narra j

Original Article

Comprehensive in silico analysis of single nucleotide polymorphism and molecular dynamics simulation of human GATA6 protein in ventricular septal defect

Taufiq Hidayat[1,2,3], Irwanto Irwanto[4,5*], Ali Rohman[6], Afrizal AA. Muhyiddin[4,5], Safira NA. Putri[7], Dedy B. Kurniawan[7], Mokhamad FR. Syaban[7], Theakirana Firdaus[7], Mahrus A. Rahman[2,3] and I KA. Utamayasa[2,3]

[1]Doctoral Program in Medical Science, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia; [2]Division of Pediatric Cardiology, Department of Pediatric, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia; [3]Division of Pediatric Cardiology, Department of Pediatric, Dr. Soetomo General Academic Hospital, Surabaya, Indonesia; [4]Department of Pediatric, Faculty of Medicine, Universitas Airlangga, Surabaya, Indonesia; [5]Department of Pediatric, Dr. Soetomo General Academic Hospital, Surabaya, Indonesia; [6]Department of Chemistry, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia; [7]Faculty of Medicine, Universitas Brawijaya, Malang, Indonesia

*Corresponding author: irwanto@fk.unair.ac.id

Abstract

Congenital heart disease (CHD) represents nearly one-third of congenital birth defects annually, with ventricular septal defect (VSD) being the most common type. The aim of this study was to explore the role of specific GATA binding protein 6 gene (*GATA6*) mutations as a potential etiological factor in the development of VSD through an in silico approach. Data were collected from the human gene databases: DisGeNET and GeneCards, with protein-protein interaction networks constructed via STRING and Cytoscape. Gene ontology and pathway enrichment analyses were conducted using DAVID, with data analysis in R with significance set at FDR $p<0.05$. Target single nucleotide polymorphisms (SNPs) of *GATA6* were obtained from NCBI dbSNP, and non-synonymous single nucleotide polymorphism (nsSNP) effects were predicted using SIFT, PolyPhen-2, I-Mutant 2.0, Fathmm, MutPred 2.0, SNP&GO, and PON-P2. Conserved regions of *GATA6* were analyzed using ConSurf, with functional classification, variant conservation, and stability changes evaluated in Google Colab. Multiple sequence alignment was performed using ClustalW. Mutation modeling and molecular dynamics analysis, using GROMACS, revealed that among 87 intersecting genes, 16 proteins were interconnected with *GATA6*, showing a centrality value of 0.4378. Gene ontology analysis highlighted atrioventricular canal development, protein-DNA complexes, and transcription factor regulation as key processes for cardiac development, especially in the ventricular septum. NsSNP and molecular dynamics analyses identified rs387906818 and rs387906820 as having the highest pathogenic potential for VSD due to amino acid structural changes.

Keywords: VSD, GATA6, in silico, nsSNP, molecular dynamic simulation

Introduction

Congenital heart disease (CHD) represents a significant global health challenge, accounting for nearly one-third of all congenital birth defects annually [1]. In 2017, the global prevalence of CHD in newborns was estimated at approximately 1.8 cases per 100 live births, with ventricular septal

https://narraj.org/

Received: August 21, 2024 - Accepted: November 25, 2024 - Published online: December 10, 2024

defect (VSD) being the most common congenital cardiac anomaly [2,3]. In Indonesia, data from Sardjito Hospital, a tertiary referral center in Yogyakarta and Central Java, indicated a 30% prevalence of VSD among 650 new CHD patients [4,5]. Large VSD defects can lead to severe complications, and children with CHD are at increased risk for developmental problems, underscoring VSD as a high-burden health issue in the country [6,7].

Advances in the study of normal cardiac development have facilitated the identification of key structural genes, transcriptional regulators, and signaling molecules involved in heart formation [8]. One such gene, the GATA binding protein 6 gene (*GATA6*), belongs to the GATA family of transcription factors and is expressed in various mesoderm and endoderm-derived tissues [9]. *GATA6* plays a critical role in heart development, including cellular differentiation and organogenesis during vertebrate development, particularly in the embryonic heart [10]. Interactions between *GATA6* and cardiac transcription factors, conserved in the atrial and ventricular myocardium, have been associated with cardiac septal defects [11]. Mutations in the *GATA6* can include missense mutations, deletions, and copy number variants [12]. These mutations have been found in both familial and isolated cases of VSD, suggesting a potential role for *GATA6* in the pathogenesis of this congenital heart defect [13].

Moreover, *GATA6* mutations have been associated with other congenital heart defects, such as atrial septal defects, atrioventricular septal defects, persistent truncus arteriosus, and tetralogy of Fallot [14]. A previous study reported two novel DNA sequence variants (DSVs), g.22169190A>T and g.22169311C>G, in VSD patients, which were absent in control subjects. The study hypothesized that *GATA6* functions in a dose-dependent manner, where alterations in GATA6 protein expression—potentially caused by changes in DNA sequence in regulatory regions—may contribute to the development of CHD [13]. In another study, nine variations of missense and nonsense mutations in the *GATA6* were tested, which led to mild congenital heart diseases such as bicuspid aortic valve, septal defects, and unspecified heart surgery [15]. These clinical tests were followed by in silico and in vivo analyses to determine which mutation variations had the highest pathogenicity. The results showed that two mutation variations in *GATA6* played an important role in the occurrence of CHDs: c.466A>G (Gln120Ter) and c.1271G>T (Ser424Ile) [15]. Despite in silico, in vivo, and clinical testing in the previous studies have explained the types of mutation variations in the *GATA6* and their impact on CHDs, there is still limited understanding of the molecular mechanisms, non-synonymous single nucleotide polymorphism (nsSNP), molecular dynamics simulation modeling, and the role of the *GATA6* gene and protein in the molecular mechanism of VSD as the most common type of congenital heart defect [12,13].

The aim of this study was to explore the role of specific *GATA6* mutations as a potential etiological factor in the development of VSD through an in silico approach. By employing bioinformatics and nsSNP analysis, further insights into the molecular mechanisms underlying VSD and the potential pathogenic contributions of *GATA6* mutations were explored. The findings may enhance understanding of the genetic basis of VSD and offer potential implications for diagnosis, prognosis, and therapeutic approaches.

Methods

Study design and study flow overview

This study used an in silico approach. Initially, data collection and screening were conducted based on two gene databases related to VSD using GeneCards and DisGeNET. A protein-protein interaction network analysis was then carried out based on the screening results from both databases, following inclusion criteria (highest closeness centrality, $p<0.05$ in Gene Ontology, and Gene Disease association score >0.05). The results of this process were followed by the collection and screening of nsSNP in the NCBI (dbSNP) and UniProt databases. Screening was conducted based on pathogenicity and disease criteria, deleterious effects on protein function, molecular mechanisms and protein stability. These results were followed by protein modeling and molecular dynamics simulations on selected nsSNPs, with analysis based on conserved protein domains, polarity and functional changes, mutation stability, and functional class. The detailed study flow diagram is presented in **Figure 1**.

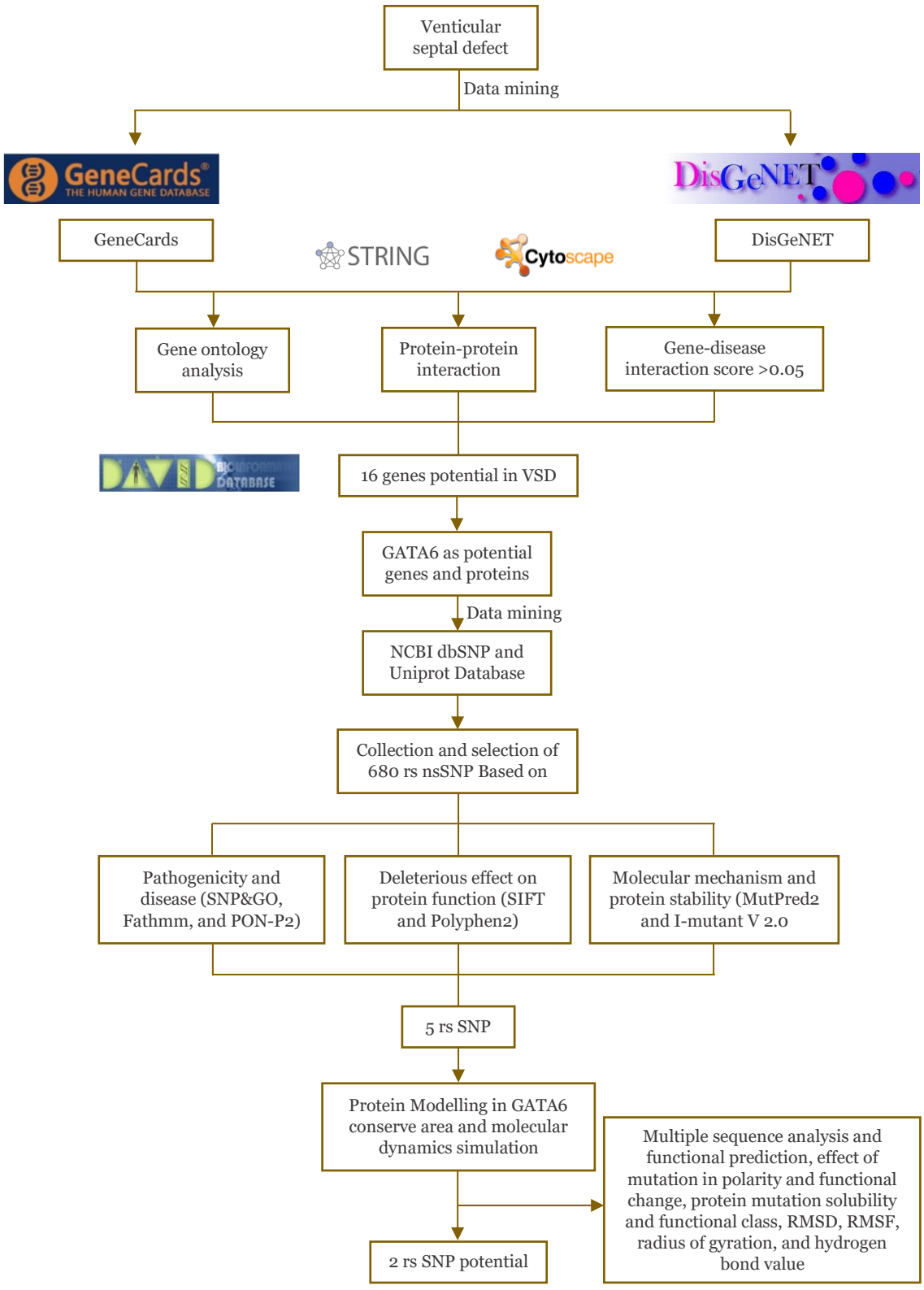


Figure 1. Flow diagram and research design.

Target data collection and selection

Data collection and gene selection related to VSD were conducted using the DisGeNET (https://disgenet.com/) and GeneCards (https://www.genecards.org/) databases [16,17]. A Venn diagram tool (https://bioinformatics.psb.ugent.be/webtools/Venn/) was utilized to identify intersecting genes from these sources.

Construction of protein-protein interaction network

Genes associated with VSD were analyzed for protein-protein interactions using the STRING web server (https://string-db.org) [18]. A total of 87 interacting proteins were identified and subsequently analyzed using Cytoscape v3.10.0 (Cytoscape Team, San Fransisco, USA) to explore the protein-protein interaction network. Cluster analysis was performed using the Cytocluster application (Cytoscape Team, San Fransisco, CA, USA) evaluating parameters including density, quality, and *p*-value [19].

Gene ontology analysis

Gene ontology analysis was conducted to identify the biological processes, molecular functions, and cellular components associated with the protein-protein interactions. The protein-protein interaction network generated from Cytoscape was further analyzed using the DAVID web server (https://david.ncifcrf.gov/home.jsp) [20]. The gene ontology results demonstrated statistical significance, with a false discovery rate (FDR) *p*-value of less than 0.05.

GATA6 protein 3D structure and domain modelling

GATA6 protein modeling was performed using the AlphaFold 3 web server (https://alphafoldserver.com/) [21]. The GATA6 protein sequence, obtained from the UniProt database (https://www.uniprot.org/) under protein code Q92908, was entered into the AlphaFold 3 web server, and the resulting model was downloaded in .pdb format [22]. The GATA6 protein domain was identified from relevant literature and case studies that examine the impact of mutations and missense SNPs on the occurrence of CHDs, particularly VSD.

Single nucleotide polymorphism (SNP) data collection and selection

Following the protein-protein interaction network analysis, an SNP analysis was conducted to evaluate gene-level changes in amino acid structure and their association with the incidence of VSD. The NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/snp/) was used to retrieve 13,808 reference sequences of *GATA6*, including 680 missense variants at the coding level. SNPs were selected based on their established clinical significance [23].

Identification and assessment of protein consequences in missense mutation

The selected SNP data were further evaluated through a functional analysis of the GATA6 protein structure using several bioinformatic tools, including SIFT (The Institute of Genomic Research, San Diego, USA), Polyphen-2 (Harvard University, Massachusetts, USA), I-Mutant v2.0 (BioFoID-University of Bologna, Bologna, Italy), MutPred2 (Northeastern University-Boston, University of Washington-Seattle, Indiana University-Bloomington, UC San Diego-San Diego, USA), SNP&GO (BioFoID-University of Bologna, Bologna, Italy), Fathmm (MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK), and PON-P2 (Lund University, Lund, Sweden). SIFT (https://sift.bii.a-star.edu.sg/) and Polyphen-2 (http://genetics.bwh.harvard.edu/pph2/) were used to assess missense mutations and predict their impact on GATA6 protein function and structure, employing a sequence homology and alignment-based approach to natural nsSNPs [24,25]. A SIFT score below 0.05 indicates a deleterious effect on protein function. Polyphen-2 provides three outcome categories: probably damaging (score >0.85), possibly damaging (score >0.15), and benign (score <0.15) [26].

I-Mutant 2.0 (https://folding.biofold.org/) and MutPred2 (http://mutpred.mutdb.org/) algorithms were employed to predict the stability and pathogenicity of the protein structure following mutations [27,28]. I-Mutant 2.0 predictions were performed under conditions of pH 7 and a temperature of 25°C. A ΔΔG (DDG) score less than 0 indicates a decrease in protein stability, while a score greater than 0 suggests increased stability. MutPred2, using a machine learning-based approach, integrates genetic and molecular data to generate statistical predictions for mutations, where a score above 0.5 indicates pathogenicity and a score below 0.5 indicates a benign mutation. A $p<0.05$ in Mendelian disease-related predictions suggests significant pathogenicity [29].

SNP&GO, Fathmm, and PON-P2 algorithms were used to predict the impact of amino acid mutations on disease occurrence, pathogenicity, and protein function, based on structural homology and gene ontology assessment. In SNP&GO (https://snps.biofold.org/snps-and-

go/snps-and-go.html), a prediction score greater than 0.5 indicates disease association, while a score below 0.5 indicates neutrality. The reliability index (RI) score ranges from 0 to 10, with the lowest scores (0–2) suggesting that the prediction requires scrutiny and further validation, while scores of 9–10 mean the most confident and likely to be biologically meaningful. Fathmm evaluates mutations based on their predicted effect on protein function, with a focus on the likelihood of significant damage [30]. PON-P2 categorizes mutations as pathogenic, benign, or unknown [31].

Multiple sequence alignment and GATA6 protein functional prediction analysis

Sequential alignment and protein functional prediction analysis were conducted to investigate the role of *GATA6* mutations in the pathogenesis of VSD. Consurf (https://consurf.tau.ac.il/consurf_index.php) analysis was utilized to identify conserved regions of the GATA6 protein across multiple species, revealing highly conserved areas that underscore their functional significance. Additionally, protein functional prediction analysis was performed to evaluate the potential impact of missense mutations within the *GATA6* gene. Conservation scores of amino acid residues were represented on an integer scale from 1 to 9, with a corresponding color scheme: magenta indicates the most variable positions (grade 1), white represents intermediately conserved positions (grade 5), and dark green denotes the most conserved positions (grade 9) [32].

A multiple sequence alignment of the GATA6 protein, in comparison with other GATA family domains and species, was carried out using the CLUSTALW web server (https://www.genome.jp/tools-bin/clustalw). The parameters for alignment included a CLUSTAL output with pairwise alignment using FAST/APPROXIMATE. GATA family members (GATA1–6) and GATA6 protein sequences were compared across species, including rat, mouse, pig, and chicken, to identify homologous regions relevant to the pathogenicity of VSD [33].

Effect on amino acid mutation in polarity and functional change

Changes in the DNA structure within the exon due to the presence of SNPs lead to mutations in the expressed amino acids. These amino acid changes result in alterations in the polarity and functionality of the protein structure. The evaluation of amino acid changes and functionality is based on the polarity properties of specific amino acids, such as: (1) non-polar (glycine (Gly), alanine (Ala), valine (Val), leucine (Leu), isoleucine (Ile), methionine (Met), phenylalanine (Phe), tryptophan (Trp), and proline (Pro)); (2) polar (serine (Ser), threonine (Thr), cysteine (Cys), tyrosine (Tyr), asparagine (Asn), and glutamine (Gln)); (3) basic polar (Lysine (Lys), arginine (Arg), and histidine (His)); and (4) acidic polar (aspartic acid (Asp) and glutamic acid (Glu)). The evaluation of functional changes is based on the properties of the side-chain structure, with seven being hydrophobic (glycine (Gly), alanine (Ala), valine (Val), leucine (Leu), isoleucine (Ile), phenylalanine (Phe), and proline (Pro)); nine being hydrophilic (serine (Ser), threonine (Thr), cysteine (Cys), aspartic acid (Asp), asparagine (Asn), glutamine (Gln), arginine (Arg), histidine (His), and glutamic acid (Glu)); and four being amphipathic (lysine (Lys), tryptophan (Trp), tyrosine (Tyr), and methionine (Met)).

Modelling of GATA6 protein and mutation in the conserved area

The structure of the GATA6 protein was constructed based on conserved regions identified from the protein sequence alignment results. The conserved region spanning amino acids 385–496 was modeled using AlphaFold 3. To create the GATA6 protein mutation model, Yasara Dynamics software (Yasara Bioscience, WHAT IF Foundation, Nijmegen, Netherlands) was employed with the Foldx menu program, allowing for the selection of specific amino acids for residue mutation. After building the mutant GATA6 protein, the resulting structure was downloaded in .pdb format [34]. Validation of the model was performed using the Procheck web server (https://saves.mbi.ucla.edu/) through Ramachandran plot analysis. A model is considered valid if more than 80% of the residues fall within the most favored regions [35].

Evaluation of GATA6 protein mutation solubility and functional class

Prediction of GATA6 protein mutation solubility was conducted to evaluate the impact of mutations on protein solubility and the potential for aggregation associated with low solubility

values. This analysis utilized the AggreScan 3D web server version 2.0 (https://biocomp.chem.uw.edu.pl/A3D2/). The parameters assessed included the average protein solubility score, with more negative values indicating normal solubility, and the energy difference between the wild type and mutated proteins, with more positive values suggesting a destabilizing mutation and more negative values indicating a stabilizing effect [36,37].

Molecular dynamics and visualization of variant/residue effect predictions

Molecular dynamics is a simulation technique employed to predict the movement of individual atoms in a protein or other molecular systems, grounded in fundamental physical principles of interatomic interactions. This simulation is performed under conditions that replicate biomolecular systems found in living organisms, such as proteins embedded in aqueous environments and lipid bilayers. Molecular dynamics simulation in this study utilized the WebGro web server (https://simlab.uams.edu/index.php) for a protein-in-water simulation. The GATA6 wild type protein served as the control, with comparisons made against GATA6 protein mutants R456C, N466H, A467T, T452A, and N466D to assess changes in protein stability and conformation under simulated biological conditions. The simulation employed the GROMOS96 43a1 force field, with a temperature of 310 K, the addition of 0.15 M NaCl, 5,000 steps of energy minimization, and a simulation duration of 50 ns. Parameters measured during this simulation included root mean square deviation (RMSD), root mean square fluctuation (RMSF), the number of hydrogen bonds formed, and the radius of gyration [35,38]. Furthermore, visualization of variant or residue effect prediction parameters—such as functional classification, variant conservation score, and variant stability change score—was conducted using the Google Colab web server (https://colab.research.google.com/github/KULL-Centre/_2024_cagiada-jonsson-func/blob/main/Download_predictions.ipynb).

Statistical analysis

Statistical test results were analyzed using the R programming environment (Frederick National Laboratory, Maryland, USA) on the DAVID web server (https://davidbioinformatics.nih.gov/tools.jsp). The gene ontology, including biological processes, cellular components, and molecular functions, was performed using Fisher's exact test between target genes and the total genome, with a false discovery rate (FDR) applied to the DAVID database. A significance threshold was set at $p<0.05$.

Results

Identification of potential genes and proteins in VSD

The results of the selection test for potential genes associated with VSD from two databases are presented in **Figure 2**. The results showed that from a total of 313 genes associated with VSD in the DisGeNET database and 282 genes associated with VSD in the GeneCards database, 87 potential genes were identified that may be linked to VSD. These 87 genes were then analyzed comprehensively further using in silico testing.

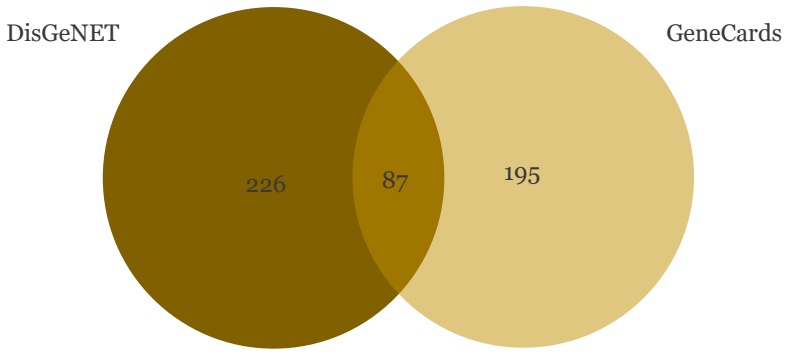


Figure 2. A total of 87 intersecting genes associated with ventricular septal defects (VSD) from two databases: DisGeNET and GeneCards.

Protein-protein interaction network of VSD

The 87 intersecting genes were analyzed using the STRING web server, which identified 87 interacting proteins (**Figure 3**). However, not all 87 proteins in this network interact directly or show a protein-protein interaction network related to the molecular mechanisms of ventricular septum formation. Therefore, further specific testing was required using Cytoscape software to identify protein-protein interaction networks that play a specific role in the occurrence of VSD. The results of analysis using Cytoscape identified 16 protein-protein interaction networks that specifically contribute to the occurrence of VSD, as presented in **Figure 4**. These results showed that GATA6 is at the center of the 16 protein-protein interaction networks, based on closeness centrality score analysis. Closeness centrality refers to how close a node is to all other nodes or how central it is within a network. The GATA6 protein has the highest closeness centrality score (**Table 1**). These results indicate that the GATA6 protein plays a central role in the protein-protein interaction network, and mutations in this protein may increase the risk or potential for the occurrence of VSD.

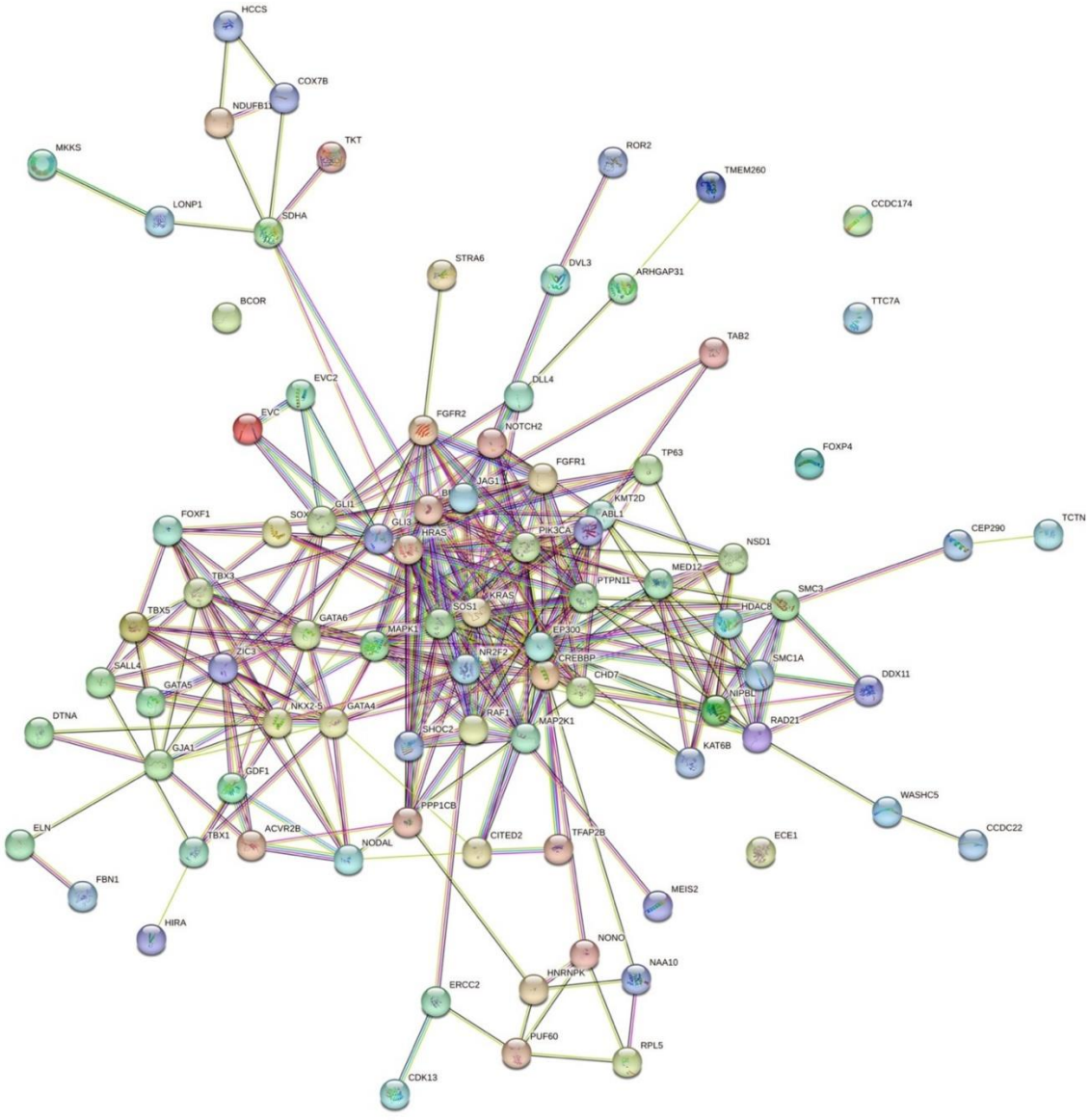


Figure 3. Protein-protein interactions in ventricular septal defect (VSD) as analyzed using the STRING web server.

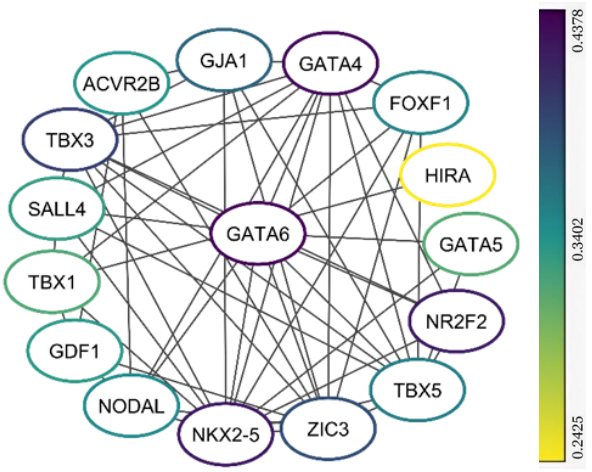


Figure 4. Sixteen protein-protein interaction networks in ventricular septal defect (VSD) were analyzed using Cytoscape software. The darker the color (black), the higher the closeness centrality value (0.4378), light green indicates a medium closeness centrality score (0.3402), while lighter colors (yellow) represent the lowest value (0.2425).

Table 1. Closeness centrality scores were obtained using Cytoscape software, indicating that the GATA6 protein exhibited the highest score

Abbreviation	Protein name	Closeness centrality
GATA6	GATA binding factor 6	0.4378
GATA4	GATA binding factor 4	0.4331
NKX2-5	NK2 homeobox 5	0.4218
NR2F2	Nuclear receptor subfamily 2 group F member 2	0.4196
TBX3	T-Bbx protein 3	0.3970
ZIC3	ZIC family member 3	0.3875
GJA1	Gap junction alpha-1	0.3715
TBX5	T-box protein 5	0.3552
FOXF1	Forkhead box F1	0.3506
NODAL	Nodal growth differentiation factor	0.3461
ACVR2B	Activin A receptor type 2B	0.3375
GDF1	Growth differentiation factor 1	0.3347
SALL4	Sal-like protein 4	0.3306
TBX1	T-box protein 1	0.3188
GATA5	GATA binding factor 5	0.3164
HIRA	Histone chaperone	0.2425

Gene ontology and fold enrichment

Gene ontology enrichment analysis revealed that the biological process (BP) parameters, including atrioventricular canal development, the cellular component (CC) parameters, specifically the protein-DNA complex, and the molecular function (MF) parameters related to transcription regulatory regions, each demonstrated the highest FDR and enrichment pathways (**Figure 5**).

GATA6 protein 3D structure visualization and domain

The visualization results of the GATA6 protein, generated using AlphaFold 3, are presented in 3D animation and a structural image highlighting the specific functional domain (**Figure 6A** and **6B**). The results show the location of 17 nsSNPs and mutant proteins in the GATA6 protein. Seven structures (Gly15Arg, Ala21Gly, Leu118Phe, Ala178Val, Ser184Asp, and Leu198Val) are located in the transcription activation domain 1 (TAD1) region, presented in purple in the 3D image (**Figure 6A** and **6B**). Seven other structures (Arg456Cys, Asn466Asp, Ala467Thr, Asn466His, Ala459Thr, Arg456His, and Thr452Ala) are located in the zinc finger 2 (ZF2) domain, presented in dark brown in the 3D image (**Figure 6A** and **6B**). Finally, two structures (Pro555Ala and Ala575Pro) are located in the nuclear localization signal (NLS) structure, presented in light brown in the 3D image (**Figure 6A** and **6B**).

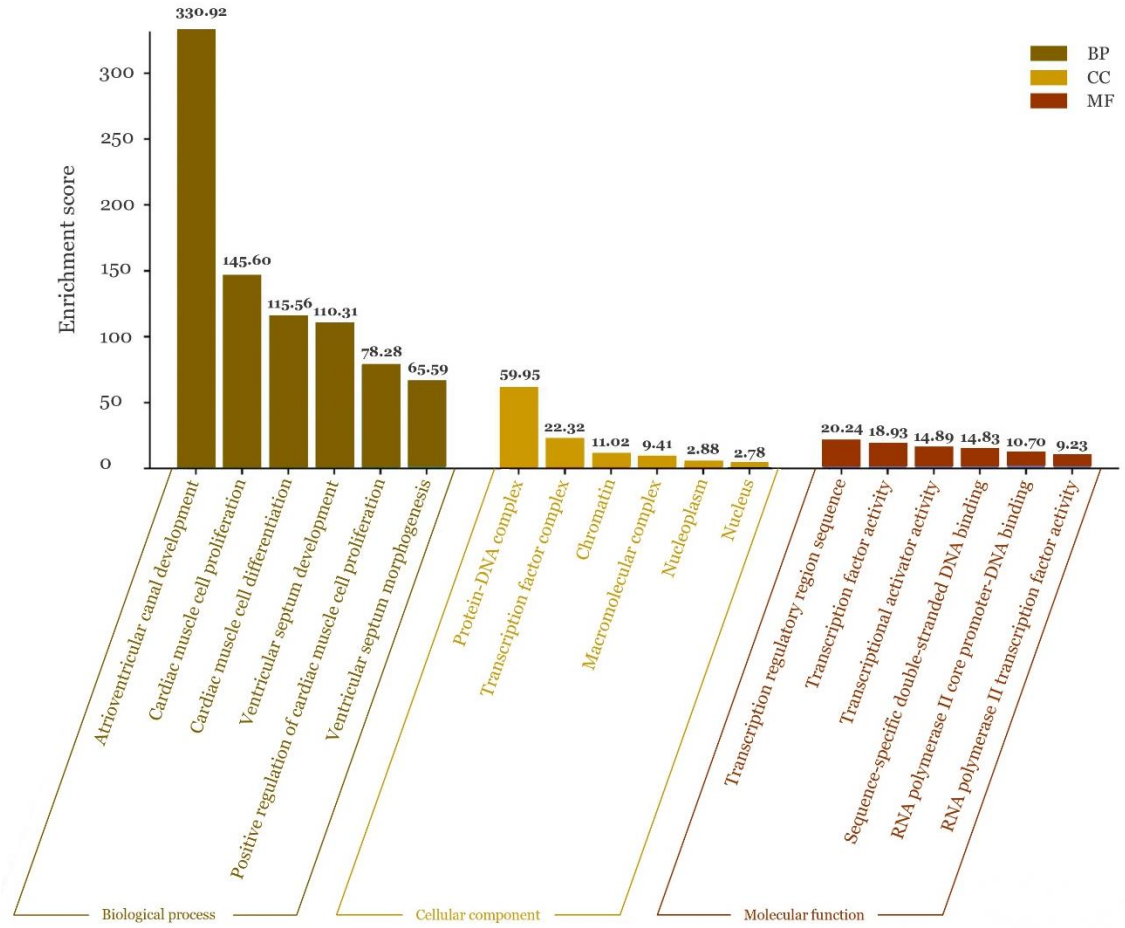


Figure 5. Gene ontology results for the 16 genes were obtained using the DAVID web server. The categories assessed include biological process (BP), cellular component (CC), and molecular function (MF).

Identification and assessment of protein consequences in missense mutation

Screening of 680 missense SNPs in the *GATA6* gene, based on clinically significant criteria (benign, likely benign, likely pathogenic, and pathogenic), identified 17 potential SNPs. A predictive assessment of these SNPs was conducted using the SIFT, PolyPhen-2, MutPred2, Fathmm, SNP&GO, I-Mutant v2.0, and PON-P2 algorithms to evaluate their impact on the GATA6 protein structure (**Tables 2–4**). The screening results indicate that there were four reference sequences (rs) of nsSNP with the highest scores and strong potential to cause VSD: rs387906813 (N466H and N466D), rs387906817 (T452A), rs387906818 (R456C), and rs387906820 (A467T).

Effect on amino acid mutation in polarity and functional change

Each type of amino acid exhibits distinct properties, characteristics, molecular weights, and functions. In the present study, mutation prediction tests were performed on protein amino acid structures, focusing on parameters such as changes in the polarity effects of amino acids, influenced by the specific amino acid being altered, and functional changes, including hydrophilic and hydrophobic properties (**Table 5**). Alterations in polarity effects and functional properties of amino acids could result in significant protein malfunctions and changes in solubility. Various mutations, including E68V, N466H, N466D, and R456C, modify amino acid polarity (**Table 5**). Among these predictions, four mutations—G15R, A459T, T452A, and A467T—indicate substantial potential for dysfunction and alterations in solubility within the GATA6 protein (**Table 5**).

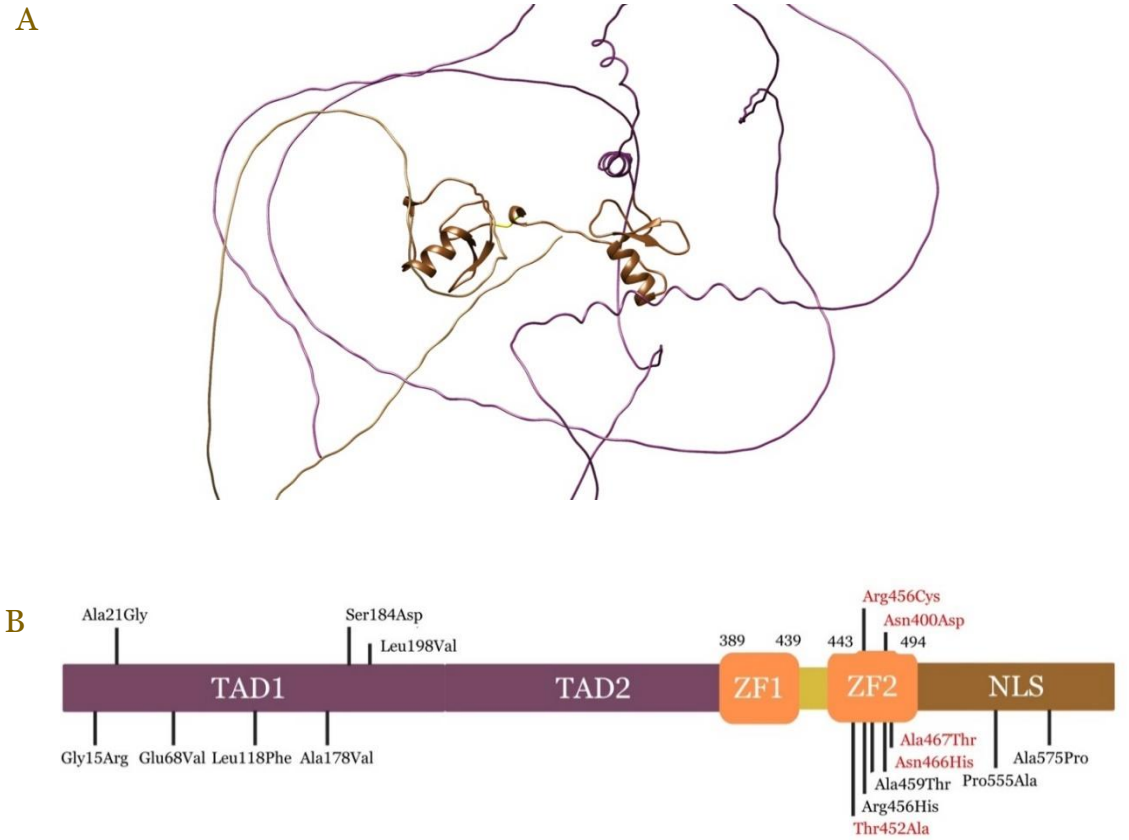
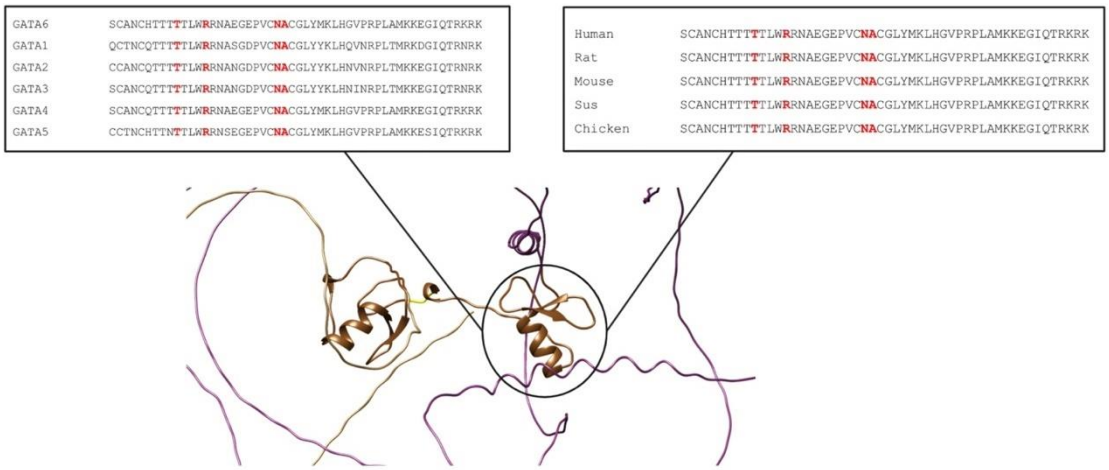


Figure 6. (A) Three-dimensional (3D) visualization of the GATA6 protein generated using AlphaFold 3. (B) Functional domains of the GATA6 protein identified in this study. NLS: nuclear localization signal; TAD: transcription activation domain; ZF: zinc finger.

Prediction of amino acid conservation status and multiple sequence alignment of GATA6 amino acid

Sequence alignment results obtained using the Genome JP web server indicated that amino acid residues 385–496 in the GATA6 protein reside within a conserved region when compared to other GATA family members and mammalian species, excluding humans (**Figure 7A**). These results indicate that amino acid residues 452, 456, 466, and 467, selected based on the screening results and the highest probabilities from the identification of missense mutations, are located in a crucial region, particularly as structural and functional units of the GATA6 protein. Mutations in these four amino acids are predicted to have a strong potential, both molecularly and clinically, for the occurrence of VSD. ConSurf analysis further demonstrated that specific areas of the GATA6 protein are highly conserved across different species, highlighting their functional significance (**Figure 7B**). These results explain that the positions of amino acid residues 452, 456, 466, and 467 are part of a crucial domain responsible for the binding between GATA6 protein and its target DNA. If mutations occur in these residues, it could disrupt the transcriptional process of the target DNA for the GATA6 protein, thereby increasing the likelihood of VSD. Additionally, protein function prediction analysis identified several missense mutations in the *GATA6* gene that are predicted to have a deleterious impact on protein function. These results provide evidence that mutations in the *GATA6* gene may disrupt its normal function, potentially leading to the pathogenesis of VSD and other cardiac abnormalities.

A



B

e - An exposed residue according to the NACSES algorithm
b - A buried residue according to the NACSES algorithm
f - A predicted functional residue (highly conserved and exposed)
s - A predicted structural residue (highly conserved and buried)
x - Insufficient data – the calculation for this site was performed on less than 10% of the sequences

Figure 7. (A) Sequence alignment of the GATA6 protein with other GATA family members in humans and with the GATA6 protein in other species. Red highlights: amino acids 452, 456, 466, and 467. (B) Results of the functional analysis using the ConSurf web server. The results indicated that amino acid residues 452, 456, 466, and 467 are conserved.

Table 2. Prediction results of non-synonymous single nucleotide polymorphisms (nsSNP) using SIFT, Polyphen-2, and Fathmm web servers

No	rs number	Amino acid change	SIFT	Polyphen-2				Fathmm	
			Prediction	Score	Sensitivity	Specificity	Prediction	Prediction	Score
1	rs116262672	G15R	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-5.37
2	rs139666654	A21G	Tolerated	0.930	0.81	0.94	Possibly damaging	Damaging	-4.84
3	rs146243018	P555A	Tolerated	0.826	0.84	0.93	Possibly damaging	Damaging	-4.66
4	rs149569288	A575P	Tolerated	0.800	0.84	0.93	Possibly damaging	Damaging	-4.59
5	rs151176879	A459T	Tolerated	0.789	0.85	0.93	Possibly damaging	Damaging	-5.99
6	rs185325359	A284G	Tolerated	0.733	0.85	0.92	Possibly damaging	Damaging	-4.74
7	rs200483324	E68V	Tolerated	0.736	0.85	0.92	Possibly damaging	Damaging	-4.82
8	rs368297251	L118F	Tolerated	0.998	0.21	0.99	Probably damaging	Damaging	-5.44
9	rs387906813	N466H	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-7.29
10	rs387906813	N466D	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-7.13
11	rs387906814	L198V	Tolerated	0.877	0.83	0.94	Possibly damaging	Damaging	-4.85
12	rs387906815	A178V	Deleterious	0.991	0.71	0.97	Probably damaging	Damaging	-4.72
13	rs387906816	S184N	Tolerated	0.001	0.99	0.15	Benign	Damaging	-5.07
14	rs387906817	T452A	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-7.32
15	rs387906818	R456C	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-7.6
16	rs387906819	R456H	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-7.54
17	rs387906820	A467T	Deleterious	1.000	0.00	1.00	Probably damaging	Damaging	-6.74

Table 3. Prediction results of non-synonymous single nucleotide polymorphisms (nsSNP) using MutPred2 and I-Mutant v2.0

No	MutPred2				I-mutant v 2.0			
	Score	Molecular mechanism	Probability	*p*-value	pH	Temperature (Celsius)	DDG score (kcal/mol)	Interpretation
1	0.640	Altered disordered interface	0.40	0.0043	7	25	-1.45	Decrease stability
2	0.279	-	-	-	7	25	-1.59	Decrease stability
3	0.136	-	-	-	7	25	-1.22	Decrease stability
4	0.514	-	-	-	7	25	0.10	Increase stability
5	0.411	-	-	-	7	25	-0.90	Decrease stability
6	0.900	-	-	-	7	25	-0.75	Decrease stability
7	0.199	-	-	-	7	25	0.49	Increase stability
8	0.688	Loss of helix	0.27	0.05	7	25	-0.68	Decrease stability
9	0.901	Altered metal binding	0.51	0.0047	7	25	-2.25	Decrease stability
10	0.898	Altered metal binding	0.40	0.0084	7	25	-2.07	Decrease stability
11	0.071	-	-	-	7	25	-0.68	Decrease stability
12	0.245	-	-	-	7	25	1.18	Increase stability
13	0.146	-	-	-	7	25	1.26	Increase stability
14	0.859	Altered metal binding	0.24	0.03	7	25	-1.16	Decrease stability
15	0.791	Loss of loop	0.27	0.05	7	25	-0.19	Decrease stability
16	0.733	Altered transmembrane protein	0.19	0.0073	7	25	0.05	Increase stability
17	0.811	Loss of disulfide linkage at C456	0.11	0.04	7	25	-0.87	Decrease stability

Table 4. Prediction results of non-synonymous single nucleotide polymorphisms (nsSNP) using the SNP&GO and PON-P2 web servers

No	SNP&GO				PON-P2	
	Amino acid change	Prediction	Reliable index score	Probability	Probability for pathogenicity	Prediction
1	G15R	Disease	2	0.62	0.63	Unknown
2	A21G	Neutral	1	0.46	0.31	Unknown
3	P555A	Neutral	6	0.22	0.42	Unknown
4	A575P	Disease	2	0.61	0.72	Unknown
5	A459T	Neutral	5	0.23	0.80	Pathogenic
6	A284G	Neutral	9	0.07	0.37	Unknown
7	E68V	Neutral	3	0.32	0.31	Unknown
8	L118F	Neutral	1	0.44	0.21	Neutral
9	N466H	Disease	8	0.91	0.90	Pathogenic
10	N466D	Disease	9	0.94	0.78	Unknown
11	L198V	Disease	0	0.51	0.24	Neutral
12	A178V	Disease	3	0.62	0.55	Unknown
13	S184N	Neutral	5	0.25	0.44	Unknown
14	T452A	Disease	8	0.88	0.89	Pathogenic
15	R456C	Disease	8	0.90	0.98	Pathogenic
16	R456H	Disease	7	0.84	0.85	Unknown
17	A467T	Disease	8	0.91	0.89	Pathogenic

Reliable index score indicates how strong association of SNP with a phenotype or function. Score 0–2: very low reliability and confidence; score 3–4: low reliability and confidence; score 5–6: moderate reliability and confidence; score 7–8: high reliability and confidence; and score 9–10: very high reliability and confidence.

Table 5. GATA6 amino acid mutations and their effects on polarity and functional changes

No	Mutation	Effect in polarity	Change in polarity	Functional change
1	G15R	Non-polar to basic polar	Yes	Hydrophobic to hydrophilic
2	A21G	Non-polar to non-polar	No	Hydrophobic to hydrophobic
3	P555A	Non-polar to non-polar	No	Hydrophobic to hydrophobic
4	A575P	Non-polar to non-polar	No	Hydrophobic to hydrophobic
5	A459T	Non-polar to polar	Yes	Hydrophobic to hydrophilic
6	A284G	Non-polar to non-polar	No	Hydrophobic to hydrophobic
7	E68V	Acidic polar to non-polar	Yes	Hydrophilic to hydrophilic
8	L118F	Non-polar to non-polar	No	Hydrophobic to hydrophobic
9	N466H	Polar to basic polar	Yes	Hydrophilic to hydrophilic
10	N466D	Polar to acidic polar	Yes	Hydrophilic to hydrophilic
11	L198V	Non-polar to non-polar	No	Hydrophobic to hydrophobic
12	A178V	Non-polar to non-polar	No	Hydrophobic to hydrophobic
13	S184N	Polar to polar	No	Hydrophilic to hydrophilic
14	T452A	Polar to non-polar	Yes	Hydrophilic to hydrophobic
15	R456C	Basic polar to polar	Yes	Hydrophilic to hydrophilic
16	R456H	Basic polar to basic polar	No	Hydrophilic to hydrophilic
17	A467T	Non-polar to polar	Yes	Hydrophobic to hydrophilic

Functional classification, variant conservation, and variant stability change score of GATA6 protein

The results of the functional classification test on the GATA6 protein indicate that the T452A, R456C, N466D, N466H, and A467T mutations are structurally critical (**Figure 8A**). The results from the variant conservation score parameters indicate that the mutations in the amino acids T452A and R456C have a score of -14, N466D has a score of -8, and A467T has a score of -12. These scores suggest that these mutations are detrimental to the structure and function of the GATA6 protein because they are located in a conserved domain (**Figure 8B**). The variant stability change score results indicate that the mutation in amino acid residue T452A has a score of -14, R456C has a score of -12, N466D has a score of -4, N466H has a score of -6, and A467T has a score of -8. Therefore, the T452A mutation has a strong potential to cause changes in the stability of the GATA6 protein (**Figure 8C**).

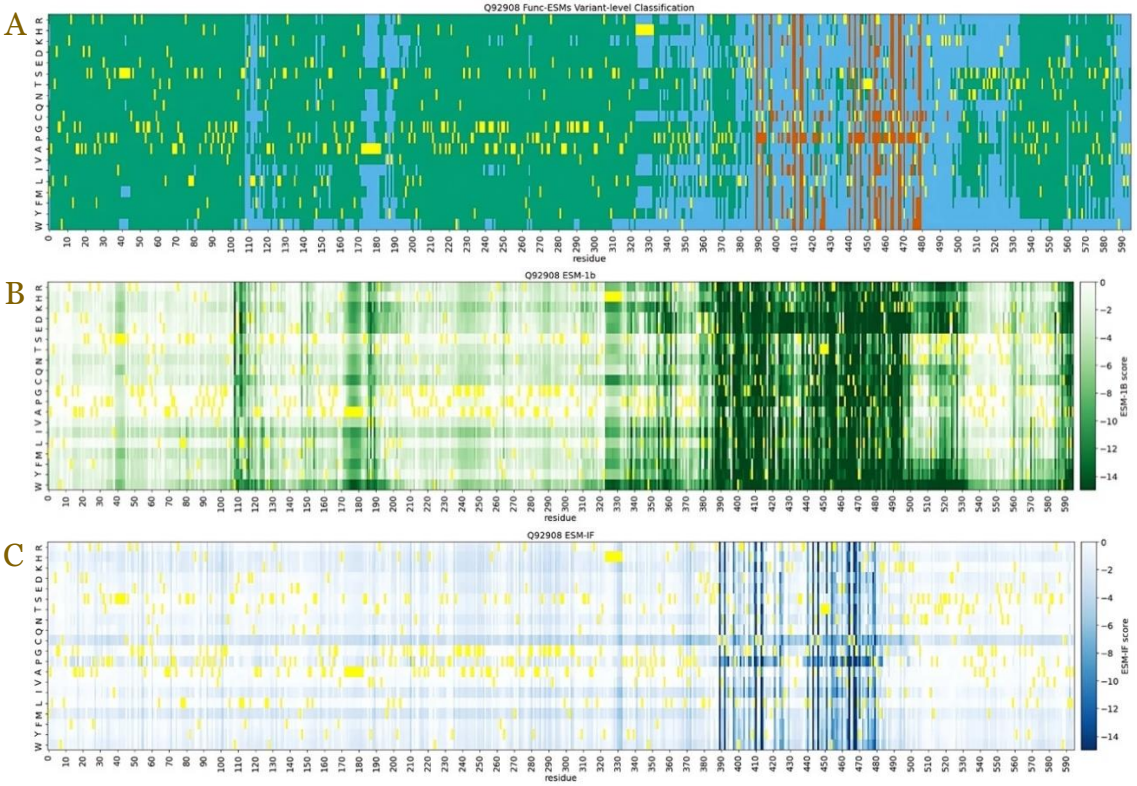
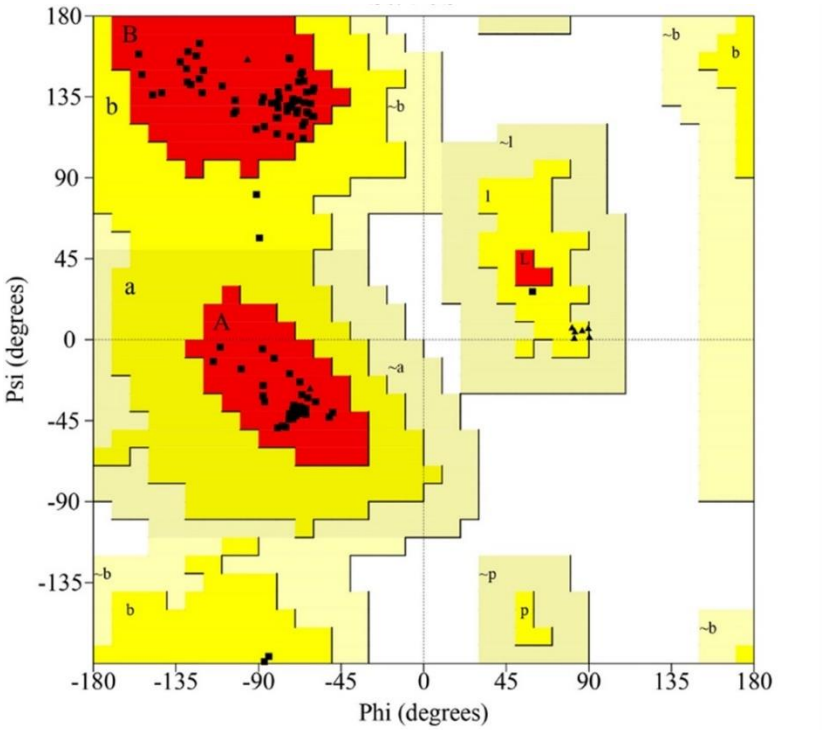


Figure 8. Functional classification, variant conservation and stability change scores between wild type (WT) vs mutation of GATA6. (A) Functional classification of GATA6 protein wild type (WT) vs mutation. Green: tolerant (variant) or wild type-like (residue); blue: stable-but-inactive (variant) or functionally relevant (residue); red: structurally critical (variant) or total loss (residue). (B) Variant conservation score of GATA6 protein WT vs mutation. Green: tolerant (variant) or wild type-like (residue); blue: stable-but-inactive (variant) or functionally relevant (residue); grey: class not assigned (residue); yellow: wild type amino acid; light-shaded variants: non-detrimental; dark-shaded variants: detrimental. (C) Variant stability change score of GATA6 protein WT vs mutation. Grey: class not assigned (residue); yellow: wild type amino acid; light-shaded variants: non-detrimental; dark-shaded variants: detrimental.

Protein modeling of GATA6 in the conserved area

Protein modeling of GATA6 focuses on the functional domain region (amino acids 385–496) because this domain is conserved and plays a crucial role as the zinc finger 2 domain, which facilitates the binding between the transcription factor GATA6 and its target DNA (**Figure 9**). Validation testing of the GATA6 protein domain is necessary to assess whether the modeled protein in its tertiary folding closely approximates its natural structure. This test is considered valid if the Ramachandran plot score for residues in the most favored region is greater than 90%. Validation of this model demonstrated a score of 94.6% for residues in the most favored region, confirming its validity (**Figure 9**).



Plot statistics

Residues in most favoured regions [A,B,L]	87	94.6%
Residues in additional allowed regions [a,b,l,p]	5	5.4%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	92	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	10	
Number of proline residues	8	
Number of residues	112	

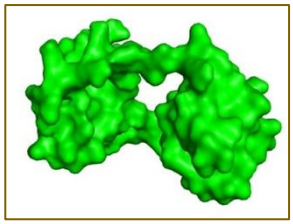


Figure 9. Ramachandran plot results and GATA6 protein modeling. Ramachandran plot results assessing the quality of GATA6 protein modeling, with a score of 94.6%. Insert showed the results of GATA6 protein modeling in the conserved and functional domain (amino acids 385–496) using AlphaFold 3 protein modeling.

Molecular dynamics simulation

The results of the molecular dynamics simulation over 50 ns between the GATA6 wild type protein and its mutants indicate changes in conformation and flexibility (**Figure 10**). These results suggest that mutations at amino acid residues 452, 456, 466, and 467 affect the conformation and flexibility of the GATA6 protein. Measurements of the aggregation score between the GATA6 wild-type protein and its mutants showed that mutations at amino acid residues 452, 456, 466, and 467 have higher aggregation scores compared to the wild type, indicating that these mutations increase the GATA6 protein susceptibility to aggregation.

The energy difference score measurements between the GATA6 wild type and mutants reveal that the N466D mutation has a negative effect, reducing protein stability, while the T452A, R456C, N466H, and A467T mutations have a positive effect, increasing the stability of the GATA6 protein (**Table 6**). The results of root mean square deviation (RMSD) from the molecular dynamics simulation indicate that the average RMSD value for the GATA6 mutant protein is lower than that of the wild type, with the R456C mutation having the lowest value (**Figure 11A-E**). The root mean square fluctuation (RMSF) parameter indicates fluctuations in amino acid

residues between the GATA6 mutants and the wild-type. The results show that the A467T and R456C mutations have higher fluctuations compared to the wild type (**Figure 11F-J**).

The calculation of the radius of gyration (Rg) shows that two mutations (R456C and N466D) have a higher average Rg than the wild-type, while two mutations (T452A and N466H) have a lower average value than the wild type, and one mutation (A467T) has an average value equal to the wild type (**Figure 12A-E**). These results indicate changes in the compactness during the molecular dynamics simulation of the GATA6 protein mutants. The calculation of the number of hydrogen bonds shows no significant difference between the GATA6 mutants and the wild type (**Figure 12F-J**). This indicates that mutations at amino acid residues 452, 456, 466, and 467 do not alter the hydrogen bonds within the secondary structure of the GATA6 protein.

Table 6. Average aggregation score and energy difference of GATA6 protein between wild type versus five mutations

GATA6 protein	Average protein solubility score	Energy difference between wild type vs mutated protein
Wild type	-0.8512	0 kcal/mol
T452A	-0.8446	0.407 kcal/mol
R456C	-0.7852	1.11 kcal/mol
N466D	-0.8406	-0.66 kcal/mol
N466H	-0.8504	0.197 kcal/mol
A467T	-0.846	1.019 kcal/mol

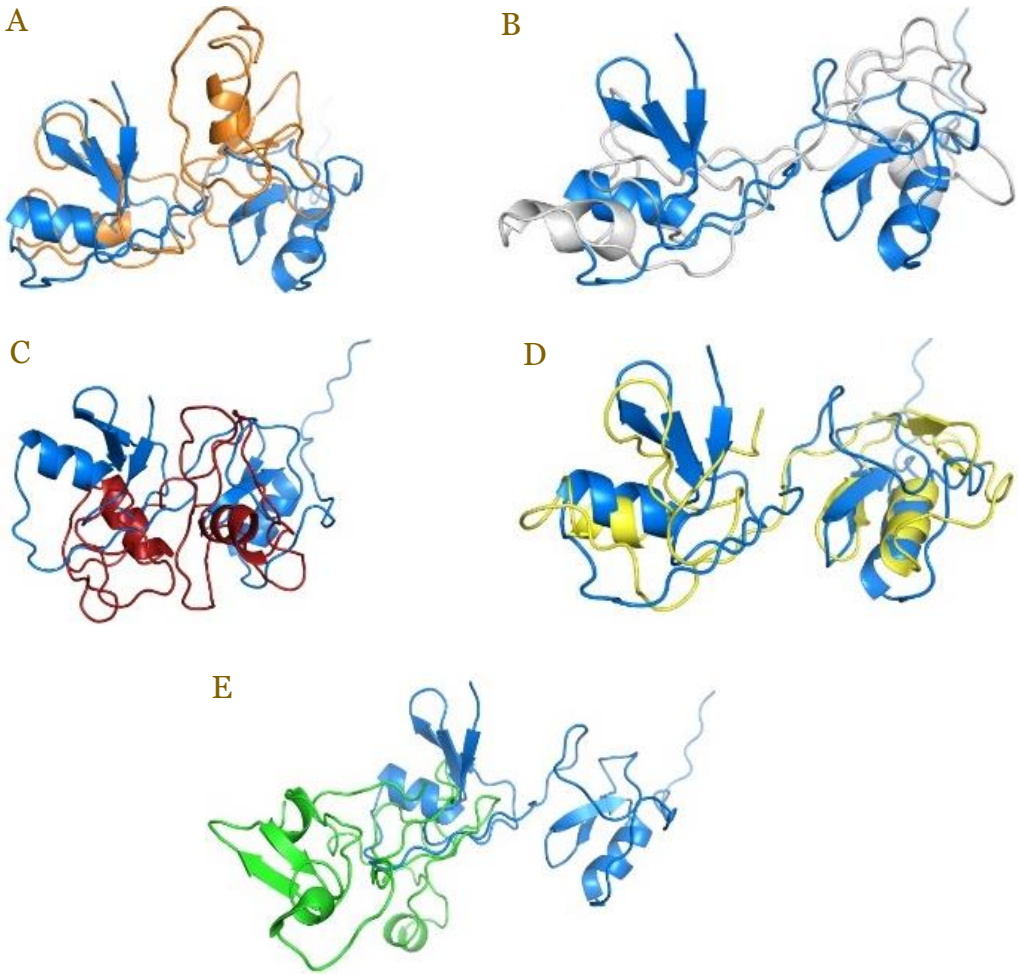


Figure 10. Results of the molecular dynamics simulation of GATA6 protein after 50 ns: (A) wild type vs T452A, (B) wild type vs R456C, (C) wild type vs N466H, (D) wild type vs N466D, and (E) wild type vs A467T.

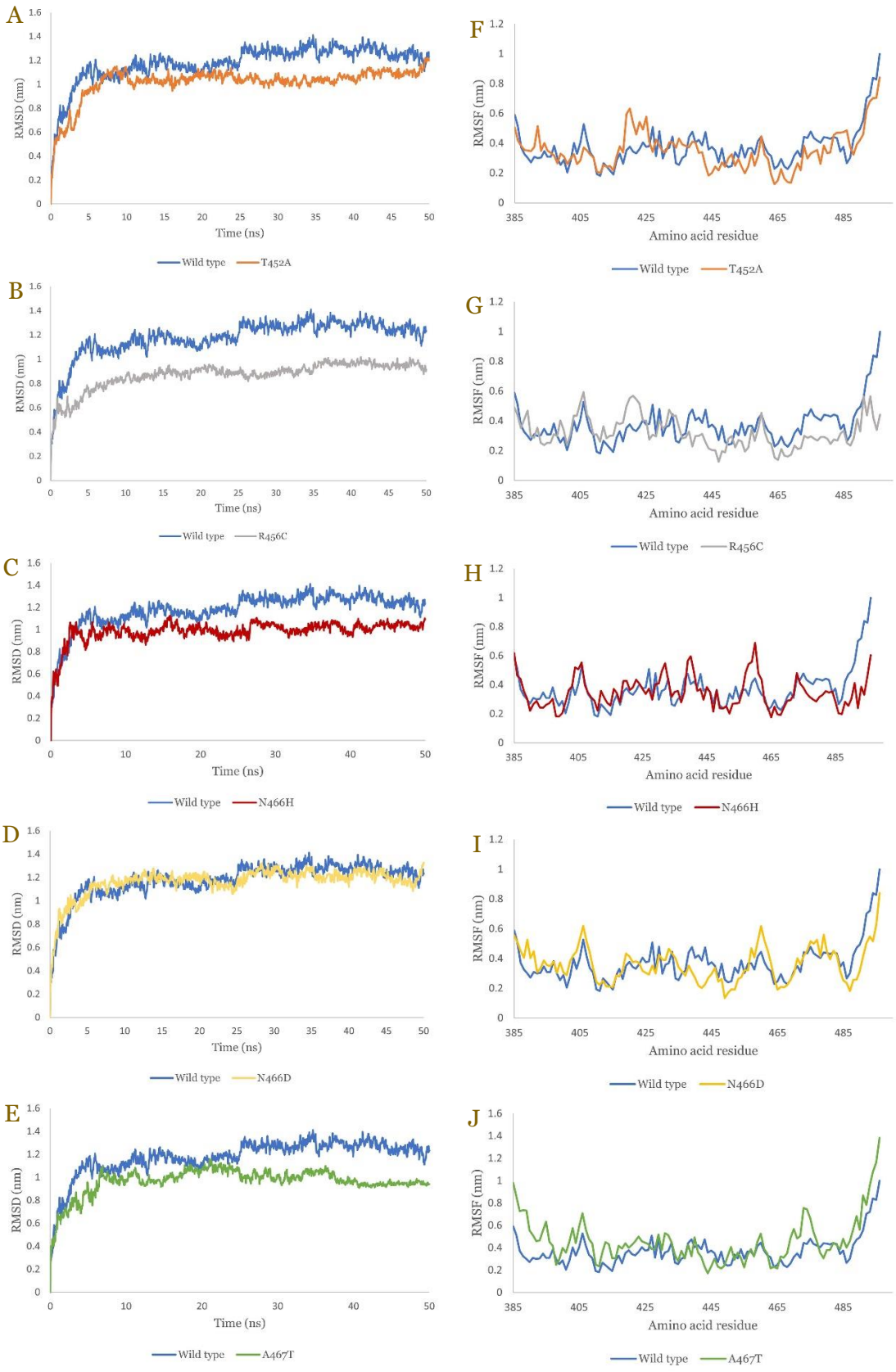


Figure 11. The results of root mean square deviation (RMSD) (A-E) and the root mean square fluctuation (RMSF) (F-J) of GATA6 protein after 50 ns simulation. RMSD of (A) wild type vs T452A; (B) wild type vs R456C; (C) wild type vs N466H; (D) wild type vs N466D; (E) wild type vs A467T. RMSF of (F) wild type vs T452A; (G) wild type vs R456C; (H) wild type vs N466H; (I) wild type vs N466D; (J) wild type vs A467T.

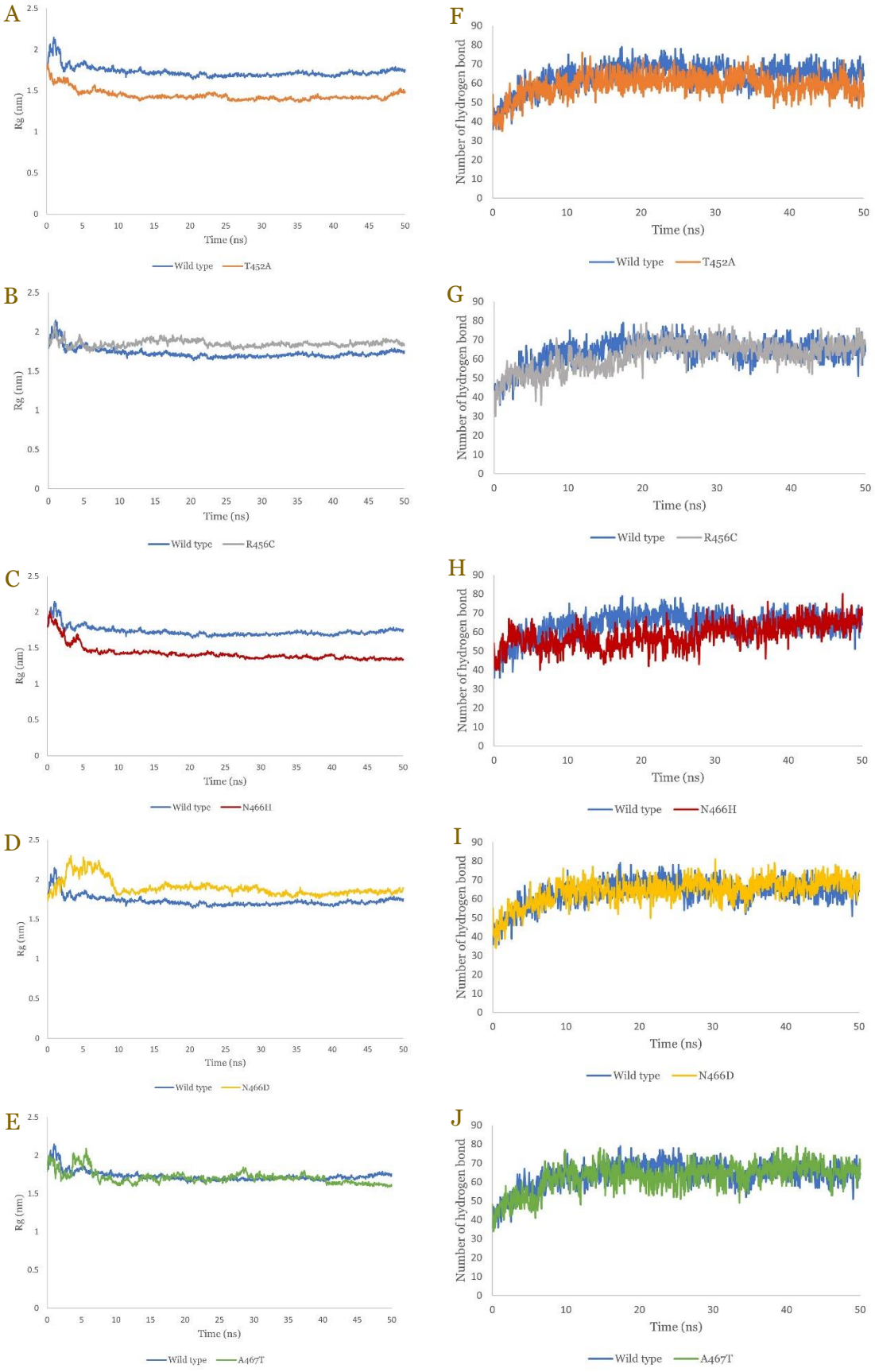


Figure 12. The results of GATA6 protein radius of gyration (Rg) (A-E) and GATA6 protein hydrogen bonds (F-J) after 50 ns simulation: (A) wild type vs T452A; (B) wild type vs R456C; (C) wild type vs N466H; (D) wild type vs N466D; (E) wild type vs A467T. Number of hydrogen bonds (F) wild type vs T452A; (G) wild type vs R456C; (H) wild type vs N466H; (I) wild type vs N466D; (J) wild type vs A467T.

Discussion

The *GATA6* gene is a member of the GATA family of zinc finger transcription factors, located on chromosome 18q11.1-q11.2 [39]. It plays a crucial role in regulating cell differentiation and organogenesis across the three primary germ layers: endoderm, ectoderm, and mesoderm, with a particular emphasis on cardiac development [40,41]. The presence of SNPs in this gene may clinically increase the risk of genetic disorders, including congenital heart diseases such as atrial septal defect (ASD), VSD, atrioventricular septal defect (AVSD), conotruncal heart malformations, and tetralogy of Fallot, as well as syndromic pancreatic agenesis and heart rhythm disturbances, such as atrial fibrillation [11].

Our study found 16 proteins linked through protein-protein interaction network analysis, and this protein interaction network was involved in the formation of the atrioventricular septum, the proliferation and differentiation of cardiac muscle cells, and the development of the ventricular septum. GATA6 protein is the most important role within this 16 protein-protein interaction network. The GATA6 interactions demonstrate considerable strength and act as a central hub within the network [42]. Gene ontology analysis revealed that these 16 proteins were involved in the formation of the atrioventricular septum, the proliferation and differentiation of cardiac muscle cells, and the development of the ventricular septum [43].

The essential roles of GATA4 and GATA6 in the precise regulation of heart formation processes are supported by findings from the protein-protein interaction network analysis and corroborated by in vivo studies conducted on rat fetuses [44]. Inactivation or mutations in these two genes have been shown to lead to congenital heart disease and result in lethality in fetuses [44,45]. Mutation testing of the *GATA6* in rat embryos revealed a failure in the formation of neural crest-derived smooth muscle cells within blood vessels [46]. Defects in the morphogenesis of the ventricular septum were observed in heart progenitors, leading to congenital heart defects such as truncus arteriosus and VSD [41].

The interaction between the T-box transcription factor 5 (TBX5) protein and GATA family transcription factors constitutes a crucial biomolecular process for heart morphogenesis [47]. In patients with CHD, mutations in the *GATA* and *TBX5* are known to disrupt this interaction [48]. Haploinsufficiency of *GATA4* and *TBX5* results in two significant defects during the morphogenesis process: the formation of the atrioventricular septum and myocardial development [49]. In contrast, haploinsufficiency of *GATA6* and *TBX5* leads to mild defects in myocardial development [50]. Additionally, genome-wide association study revealed interactions among the *GATA*, *Nkx2.5*, and *TBX5* genes in heart formation, indicating that mutations within this complex interaction can result in congenital heart defects [15].

Our in silico analysis of 680 missense SNPs identified four potential SNPs with the highest likelihood of causing VSD: rs387906813, rs387906817, rs387906819, and rs387906820. These SNPs were evaluated using the SIFT, PolyPhen-2, and Fathmm web servers, which predicted them to be harmful and damaging to their respective amino acid products due to DNA base alterations. The MutPred2 web server elucidated mechanisms underlying the predicted detrimental properties of these SNPs, including disrupted metal binding, loss of loop structure in amino acids, and loss of disulfide bonds in the GATA6 protein [28]. The DNA base changes and subsequent disruptions in amino acid production contribute to the decreased stability of the GATA6 protein, as assessed by I-Mutant v2.0 software in the present study. Screening with the SNP&GO and PON-P2 web servers indicated that these SNPs are pathogenic and may lead to disease [31,51].

The identified missense SNPs in the *GATA6* result in substitutions that alter the produced amino acids: threonine is replaced by alanine at amino acid 452, arginine is replaced by cysteine at amino acid 456, asparagine is substituted with aspartic acid and histidine at amino acid 466, and alanine is replaced by threonine at amino acid 467 [52]. These substitutions could modify the polarity of the amino acids, potentially impacting the functional structure of the GATA6 protein and reducing its stability and effectiveness in interacting with target genes [24,53,54].

Functional prediction using the ConSurf web server evaluated the changes in these five non-synonymous SNPs and their resulting amino acid residues, revealing that these reference sequences reside in highly conserved and exposed regions [33,55,56]. Multiple sequence alignment tests across the six GATA protein families and four mammalian species, in addition to

humans, confirmed that these amino acid residue changes occur in conserved regions that are integral to the functional residues of the GATA6 protein [57]. Notably, amino acid residues 452, 456, 466, and 467 are part of the zinc finger domain of the GATA6 protein, which binds to the DNA target gene at the enhancer region and serves as a crucial domain for protein-protein and GATA6 protein-DNA interactions [52,58]. These findings suggest that mutations within these conserved regions and the zinc finger domain of the *GATA6* gene may lead to functional disruptions, potentially contributing to the development of ventricular septal defect [59].

Our molecular dynamics simulation on the zinc finger domain and functional residues of the GATA6 protein, with a focus on the RMSD parameter, indicated that the wild type protein stabilized after 5 ns. In comparison, the mutations T452A, R456C, N466H, and A467T exhibited lower RMSD values than the wild type, indicating structural instability [60]. The Rg values revealed that the R456C and N466D mutations had higher Rg values than the wild type after 3 ns, suggesting a reduction in the compactness of the GATA6 protein [61]. Conversely, the T452A and N466H mutations had lower Rg values than the wild type, indicating that these mutations increased the compactness of the GATA6 protein [62].

In the present study, the RMSF parameter demonstrated that the A467T mutation exhibited the highest residue fluctuation compared to the wild type and other mutations, resulting in significant flexibility. These findings suggest that mutations within this domain lead to overall denaturation and reduced compactness of the GATA6 protein relative to the control, adversely affecting its function and structure [63-65]. Solubility tests in the present study further supported these results, revealing that all five mutations displayed lower solubility values than the control. Notably, the R456C (rs387906818) and A467T (rs387906820) mutations showed a propensity for instability and increased GATA6 protein aggregation [37]. This aggregation is attributed to alterations in amino acid types, as well as changes in polarity (from basic polar to polar and from non-polar to polar) and functional properties (from hydrophilic to hydrophobic) [66,67]. These alterations impact the secondary folding structure of the protein, initially exposing it to water—a crucial characteristic for biological function—but subsequently rendering it permeable to water, thereby disrupting the protein's secondary structure [68]. This disruption increases the solubility value of the protein and promotes protein aggregation [69,70]. The discovery of structural disruptions due to changes in polarity properties, structural instability, reduced compactness, increased denaturation, and increased aggregation, along with the position in important protein domains (zinc finger domain), in silico tests of mutant GATA6 proteins, leads to a decrease in the physiological performance or function of GATA6 protein [71]. This disruption primarily occurs in the GATA6-GATA4 interaction pathway, which works synergistically to enhance signal transduction, control the onset of cardiac myocyte differentiation through the Wnt/β-catenin-independent pathway and atrial natriuretic factor protein that initiates the heart field marker, resulting in the activation of the GATA5 transcription factor and triggering the process of cardiomyocyte cell differentiation and proliferation [46,72,73]. Ultimately, mutations at these residues result in loss of function of the GATA6 protein, influencing the transcription, translation, and regulation of cardiogenesis factor proteins and affecting proliferation and differentiation in cardiomyocyte cells [74].

In this study, comprehensive in silico testing showed that GATA6 plays a central role in the protein-protein interaction network, and mutations in this protein's functional domain are significant in the process of VSD occurrence. However, this testing was conducted using computational approaches, some conditions in the complex body systems were not considered, such as the role of transport proteins, interactions with the target DNA of the GATA6 protein, and the role of cell membranes, among others. Therefore, further in vivo, in vitro, and clinical tests should be conducted to provide a detailed explanation of the molecular mechanisms involved in a more complex system.

Conclusion

GATA6 protein plays a crucial and central role in cardiac organogenesis, particularly in the formation of the atrioventricular canal of the heart. The screening for pathogenic nsSNPs, coupled with assessments of *GATA6* mutation characteristics, structural and polarity changes, and the properties of amino acids in the GATA6 protein, identified four potential SNPs:

rs387906813, rs387906817, rs387906818, and rs387906820. Further analyses, including functional domain analysis, multiple sequence alignment, and molecular dynamics simulations, identified two SNPs—rs387906818 and rs387906820—as having the potential to cause VSD.

Ethics approval

Not required.

Acknowledgments

None.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

The present study received support from Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM), Universitas Airlangga, Surabaya, Indonesia (Grant number: 9491/UN3.FK/PT.01.03/2023).

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

How to cite

Hidayat T, Irwanto I, Rohman A, *et al*. Comprehensive in silico analysis of single nucleotide polymorphism and molecular dynamics simulation of human GATA6 protein in ventricular septal defect. Narra J 2024; 4 (3): e1344 - http://doi.org/10.52225/narra.v4i3.1344.

References

1. Salari N, Faryadras F, Shohaimi S, *et al.* Global prevalence of congenital heart diseases in infants: A systematic review and meta-analysis. J Neonatal Nurs 2024;30(6):570-575.
2. Liu Y, Chen S, Zühlke L, *et al.* Global birth prevalence of congenital heart defects 1970–2017: Updated systematic review and meta-analysis of 260 studies. Int J Epidemiol 2019;48(2):455-463.
3. Pihl C, Sillesen AS, Norsk JB, *et al.* The prevalence and spontaneous closure of ventricular septal defects the first year of life. Neonatology 2024:1-10.
4. Ismail MT, Hidayati F, Krisdinarti L, *et al.* Epidemiological profile of congenital heart disease in a national referral hospital. Acta Cardiol Indones 2015;1(2):66-71.
5. Qanitha A, Qalby N, Amir M, *et al.* Clinical cardiology in South East Asia: Indonesian lessons from the present towards improvement. Glob Heart 2022;17(1):66.
6. GBD 2017 Congenital Heart Disease Collaborators. Global, regional, and national burden of congenital heart disease, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. Lancet Child Adolesc Health 2020;4(3):185-200.
7. Amelia P, Yosephine AG, Tobing TC, *et al.* Association between type of congenital heart disease with child growth and development status: A cross-sectional study in Medan, Indonesia. Narra J 2023;3(3):e414.
8. Zhao K, Nan S, Li Y, *et al.* Comprehensive analysis and characterization of the GATA gene family, with emphasis on the GATA6 transcription factor in poplar. IJMS 2023;24(18):14118.
9. Wamaitha SE, del Valle I, Cho LTY, *et al.* Gata6 potently initiates reprograming of pluripotent and differentiated cells to extraembryonic endoderm stem cells. Genes Dev 2015;29(12):1239-1255.
10. Miles ML, Cowan N, Jackson G. A nonsense GATA6 Mutation Explains History Of Congenital Heart Defects And 10 Years Of Poorly-Controlled Diabetes Lacking DKA in a non-obese 30 year-old incidentally found to have pancreatic hypoplasia. AACE Clin Case Rep 2020;6(3):e123-e126.
11. Maitra M, Koenig SN, Srivastava D, Garg V. Identification of GATA6 sequence variants in patients with congenital heart defects. Pediatr Res 2010;68(4):281-285.

12. Yue X, Luo Y, Wang J, Huang D. Monogenic diabetes with GATA6 mutations: Characterization of a novel family and a comprehensive analysis of the GATA6 clinical and genetics traits. Mol Biotechnol 2024;66(3):467-474.
13. Li C, Li X, Pang S, *et al.* Novel and functional DNA sequence variants within the GATA6 gene promoter in ventricular septal defects. Int J Mol Sci 2014;15(7):12677-12687.
14. Yasuhara J, Garg V. Genetics of congenital heart disease: A narrative review of recent advances and clinical implications. Transl Pediatr 2021;10(9):2366-2386.
15. Wu H, Wu H, He Y, *et al.* Functional characterization of GATA6 genetic variants associated with mild congenital heart defects. Biochem Biophys Res Commun 2023;641:77-83.
16. Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48(D1):D845-D855.
17. Stelzer G, Rosen N, Plaschkes I, *et al.* The GeneCards suite: From gene data mining to disease genome sequence analyses. Curr Protoc Bioinformatics 2016;54:1.30.1-1.30.33.
18. Szklarczyk D, Kirsch R, Koutrouli M, *et al.* The STRING database in 2023: Protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 2023;51(D1):D638-D646.
19. Majeed A, Mukhtar S. Protein-protein interaction network exploration using cytoscape. In: Mukhtar S, editor. Protein-protein interactions. New York: Humana Press; 2023.
20. Sherman BT, Hao M, Qiu J, *et al.* DAVID: A web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res 2022;50(W1):W216-W221.
21. Abramson J, Adler J, Dunger J, *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024;630(8016):493-500.
22. The UniProt Consortium. UniProt: The universal protein knowledgebase in 2021. Nucleic Acids Res 2021;49(D1):D480-D489.
23. Saccone SF, Quan J, Mehta G, *et al.* New tools and methods for direct programmatic access to the dbSNP relational database. Nucleic Acids Res 2011;39(Database issue):D901-D907.
24. Vaser R, Adusumalli S, Leng SN, *et al.* SIFT missense predictions for genomes. Nat Protoc 2016;11(1):1-9.
25. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013;76:7.20.1-7.20.41.
26. Irfan M, Iqbal T, Hashmi S, *et al.* Insilico prediction and functional analysis of nonsynonymous SNPs in human CTLA4 gene. Sci Rep 2022;12(1):20441.
27. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Research 2005;33(Web Server issue):W306-W310.
28. Pejaver V, Urresti J, Lugo-Martinez J, *et al.* Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 2020;11(1):5918.
29. Research Square. Assessment of single nucleotide variants in hGPRC5A gene using in-silico tools. Available from: https://www.researchsquare.com/article/rs-3576134/v1. Accessed: 17 November 2013.
30. Rogers MF, Shihab HA, Mort M, *et al.* FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. Bioinformatics 2018;34(3):511-513.
31. Capriotti E, Calabrese R, Fariselli P, *et al.* WS-SNPs&GO: A web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics 2013;14 Suppl 3:S6.
32. Yariv B, Yariv E, Kessel A, *et al.* Using evolutionary data to make sense of macromolecules with a "face-lifted" ConSurf. Protein Sci 2023;32(3):e4582.
33. Hung JH, Weng Z. Sequence alignment and homology search with BLAST and ClustalW. Cold Spring Harb Protoc 2016;2016(11):pdb.prot093088.
34. Land H, Humble MS. YASARA: A tool to obtain structural guidance in biocatalytic investigations. In: Bornscheuer UT, Höhne M, editors. Protein engineering. New York: Humana Press; 2018.
35. Syaban M, Putra G, Vadhana R, *et al.* Molecular docking analysis and dynamics simulation of ethanol extract of *Citrus sinensis* as a Keap1 and NMDA inhibitor in brain injury. World Acad Sci J 2023;5(2):14.
36. Zambrano R, Jamroz M, Szczasiuk A, *et al.* AGGRESCAN3D (A3D): Server for prediction of aggregation properties of protein structures. Nucleic Acids Res 2015;43(W1):W306-W313.
37. Kuriata A, Iglesias V, Pujols J, *et al.* Aggrescan3D (A3D) 2.0: Prediction and engineering of protein solubility. Nucleic Acids Res 2019;47(W1):W300-W307.

38. Abraham MJ, Murtola T, Schulz R, *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1-2:19-25.
39. Pugnaloni F, Martini L, De Rose DU, *et al.* A new variant in the GATA6 gene associated with tracheoesophageal fistula, pulmonary vein stenosis and neonatal diabetes. Horm Res Paediatr 2024.
40. Gao J, Chen YH, Peterson LC. GATA family transcriptional factors: Emerging suspects in hematologic disorders. Exp Hematol Oncol 2015;4:28.
41. Lentjes MH, Niessen HE, Akiyama Y, *et al.* The emerging role of GATA transcription factors in development and disease. Expert Rev Mol Med 2016;18:e3.
42. Roy S, Manners HN, Elmsallati A, Kalita JK. Alignment of protein-protein interaction networks. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology. Amsterdam: Elsevier; 2019.
43. Gonzalez-Teran B, Pittman M, Felix F, *et al.* Transcription factor protein interactomes reveal genetic determinants in heart disease. Cell 2022;185(5):794-814.e30.
44. Maitra M, Schluterman MK, Nichols HA, *et al.* Interaction of Gata4 and Gata6 with Tbx5 is critical for normal cardiac development. Dev Biol 2009;326(2):368-377.
45. Xin M, Davis CA, Molkentin JD, *et al.* A threshold of *GATA4* and *GATA6* expression is required for cardiovascular development. Proc Natl Acad Sci U S A 2006;103(30):11189-11194.
46. Afouda BA. Towards understanding the gene-specific roles of GATA factors in heart development: Does GATA4 lead the way?. Int J Mol Sci 2022;23(9):5255.
47. Tremblay M, Sanchez-Ferras O, Bouchard M. GATA transcription factors in development and disease. Development 2018;145(20):dev164384.
48. Suluba E, Shuwei L, Xia Q, Mwanga A. Congenital heart diseases: Genetics, non-inherited risk factors, and signaling pathways. Egypt J Med Hum Genet 2020;21:11.
49. Behiry EG, Al-Azzouny MA, Sabry D, *et al.* Association of NKX2-5, GATA4, and TBX5 polymorphisms with congenital heart disease in Egyptian children. Mol Genet Genomic Med 2019;7(5):e612.
50. Misra C, Chang SW, Basu M, *et al.* Disruption of myocardial Gata4 and Tbx5 results in defects in cardiomyocyte proliferation and atrioventricular septation. Hum Mol Genet 2014;23(19):5025-5035.
51. Bolunduț AC, Lazea C, Mihu CM. Genetic alterations of transcription factors and signaling molecules involved in the development of congenital heart defects-A narrative review. Children (Basel) 2023;10(5):812.
52. Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction method for fast and reliable identification of harmful variants. PLoS One 2015;10(2):e0117380.
53. Allen HL, Flanagan SE, Shaw-Smith C, *et al.* GATA6 haploinsufficiency causes pancreatic agenesis in humans. Nat Genet 2011;44(1):20-22.
54. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins 2011;79(3):830-838.
55. Diaz DJ, Kulikova AV, Ellington AD, Wilke CO. Using machine learning to predict the effects and consequences of mutations in proteins. Curr Opin Struct Biol 2023;78:102518.
56. Pei J. Multiple protein sequence alignment. Curr Opin Struct Biol 2008;18(3):382-386.
57. Mahfuz AMUB, Khan MdA, Deb P, *et al.* Identification of deleterious single nucleotide polymorphism (SNP)s in the human *TBX5* gene & prediction of their structural & functional consequences: An in silico approach. Biochem Biophys Rep 2021;28:101179.
58. Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics 2017;109(5-6):419-431.
59. Eifes S, Chudasama KK, Molnes J, *et al.* A novel GATA6 mutation in a child with congenital heart malformation and neonatal diabetes. Clin Case Rep 2013;1(2):86-90.
60. Cassandri M, Smirnov A, Novelli F, *et al.* Zinc-finger proteins in health and disease. Cell Death Discov 2017;3:17071.
61. Islam MJ, Khan AM, Parves MR, *et al.* Prediction of deleterious non-synonymous SNPs of human STK11 gene by combining algorithms, molecular docking, and molecular dynamics simulation. Sci Rep 2019;9(1):16426.
62. Ghahremanian S, Rashidi MM, Raeisi K, Toghraie D. Molecular dynamics simulation approach for discovering potential inhibitors against SARS-CoV-2: A structural review. J Mol Liq 2022;354:118901.
63. Dash R, Ali MC, Dash N, *et al.* Structural and dynamic characterizations highlight the deleterious role of SULT1A1 R213H polymorphism in substrate binding. Int J Mol Sci 2019;20(24):6256.

64. Abdul Samad F, Suliman BA, Basha SH, *et al.* 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(5):e0153999.
65. Kumar A, Ojha KK. Molecular dynamics simulation methods to study structural dynamics of proteins. In: Saudagar P, Tripathi T, editors. Protein folding dynamics and stability. Singapore: Springer Singapore; 2023.
66. Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. Neuron 2018;99(6):1129-1143.
67. Di Rienzo L, Miotto M, Bò L, *et al.* Characterizing hydropathy of amino acid side chain in a protein environment by investigating the structural changes of water molecules network. Front Mol Biosci 2021;8:626837.
68. Hamzi H, Rajabpour A, Roldán É,Hassanali A. Learning the hydrophilic, hydrophobic and aromatic character of amino acids from thermal relaxation and interfacial thermal conductance. J Phys Chem B 2022;126(3):670-678.
69. Durell SR, Ben-Naim A. Hydrophobic-hydrophilic forces in protein folding. Biopolymers 2017;107(8):e23020.
70. Ptak-Kaczor M, Banach M, Stapor K, *et al.* Solubility and aggregation of selected proteins interpreted on the basis of hydrophobicity distribution. Int J Mol Sci 2021;22(9):5002.
71. Qadah T, Jamal MS. Computational analysis of protein structure changes as a result of nondeletion insertion mutations in human β-globin gene suggests possible cause of β-thalassemia. Biomed Res Int 2019;2019:9210841.
72. Bisson JA, Gordillo M, Kumar R, *et al.* GATA6 regulates WNT and BMP programs to pattern precardiac mesoderm during the earliest stages of human cardiogenesis 2024. Preprint. bioRxiv 2024;2024.07.09.602666.
73. Holtzinger A, Evans T. Gata5 and Gata6 are functionally redundant in zebrafish for specification of cardiomyocytes. Dev Biol 2007;312(2):613-622.
74. Sam J, Torregroza I, Evans T. Gata6 functions in zebrafish endoderm to regulate late differentiating arterial pole cardiogenesis. Development 2024;151(17):dev202895.