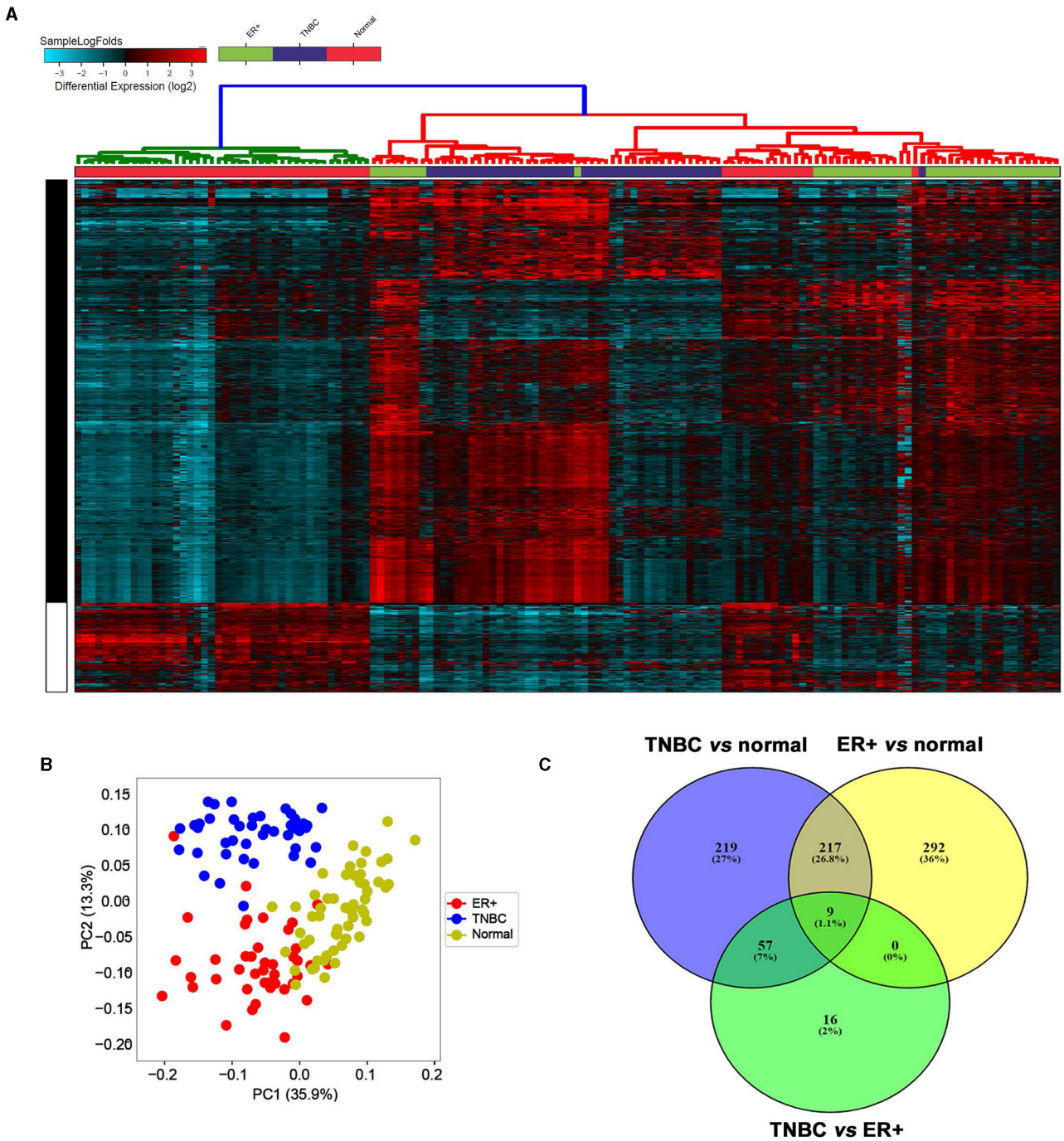


FIGURE 1
LncRNA transcriptional landscape in different breast cancer subtypes and normal breast tissue. **(A)** Hierarchical clustering of TNBC ($n = 42$), ER$^+$ breast cancer ($n = 42$) and normal breast tissue ($n = 56$) based on differentially expressed lncRNAs. Each column represents one sample, and each row represents a single lncRNA. The expression level of each lncRNA (log2) is depicted according to the color scale. **(B)** Principal component analysis (PCA) for the lncRNA transcriptome of TNBC, ER$^+$ breast cancer, and normal breast tissue. **(C)** Venn diagram depicting the overlap between upregulated lncRNAs in TNBC vs. normal, ER$^+$ vs. normal, and TNBC vs. ER$^+$.

plot (Figure 3B). Protein–protein interaction (PPI) analysis on the upregulated genes in AC099850.4^{high} vs. AC099850.4^{low} revealed strong network interaction with the highest enrichment in cell cycle-related processes, where the expression of cell cycle regulators (TRIP13, MYBL2, BRIP1, UBE2S, ANLN, NUF2, CCNB2, MELK, PLK1, TPX2, BIRC5, AURKB, TYMS, NCAPD2, FOXM1, UBE2C, IQGAP3, CENPF, NEK2, ASPM, MKI67, TTK, CEP55, KIF2C, CDC20, CKS2, PTTG1, PRC1, CDK1, KIFC1, STMN1, TOP2A, and CDKN2A) was enriched in AC099850.4^{high} (Figure 4).

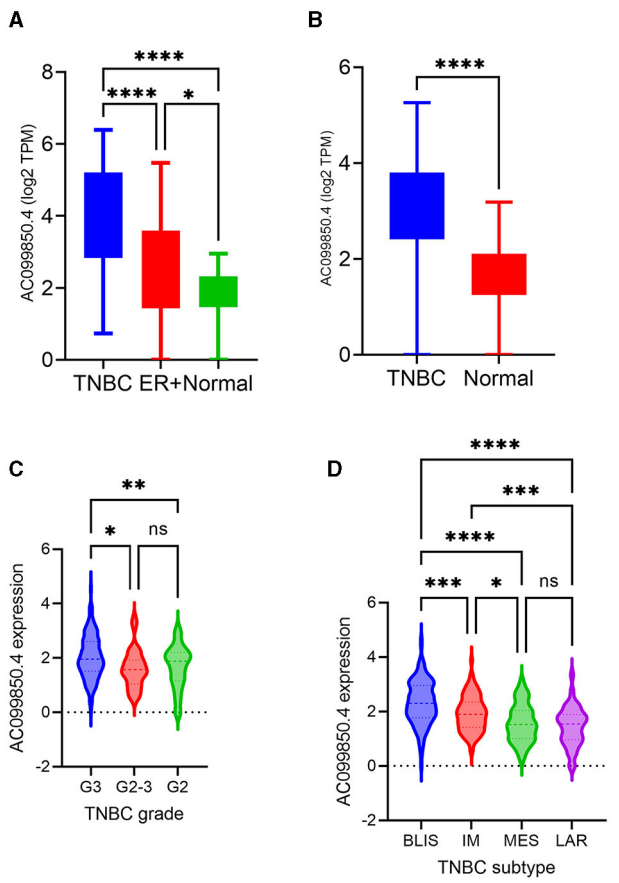


FIGURE 2
AC099850.4 expression correlated with advanced clinicopathological features of TNBC. **(A)** Box plots depicting the expression of AC099850.4 in TNBC (n = 42), ER^+ BC ($n = 42$), and normal breast tissue ($n = 56$) from the PRJNA251383. **(B)** Expression of AC099850.4 in the validation cohort (TNBC = 360 vs. normal = 88) from PRJNA486023 **(C)** mRNA-based classification **(D)** TNBC. BLIS: basal-like immunosuppressed, MES: mesenchymal, IM: immunomodulatory, LAR: luminal androgen receptor. $^*p < 0.05$, $^{**}p < 0.005$, $^{***}p < 0.0005$ and $^{****}p < 0.00005$. Expression of AC099850.4 as a function of tumor grade.

Ingenuity pathway analysis of differentially expressed genes in AC099850.4^{high} vs. AC099850.4^{low} TNBC

We subsequently used ingenuity pathway analysis to provide a better understanding of the enriched canonical, upstream regulator, and disease and function categories in AC099850.4^{high} TNBC. Canonical enrichment analysis identified activation of the kinetochore metaphase signaling pathway, pyridoxal 5'-phosphate salvage pathway, and salvage pathways of pyrimidine ribonucleotides in AC099850.4^{high} TNBC (Supplementary Table S2). Disease and function analysis identified enrichment in cell proliferation, cell movement, migration of cells, invasion of cells, cell viability, and colony formation (Figure 5A, Supplementary Table S3). Upstream regulator analysis identified enrichment in networks with predicted activation state of CKAP2L, FOXM1, RABL6, PCLAF, MITF, FOXO1, AREG, H2AZ1, E2F3, ESR1, RARA, ZNF768, KRAS, HNF1A-AS1, OGT, YAP1, KDM1A, and MYBL2 (Figure 5B, Supplementary Table S4). In contrary, TP53, NUPR1, TRPS1, CDKN1A, CTLA4, AR, KDM5B, ARID1A, ATF3, and PDCD1 were suppressed (Figure 5C, Supplementary Table S4). Taken together, our data suggested a strong correlation between AC099850.4 expression and mitotic cell cycle in clinical tumor specimens from TNBC patients.

AC099850.4 is an unfavorable prognostic biomarker for TNBC relapse-free short-term survival

We subsequently sought to assess the prognostic value of AC099850.4 in relation to RFS in TNBC. In that regard, we divided the 360 TNBC cohorts into AC099850.4^{high} and AC099850.4^{low} based on median AC099850.4 expression and performed the Kaplan–Meyer survival analysis. Interestingly, AC099850.4 expressed had a modest correlation with RFS in the long term (*log-rank* p-value = 0.4, Figure 6A). However, when we assessed the ability of AC099850.4 to predict short-term RFS (24 months), the high expression of AC099850.4 correlated with a worse prognosis (*log-rank* p-value = 0.01, Figure 6B). Those data highlighted a role for AC099850.4 as an unfavorable prognostic biomarker for short-term RFS.

Discussion

Understanding the biological roles of various lncRNAs has contributed to our knowledge of the functions of this class of epigenetic regulators in cancer. In the current study, we characterized the lncRNA transcriptome of TNBC and ER^+ breast cancers and identified 57 lncRNAs that were upregulated in TNBC vs. ER^+ and in TNBC vs. NT, but not in ER^+ vs. NT, suggesting their restricted expression in TNBC. Of particular interest, we conducted a comprehensive investigation on the expression AC099850.4 in TNBC. Interestingly, the highest expression of AC099850.4 was observed in TNBC patients with advanced tumor grade and in the BLIS subtype, which is known to have the worst prognosis among different TNBC subtypes (18). Investigating the expression of AC099850.4 in a larger cohort of TNBC (n = 360) correlated higher expression of AC099850.4 and enriched functional categories indicative of cellular proliferation and mitosis.

More in-depth computational analyses using IPA revealed activation of several functional categories in AC099850.4^{high} TNBC, including the canonical kinetochore metaphase signaling pathway, pyridoxal 5'-phosphate salvage pathway, and salvage pathways of pyrimidine ribonucleotides. Additionally, upstream regulator analysis predicted activation of CKAP2L, FOXM1, RABL6, PCLAF, and MITF and suppression of TP53, NUPR1, TRPS1, and CDKN1A in AC099850.4^{high} TNBC. Nonetheless, our data highlighted AC099850.4 as an unfavorable prognostic biomarker predicting short-term TRFS in TNBC. In agreement with our data, AC099850.4 was recently identified among 8 lncRNA biomarker panels in head and neck squamous cell carcinoma

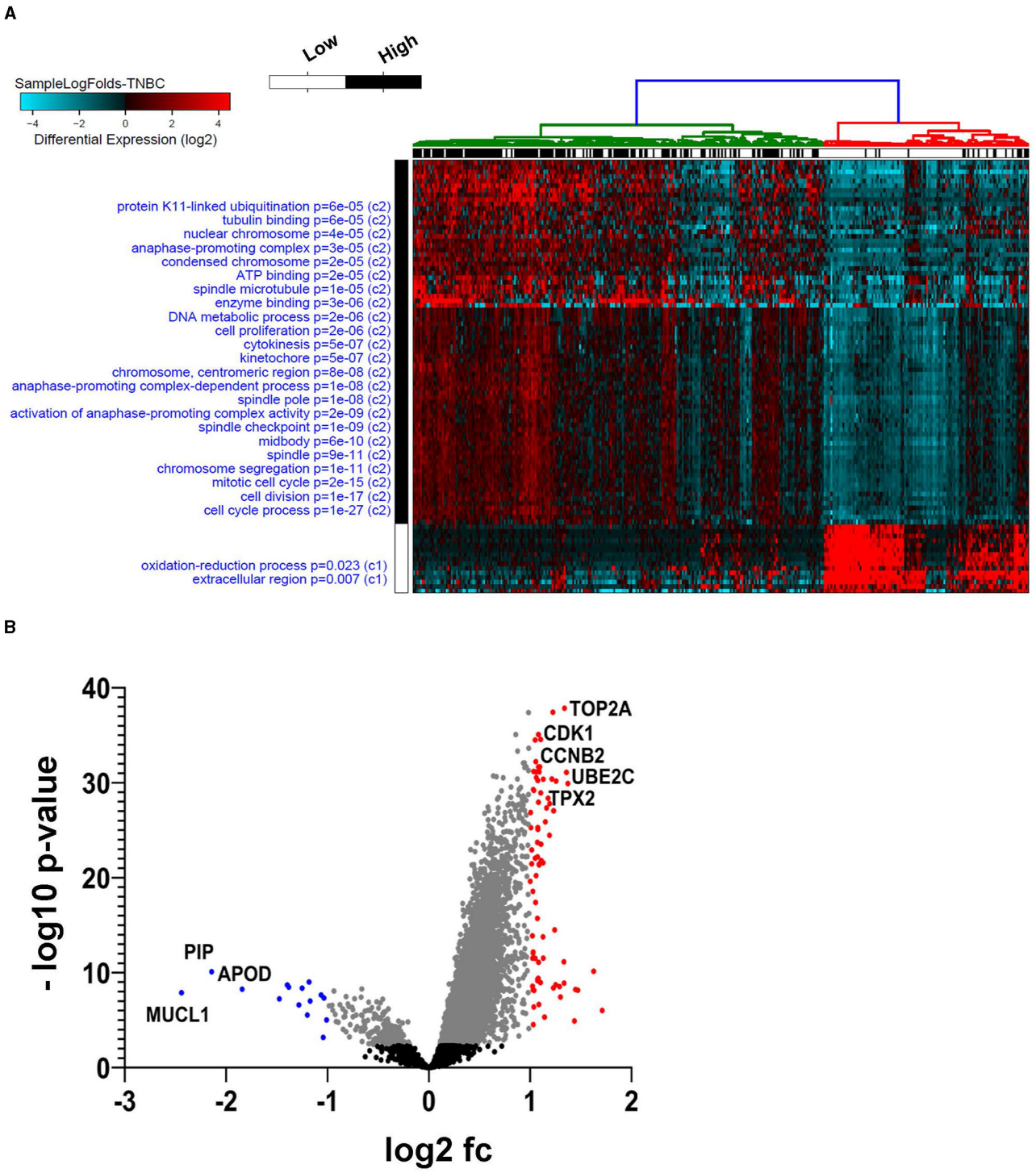


FIGURE 3
Functional enrichment in AC099850.4^{high} vs. AC099850.4^{low} TNBC. Expression data from 360 TNBC were grouped into "AC099850.4^{high}" and "AC099850.4^{low}" according to AC099850.4 median expression and were subjected to differential expression analysis. **(A)** Heatmap depicting the clustering of the AC099850.4^{high} vs. AC099850.4^{low} TNBC with the enriched gene ontology (GO) categories indicated on the left side and the corresponding enrichment *p*-value. **(B)** Volcano plot depicting the upregulated (red) and downregulated (blue) genes in AC099850.4^{high} vs. AC099850.4^{low} TNBC.

(19). Similarly, the elevated expression of AC099850.4, an m6A-related lncRNA, was reported in patients with oral squamous cell carcinoma (20), and the elevated expression of AC099850.4 was also correlated with worse survival in lung cancer (21). Recently, AC099850.4 was reported to be highly expressed and correlated with a worse prognosis in non-small cell lung cancer

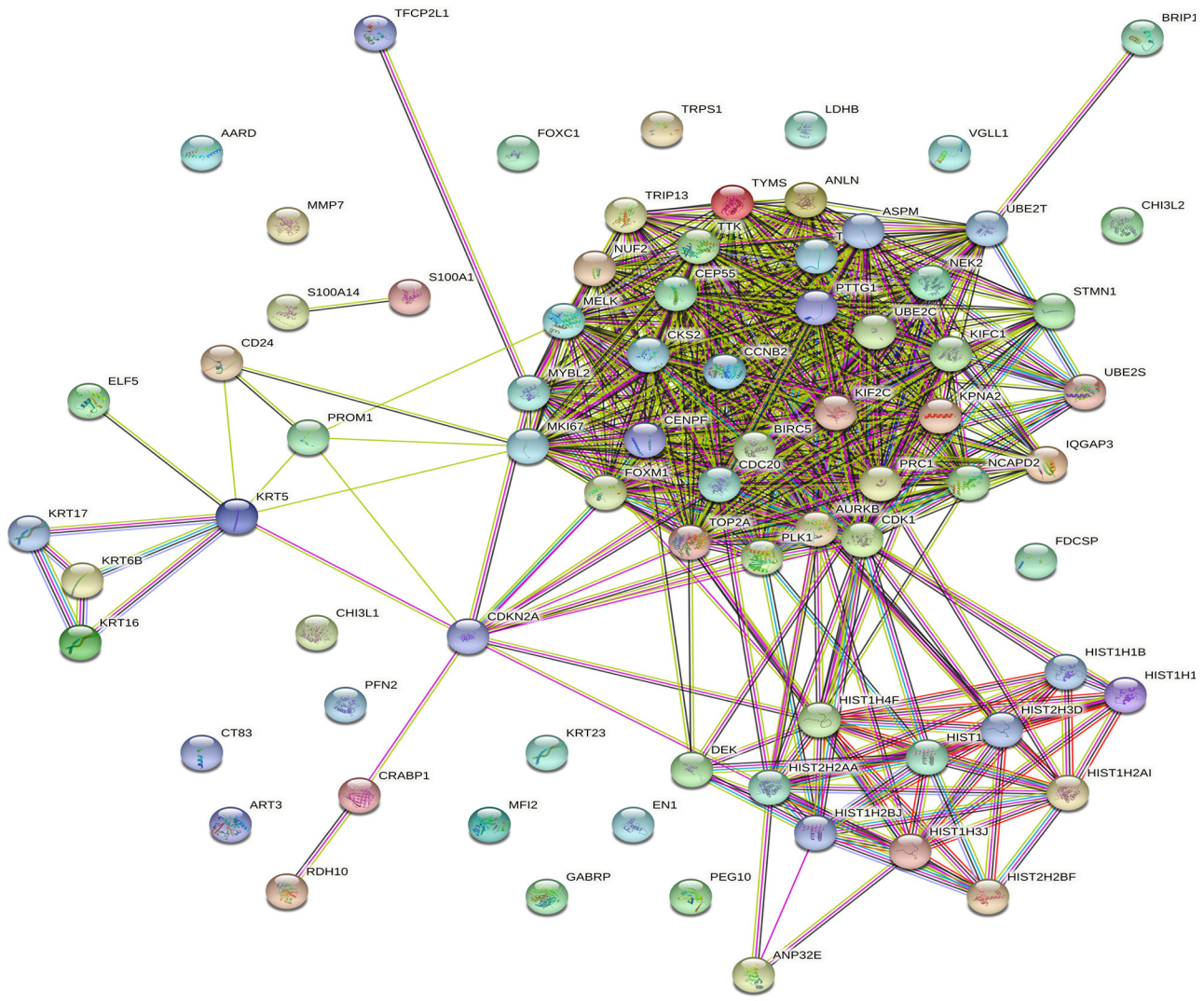


FIGURE 4
Protein-protein interaction (PPI) network analysis of upregulated genes in AC099850.4^{high} vs. AC099850.4^{low} TNBC. PPI network based on STRING analysis of upregulated genes in AC099850.4^{high} vs. AC099850.4^{low}. Network statistics: number of nodes: 76, number of edges: 585, expected number of edges: 79, average node degree: 15.4, avg. local clustering coefficient: 0.647, PPI enrichment p-value: $< 1.0 \times 10^{-16}$.

(22). Similarly, a recent study on hepatocellular carcinoma (HCC), which included 374 HCC and 160 non-HCC samples, identified five immune-related lncRNA prognostic panels, including AC099850.3. Silencing of AC099850.3 inhibited HCC cell proliferation and migration and led to significant inhibition of PLK1, TTK, CDK1, and BULB1 cell cycle molecules and CD155 and PDL1 immune receptors (23). Numerous recent studies revealed intriguing aspects of AC099850.4 as immuno-autophagy-related lncRNA (24), epithelial-mesenchymal transition-related lncRNA (25), and cancer cell stemness-associated lncRNA (26) in HCC. Those reports further support an oncogenic role for AC099850.4 in various human cancers, which remains to be validated in TNBC.

While several studies implicated AC099850.4 in various other cancer types, our data are the first to implicate this lncRNA in TNBC prognosis. Our data suggest the potential use of AC099850.4 as a prognostic biomarker and therapeutic target in TNBC, which warrants further investigation.

Conclusion

Our data are the first to identify AC099850.4 as a novel prognostic biomarker for TNBC, correlating with advanced disease stage and patient survival.

Limitations of the study

Our data provide solid evidence implicating AC099850.4 as a prognostic biomarker in TNBC. One limitation of the current study is that the cohort we analyzed has only ER$^+$ and TNBC, but none of the patients were HER2$^+$; hence, the expression of AC099850.4 in HER2$^+$ BC remains to be assessed. Although our study was initially based on patients' transcriptomic data, the potential to utilize this lncRNA for patient prognosis remains to be validated in multiple TNBC cohorts. The functional

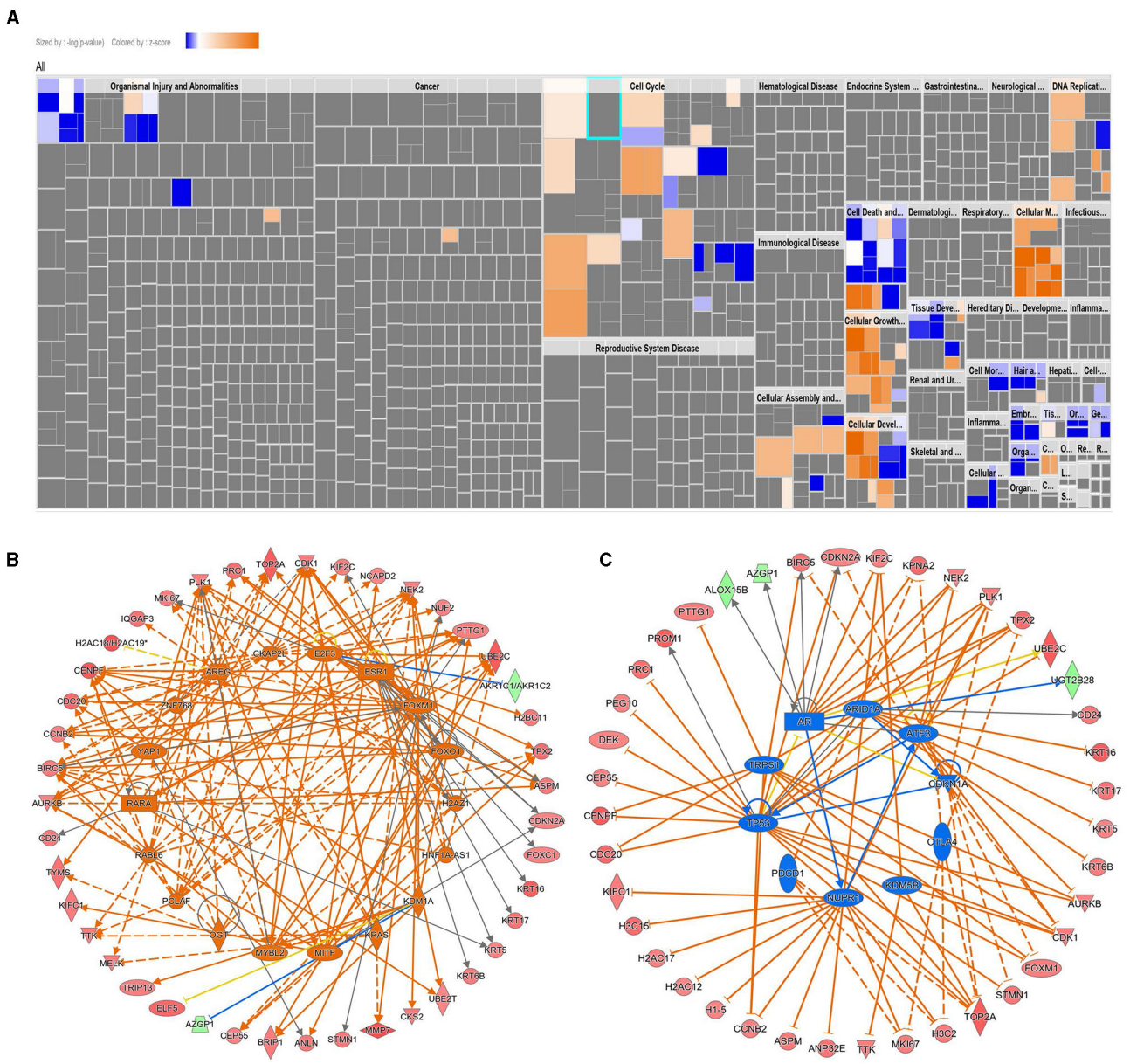


FIGURE 5
Ingenuity pathway analysis of differentially expressed genes in AC099850.4^{high} vs. AC099850.4^{low} TNBC. **(A)** Tree map (hierarchical heatmap) depicting affected functional categories based on differentially expressed genes in AC099850.4^{high} vs. AC099850.4^{low} where the major boxes represent a category of diseases and functions. Upstream regulator analysis depicting activated **(B)** and inhibited **(C)** networks in AC099850.4^{high} vs. AC099850.4^{low} TNBC.

consequences of AC099850.4 depletion in TNBC cell models remain to be validated *in vitro*, and the potential use of RNA-based therapeutics to target AC099850.4 systemically remains also to be addressed *in vivo*. Our data highlighted multiple enriched GO and networks in AC099850.4^{high} vs. AC099850.4^{low} TNBC; however, the exact mechanism by which AC099850.4 exerts its biological functions and its interacting protein partners remains to be identified using biochemical approaches, such as comprehensive identification of RNA-binding proteins by mass spectrometry, ChIRP-MS (27).

Materials and methods

RNA-Seq data analysis and bioinformatics

Raw RNA sequencing data were retrieved from the sequence read archive (SRA) database under accession no. PRJNA251383, consisting of 42 TNBC, 42 ER^{+}HER2^{-}, and 56 normal breast tissue samples. The Kallisto index was constructed by creating a de Bruijn graph employing the GENCODE release (V33) reference transcriptome and 31 length k-mer. FASTQ files were subsequently

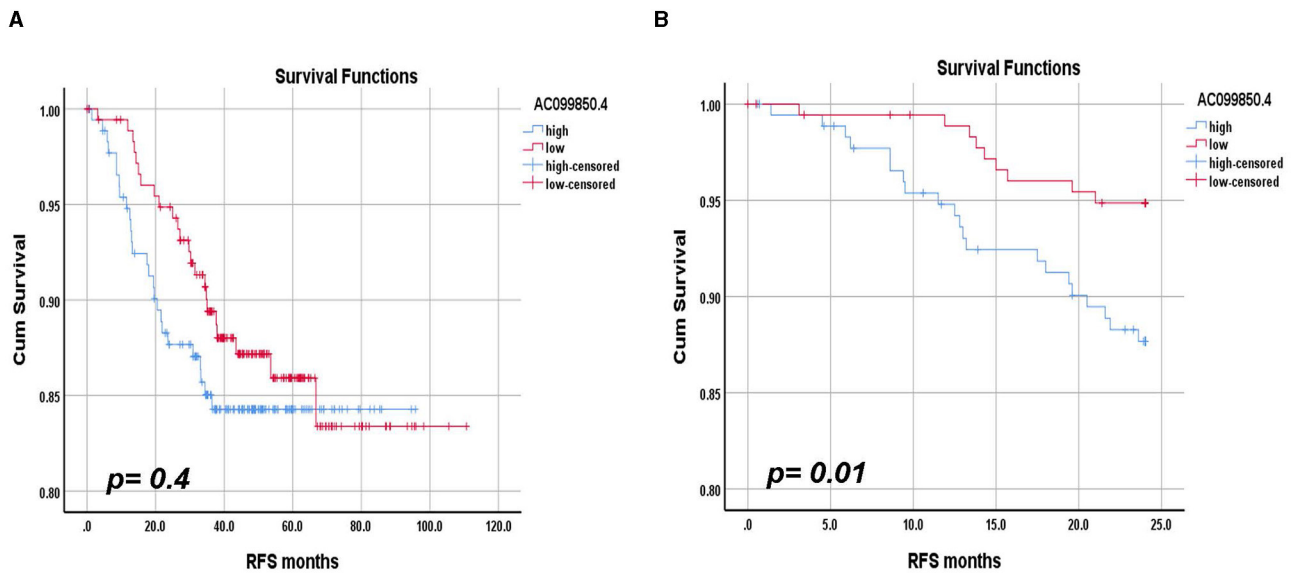


FIGURE 6
Relapse-free survival (RFS) analysis according to AC099850.4 expression. **(A)** Long-term RFS analysis in a cohort of 360 TNBC based on median AC099850.4 expression. **(B)** Short-term RFS analysis in a cohort of 360 TNBC based on median AC099850.4 expression. The log-rank test was used to compare groups.

pseudo-aligned to the generated index using KALLISTO 0.4.2.1, as previously described (3, 28). Normalization (TPM, transcript per million) was conducted using KALLISTO 0.4.2.1. A detailed description of the study subjects can be found in Ref. (29). Normalized expression data (TPM) were sequentially imported into AltAnalyze v.2.1.3 software for differential expression and PCA analysis using 2.0-fold change and adjusted cut-off *p*-value of <0.05 (30). Low abundant transcripts (<1.0 TPM raw expression value) were excluded from the analysis. The Benjamini–Hochberg method was used to adjust for the false discovery rate (FDR). The marker finder prediction was carried out as previously explained. PRJNA486023 (360 TNBC and 88 normal samples) was retrieved from the SRA databases using the SRA toolkit v2.9.2 as previously described (31, 32) and was mapped to GENCODE release (v33) as mentioned above and was used to confirm our findings. Detailed information on the study subjects in this validation cohort can be found in Jiang et al. (33).

Protein-protein interaction and KEGG network analysis

Upregulated genes in AC099850.4^{high} TNBC (n = 180) were subject to PPI network analysis using the STRING (STRING v10.5) database to illustrate the interacting genes/proteins based on knowledge and predication as described before (34). KEGG pathway analysis was conducted using DAVID as described earlier (35).

Gene set enrichment and modeling of gene interactions networks

Upregulated genes in AC099850.4^{high} were imported into the Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems; http://www.ingenuity.com/) and were subjected to functional annotations and regulatory network analysis using upstream regulator analysis (URA), downstream effects analysis (DEA), mechanistic network (MN) and causal network analysis (CNA) prediction algorithm. IPA uses precision to predict functional regulatory networks from gene expression data and provides a significance score for each network according to the fit of the network to the set of focus genes in the database. The *p*-value is the negative log of P and represents the possibility of focus genes in the network being found together by chance.

Survival and statistical analysis

The Kaplan–Meier survival analysis and plotting were conducted using IBM SPSS version 26 software. For survival analysis, patients were grouped into high or low based on the corresponding lncRNA median expression. The log-rank test was used to compare the outcome between expression groups. GraphPad Prism 9.0 software (San Diego, CA, USA) was used to compare the lncRNA expression as a function of tumor grade and LN status. An unpaired two-tailed *t*-test was used to compare two groups, while a one-way ANOVA was used to compare multiple groups. The Benjamini–Hochberg method was used to adjust for the false discovery rate (FDR). The *p*-value of < 0.05 was considered statistically significant.

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

RV performed the experiments and manuscript writing. NA obtained funding, concept, design, data analysis, and finalized the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the Qatar Biomedical Research Institute (grant no. QB13) for NA.

Acknowledgments

Open Access funding provided by the Qatar National Library.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1149860/full#supplementary-material

References

1. Zardavas D, Irrthum A, Swanton C, Piccart M. Clinical management of breast cancer heterogeneity. *Nat Rev Clin Oncol.* (2015) 12:381–94. doi: 10.1038/nrclinonc.2015.73

2. Liu YR, Jiang YZ, Xu XE, Yu KD, Jin X, Hu X. Comprehensive transcriptome analysis identifies novel molecular subtypes and subtype-specific RNAs of triple-negative breast cancer. *Breast Cancer Res.* (2016) 18:33. doi: 10.1186/s13058-016-0690-8

3. Shaath H, Elango R, Alajez NM. Molecular classification of breast cancer utilizing long non-coding RNA (lncRNA) transcriptomes identifies novel diagnostic lncRNA Panel for triple-negative breast cancer. *Cancers.* (2021) 13:350. doi: 10.3390/cancers13215350

4. Iorio MV, Casalini P, Tagliabue E, Menard S, Croce CM. MicroRNA profiling as a tool to understand prognosis, therapy response and resistance in breast cancer. *Eur J Cancer.* (2008) 44:2753–9. doi: 10.1016/j.ejca.2008.09.037

5. Vishnubalaji R, Shaath H, Elango R, Alajez NM. Noncoding RNAs as potential mediators of resistance to cancer immunotherapy. *Semin Cancer Biol.* (2020) 65:65–79. doi: 10.1016/j.semcancer.2019.11.006

6. Huang Y, Wang X, Zheng Y, Chen W, Zheng Y, Li G. Construction of an mRNA-miRNA-lncRNA network prognostic for triple-negative breast cancer. *Aging.* (2021) 13:1153. doi: 10.18632/aging.202254

7. Song X, Liu Z, Yu Z. LncRNA NEF is downregulated in triple negative breast cancer and correlated with poor prognosis. *Acta Biochim Biophys Sin.* (2019) 51:386–92. doi: 10.1093/abbs/gmz021

8. Jin H, Du W, Huang W, Yan J, Tang Q, Chen Y, et al. lncRNA and breast cancer: progress from identifying mechanisms to challenges and opportunities of clinical treatment. *Mol Ther Nucleic Acids.* (2021) 25:613–37. doi: 10.1016/j.omtn.2021.08.005

9. Liu L, Zhang Y, Lu J. The roles of long noncoding RNAs in breast cancer metastasis. *Cell Death Dis.* (2020) 11:749. doi: 10.1038/s41419-020-02954-4

10. Shaath H, Vishnubalaji R, Elango R, Kardousha A, Islam Z, Qureshi R. Long non-coding RNA and RNA-binding protein interactions in cancer: experimental and machine learning approaches. *Semin Cancer Biol.* (2022) 86:325–45. doi: 10.1016/j.semcancer.2022.05.013

11. Winkle M, El-Daly SM, Fabbri M, Calin GA. Noncoding RNA therapeutics - challenges and potential solutions. *Nat Rev Drug Discov.* (2021) 20:629–51. doi: 10.1038/s41573-021-00219-z

12. Jiang MC, Ni JJ, Cui WY, Wang BY, Zhuo W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am J Cancer Res.* (2019) 9:1354–66.

13. Venkatesh J, Wasson MD, Brown JM, Fernando W, Marcato P. LncRNA-miRNA axes in breast cancer: novel points of interaction for strategic attack. *Cancer Lett.* (2021) 509:81–8. doi: 10.1016/j.canlet.2021.04.002

14. Wang J, Katsaros D, Biglia N, Fu Y, Benedetto C, Loo L. LncRNA ZNF582-AS1 expression and methylation in breast cancer and its biological and clinical implications. *Cancers.* (2022) 14:788. doi: 10.3390/cancers14112788

15. Bjorklund SS, Aure MR, Hakkinen J, Vallon-Christersson J, Kumar S, Evensen KB. Subtype and cell type specific expression of lncRNAs provide insight into breast cancer. *Commun Biol.* (2022) 5:834. doi: 10.1038/s42003-022-03559-7

16. Shaath H, Vishnubalaji R, Elango R, Khattak S, Alajez NM. Single-cell long noncoding RNA (lncRNA) transcriptome implicates MALAT1 in triple-negative breast cancer (TNBC) resistance to neoadjuvant chemotherapy. *Cell Death Discov.* (2021) 7:23. doi: 10.1038/s41420-020-00383-y

17. Vishnubalaji R, Elango R, Alajez NM. LncRNA-based classification of triple negative breast cancer revealed inherent tumor heterogeneity and vulnerabilities. *Non-Coding RNA.* (2022) 8:44. doi: 10.3390/ncrna8040044

18. Yin L, Duan JJ, Bian XW, Yu SC. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Res.* (2020) 22:61. doi: 10.1186/s13058-020-01296-5

19. Mao R, Chen Y, Xiong L, Liu Y, Zhang T. Identification of a nomogram based on an 8-lncRNA signature as a novel diagnostic biomarker for head and neck squamous cell carcinoma. *Aging.* (2020) 12:20778–800. doi: 10.18632/aging.104014

20. Yang Q, Cheng C, Zhu R, Guo F, Lai R, Liu X. N6-methyladenosine-related long noncoding RNAs model for predicting prognosis in oral squamous cell carcinoma: association with immune cell infiltration and tumor metastasis. *Oral Oncol.* (2022) 127:105771. doi: 10.1016/j.oraloncology.2022.105771

21. Liu X, Zuo X, Ma L, Wang Q, Zhu L, Li L. Integrated analysis of the m6A-related lncRNA identified lncRNA ABALON/miR-139-3p/NOB1 axis was involved in the occurrence of lung cancer. *Cancer Manag Res.* (2021) 13:8707–22. doi: 10.2147/CMAR.S339032

22. Zhou J, Zhang M, Dong H, Wang M, Cheng Y, Wang S. Comprehensive analysis of acetylation-related lncRNAs and identified ac099850.3 as prognostic biomarker in non-small cell lung cancer. *J Oncol.* (2021) 21:4405697. doi: 10.1155/2021/4405697

23. Wu F, Wei H, Liu G, Zhang Y. Bioinformatics profiling of five immune-related lncRNAs for a prognostic model of hepatocellular carcinoma. *Front Oncol.* (2021) 11:667904. doi: 10.3389/fonc.2021.667904

24. Wang Y, Ge F, Sharma A, Rudan O, Setiawan MF, Gonzalez-Carmona MA. Gonzalez-carmona, immunoautophagy-related long

noncoding RNA (IAR-lncRNA) signature predicts survival in hepatocellular carcinoma. *Biology*. (2021) 10:301. doi: 10.3390/biology10121301

25. Xu BH, Jiang JH, Luo T, Jiang ZJ, Liu XY, Li LQ. Signature of prognostic epithelial-mesenchymal transition related long noncoding RNAs (ERLs) in hepatocellular carcinoma. *Medicine*. (2021) 100:e26762. doi: 10.1097/MD.0000000000026762

26. Zhang Q, Cheng M, Fan Z, Jin Q, Cao P, Zhou G. Identification of Cancer cell stemness-associated long noncoding RNAs for predicting prognosis of patients with hepatocellular carcinoma. *DNA Cell Biol.* (2021) 40:1087–100. doi: 10.1089/dna.2021.0282

27. Chu C, Zhang QC, Rocha STda, Flynn RA, Bharadwaj M, Calabrese JM. Systematic discovery of Xist RNA binding proteins. *Cell.* (2015) 161:404–16. doi: 10.1016/j.cell.2015.03.025

28. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol.* (2016) 34:525–7. doi: 10.1038/nbt.3519

29. Varley KE, Gertz J, Roberts BS, Davis NS, Bowling KM, Kirby MK. Recurrent read-through fusion transcripts in breast cancer. *Breast Cancer Res Treat.* (2014) 146:287–97. doi: 10.1007/s10549-014-3019-2

30. Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin BR, Albrecht M. Altanalyze and domaingraph: analyzing and visualizing exon expression data. *Nucleic Acids Res.* (2010) 38:W755–62. doi: 10.1093/nar/gkq405

31. Leinonen R, Sugawara H, Shumway M. The sequence read archive. *Nucleic Acids Res.* (2011) 39:D19–21. doi: 10.1093/nar/gkq1019

32. Elango R, Vishnubalaji R, Shaath H, Alajez NM. Molecular subtyping and functional validation of TTK, TPX2, UBE2C, and LRP8 in sensitivity of TNBC to paclitaxel. *Mol Ther Methods Clin Dev.* (2021) 20:601–14. doi: 10.1016/j.omtm.2021.01.013

33. Jiang YZ, Ma D, Suo C, Shi J, Xue M, Hu X. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell.* (2019) 35:428–40. doi: 10.1016/j.ccell.2019.02.001

34. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J. 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

35. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res.* (2022) 50:W221–34. doi: 10.1093/nar/gkac194

Frontiers in Medicine

Translating medical research and innovation into improved patient care

A multidisciplinary journal which advances our medical knowledge. It supports the translation of scientific advances into new therapies and diagnostic tools that will improve patient care.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact