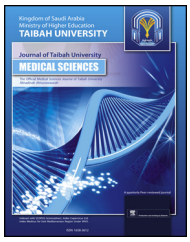
Journal of Taibah University Medical Sciences (2022) 17(6), 1074–1082

Taibah University

Journal of Taibah University Medical Sciences

www.sciencedirect.com



Original Article

In silico analysis of missense variants of the *C1qA* gene related to infection and autoimmune diseases

Mohammed Y. Behairy, Ph.D [a], ALi A. Abdelrahman, Ph.D [b], Hoda Y. Abdallah, Ph.D [c,d], Emad El-Deen A. Ibrahim, Ph.D [e], Anwar A. Sayed, Ph.D [f,g] and Marwa M. Azab, Ph.D [b,*]

[a] *Department of Microbiology and Immunology, Faculty of Pharmacy, University of Sadat City, Sadat City, Egypt*
[b] *Department of Microbiology and Immunology, Faculty of Pharmacy, Suez Canal University, Ismailia, Egypt*
[c] *Department of Histology and Cell Biology (Genetics Unit), Faculty of Medicine, Suez Canal University, Ismailia, Egypt*
[d] *Molecular Biology Unit, Center of Excellence in Molecular and Cellular Medicine, Faculty of Medicine, Suez Canal University, Ismailia, Egypt*
[e] *Department of Anesthesia, Intensive Care and Pain Management, Faculty of Medicine, Suez Canal University, Ismailia, Egypt*
[f] *Department of Medical Microbiology and Immunology, Taibah University, Almadinah Almunawwarah, KSA*
[g] *Department of Surgery and Cancer, Imperial College London, London, United Kingdom*

Received 30 December 2021; revised 28 February 2022; accepted 28 April 2022; Available online 13 May 2022

الملخص

أهداف البحث: يشكل C1q عاملا رئيسيا فى تنشيط الطريق الكلاسيكى فى الجهاز المناعى المتمم المسئول عن وظيفتى البلعمة و الطهاية بالغتي الأهمية. و يعد جين *C1qA* أحد ثلاثة جينات مسئولة عن شفرةC1q . وقد لوحظ أن الاختلالات في C1q و بالذات في C1qA لها علاقة بزيادة العدوى و تسمم الدم و كذلك مرض الذئبة الحمامية المجموعية. وهذه الاختلالات قد تنشأ عن الطفرات المغلطة بما لها من تأثير ضار على وظيفة وشكل البروتين. لذلك فإن التعرف على هذه الطفرات المغلطة ذات الخطورة العالية يعد ضرورة ملحة للتعرف على هؤلاء الأشخاص المعرضين لاتخاذ إجراءات الوقاية والعلاج الملائمة.

طرق البحث: تم إجراء دراسة متكاملة لفحص ١٨٤طفرة مغلطة موجودة فى جين *C1qA* وذلك بإستخدام أدوات مختلفة تعتمد على منهجيات ولوغاريتمات متعددة. وقد شملت الدراسة فحص تأثير هذه الطفرات المغلطة على وظيفة وشكل وثبات البروتين. كذلك فحص موقع هذه الطفرات المغلطة على نطاقات البروتين بالإضافة إلى موقعها على الهيكل الثانوي للبروتين ودرجة الثبات الفيلوجينى للأماكن التى وقعت فيها.

النتائج: بداية من ١٨٤ طفرة مغلطة، تم العثور على عشر طفرات مغلطة يتوقع لها أن تكون الاشد إضرارا بوظيفة وشكل البروتين وذلك بجميع الأدوات المستخدمة.

الاستنتاجات: تم العثور على عشر طفرات مغلطة يتوقع لها أن تكون الأشد إضرارا بوظيفة وشكل البروتين مما قد يؤدي إلى الاصابة بالعدوى وبتسمم الدم وبمرض الذئبة الحمامية المجموعية كذلك. هذه العشر طفرات مغلطة تشكل أفضل مرشح للقيام بمزيد من الدراسات التجريبية.

الكلمات المفتاحية: بروتين؛ تحاليل حاسوبية؛ العدوى؛ الذئبة الحمامية المجموعية؛ التغيرات الفردية متعددة الأشكال للنيوكليوتيد

Abstract

Objectives: C1q is a key activator of the classical pathway of the complement system and exerts consequences relating to opsonization and phagocytosis. The *C1qA* gene is one of three genes encoding the C1q molecule. Defects in C1q, and especially in C1qA, have been linked to an increased susceptibility to infection, sepsis, and systemic lupus erythematosus. These defects could arise from missense single nucleotide polymorphisms (SNPs) and their deleterious impacts on protein structure and function. Thus, identifying high-risk missense SNPs in *C1qA* has become a necessity if we are to identify appropriate measures for prevention and management of affected patients.

Methods: A comprehensive *in silico* study was conducted to screen the 184 missense SNPs in the *C1qA* gene using different tools with different algorithms and approaches. We investigated the impact of SNPs on protein function, stability, and structure. In addition, we identified the location of the SNPs on protein domains, secondary

* Corresponding address: Department of Microbiology and Immunology, Faculty of Pharmacy, Suez Canal University, Ismailia, Egypt.
E-mail: marwaazab2515@yahoo.com (M.M. Azab)
Peer review under responsibility of Taibah University.

ELSEVIER Production and hosting by Elsevier

 https://doi.org/10.1016/j.jtumed.2022.04.014

structure alignment, and the phylogenetic conservation of their positions.

Results: Of the 184 missense SNPs, 10 SNPs were predicted to be the most damaging to protein function and structure.

Conclusion: Ten missense SNPs were predicted to have the highest risk of damaging protein function and structure, thus leading to infection, sepsis, and systemic lupus erythematosus. These 10 SNPs constitute the best candidates for further experimental investigations.

Keywords: *C1qA*; In silico; Infection; SLE; SNP

Introduction

The complement system is a critical part of our immune system that protects our bodies from invading bacteria and deleterious immune complexes through its cascade of enzymes, receptors and proteins.[1] Genetic defects in the constituents of complement have been found to increase susceptibility to infections and even to autoimmune disorders. These findings highlighted the need to identify these genetic defects in different components of the complement system as a necessity to facilitate the early detection of patients with this liability as this would allow timely and appropriate prevention and management.[2,3] Single nucleotide polymorphisms (SNPs) are the most common type of these genetic variations. One of its subtypes, the missense SNPs, has attracted significant attention with regards to disease pathogenesis as alterations in amino acid sequences can lead to subsequent alterations in protein function and stability.[4]

Of the constituents of complement, C1q is a key molecule that activates the classical pathway of the complement system that leads to the initiation of opsonization and phagocytosis. C1q is coded by a cluster of three genes (*C1qA*, *C1qB* and *C1qC*) that produce the three types of polypeptide chains that compose the C1q molecule.[1] Defects in the C1q genes, especially *C1qA*, have been associated with increased susceptibility to infection, sepsis and septicemia.[3,5] Moreover, these defects were linked to the development of systemic lupus erythematosus (SLE).[2,6] Consequently, the comprehensive study of *C1qA* genetic variants has become very important so that we can identify such essential associations. Before moving to high cost and lengthy experimental studies, it is important to consider the many bioinformatics tools that have been developed to investigate genetic defects *in silico*, thus leading to better engagement and analysis.

In this study, we performed a comprehensive analysis of missense variants of the *C1qA* gene and investigated their impact on the structure and function of this remarkably important molecule.

Materials and Methods

General information

General information related to the *C1qA* gene were retrieved from Ensembl and National Center for Biotechnology Information (NCBI) databases. Gene ontology (GO) information was retrieved from Genecards.org and compartments.jensenlab.org was used as a source for subcellular localization data.

Retrieval of C1qA SNPs

C1qA SNPs were retrieved from National Center for Biotechnology Information (NCBI) through variation viewer using dbSNP as a source database (https://www.ncbi.nlm.nih.gov/variation/view/) by using *C1qA* as an entry or 712 as a gene ID. The resulting SNPs were filtered for screening and only missense variants were selected for further screening due to the consequential alterations of amino acids in the protein sequence.

Predicting the impact of SNPs on protein function

The accession numbers of all missense SNPs were uploaded into VEP (Variant Effect Predictor) (https://www.ensembl.org/Tools/VEP); then, we enabled SIFT (Sorting Intolerant from Tolerant) and PolyPhen (Polymorphism Phenotyping) to predict pathogenicity. After that, we used SIFT and PolyPhen-2 (Polymorphism Phenotyping v2) tools as well. The SNPs identified by all these databases to be deleterious and probably damaging were selected for further screening. SIFT uses sequence homology and the physical properties of amino acids to predict the impact of missense SNPs on the function of proteins. SIFT computes a score that lies within a range of 0 and 1; a deleterious effect is predicted with a score lying within the range of 0 and 0.05 (https://sift.bii.a-star.edu.sg/).[7] PolyPhen-2 relies on physical and comparative approaches to predict the impact of amino acid alterations. PolyPhen-2 predicts the effect of mutation and also generates a score ranging from 0 (a benign SNP) to 1 (a damaging SNP) (http://genetics.bwh.harvard.edu/pph2).[8,9]

Next, the resulting SNPs were further analyzed by four other bioinformatics tools (PROVEAN (Protein Variation Effect Analyzer), SNP&GO, PHD-SNP and SNAP) to increase the accuracy and strength of our results. PROVEAN (http://provean.jcvi.org/seq_submit.php) relies on blast hits to calculate the delta alignment value and computes the final PROVEAN score using −2.5 as a cutoff.[10] SNPs&GO relies on the functional annotation of proteins for predicting the effect of variations (https://snps.biofold.org/snps-and-go/snps-and-go.html).[11] PHD-SNP (http://snps.biofold.org/phd-snp/phd-snp.html) depends on support vector machines (SVMs) to predict the relationships between the new phenotype resulting from the missense variation and the human genetic disorders.[12] The SNAP2 tool (https://rostlab.org/services/snap/) uses a new neural network method to differentiate the effect variants from the neutral variants. The SNAP2 produces a score that ranges from −100 that indicates a prediction of strong neutral to +100 that

indicates a strong effect.[13] The SNPs predicted by all these tools to be deleterious were selected for further analysis.

Prediction of the impact of SNPs on protein stability

We used the I-Mutant 2.0 tool to predict changes in the protein stability associated with our missense SNPs (https://folding.biofold.org/i-mutant/i-mutant2.0.html). This is a support vector machine that has the ability to predict the direction of protein stability after mutation as well as the related free energy change of protein stability (DDG).[14]

The identification of missense SNPs on protein domains

We used the InterPro tool to identify the location of missense SNPs on the conserved domains of our protein (https://www.ebi.ac.uk/interpro/). The InterPro tool can provide functional analysis of a selected protein so that its conserved sites and important domains could be identified.[15]

The identification of evolutionarily conserved positions in our protein

The ConSurf server (https://consurf.tau.ac.il) was used to identify functional residues by analyzing phylogenetic relationships found between different homologous sequences and investigating the phylogenetic conservation of protein sequences. A conservation score was calculated for each residue that ranges from 1 to 9. The grade begins with the variable positions then the residues with intermediate conservation and ends with highly conserved residues. Moreover, each residue is predicted as structural or functional.[16,17]

Predicting the secondary structures of the C1qA chain

The SOPMA tool was used to predict the secondary structure of the C1qA chain and the alignment of the mutated residues in its secondary structure (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html). In addition, the secondary structure was predicted with deleterious missense mutations. SOPMA is an improved version of the self-optimized prediction method (SOPM) that could analyze the multiple alignments of a specific protein sequence to predict its secondary structure.[18]

Gene–gene interaction

GeneMANIA was used to produce a network for gene–gene interaction analysis related to the *C1qA* gene (http://www.genemania.org). GeneMANIA predicts genes with strong interactions depending on different types of resources and data, including physical interaction data, co-expression data, predicted functional relationships, co-localization information, genetic interaction information, pathway data and information relating to protein domains.[19]

Identification of the impact of SNPs on the three-dimensional structure of the protein

The effects of the most damaging SNPs on protein 3D structure were analyzed using the HOPE tool (https://www3.cmbi.umcn.nl/hope/). The HOPE tool collects data from numerous sources including the UniProt database and DAS services. In addition, HOPE can build homology models by YASARA to predict the impact of SNPs on the structure and function of the related protein.[20]

Results

General information

C1qA gene is located at 1p36.12 and is a protein-coding gene that consists of three exons. *C1qA* is 3216 nucleotides in length (NCBI Gene ID: 712) and has three transcripts (ensemble.org). This gene encodes the A chain of serum complement C1q that composes the first element of the complement system with C1s and C1r. C1q is composed of 18 chains, including 6 A-chains (https://www.ncbi.nlm.nih.gov/gene/712). Figure S1A shows the subcellular localization of C1qA and Figure S1B shows gene ontology analysis (Gene cards.org).

SNPs in the C1qA gene

In *C1qA*, we identified 1285 single nucleotide variations (accessed 19 November 2021). Of these, there were 184 missense variants, 96 synonymous variants, 148 SNPs in the 5′-untranslated region (UTR), 117 SNPs in the 3′-UTR, and 618 intron variants, in addition to other upstream and downstream variants.

The predicted impact of SNPs on protein function

VEP, SIFT and Polyphen-2 databases showed that 35 SNPs were deleterious (as determined by all three databases). After further analysis of these 35 SNPs with four other *in silico* tools (PROVEAN, SNP&GO, PHD-SNP, SNAP), 20 SNPs were predicted by all these tools to be deleterious and disease-causing. The predictions and scores for all 20 SNPs are shown in Table 1. Of these SNPs, four involved two mutant alleles each, resulting in missense mutations: rs1250029890, rs528301944, rs749595647 and rs754597784. These 20 SNPs were selected for further analysis steps.

Predicting the impact of SNPs on protein stability

The I-Mutant 2.0 server was used to analyze the effects of the twenty selected SNPs on protein stability by estimating the free energy change values (DDG) and the reliability index (RI). Thirteen missense SNPs were found to reduce protein stability as shown in Table 2. The impact of these SNPs on protein stability was predicted to be more damaging and were selected for further analysis.

Identifying the locations of the missense SNPs within protein domains

The InterPro server predicted that C1qA (InterPro entry: IPR037572) contains a C1q domain (Interpro entry: IPR001073) and a collagen triple helix repeat (Interpro entry: IPR008160). The positions of the thirteen high risk SNPs are shown in Table 3. Rs1269727956 and rs902565316 are

Table 1: Predictions and scores for damaging missense SNPs identified by six bioinformatics tools.

SNP Id	AA change	SIFT Prediction (score)	PolyPhen-2 Prediction (score)	PROVEAN Prediction (score)	SNP&GO Prediction (RI score)	PHD-SNP Prediction (RI score)	SNAP2 Prediction (score)
rs369062665	G31R	Deleterious (0.00)	Probably damaging (0.999)	Deleterious (−5.414)	Disease (2)	Disease (6)	Effect (64)
rs1250029890	G46A	Deleterious (0.00)	Probably damaging (0.999)	Deleterious (−4.814)	Disease (4)	Disease (2)	Effect (49)
	G46V	Deleterious (0.00)	Probably damaging (0.999)	Deleterious (−7.187)	Disease (5)	Disease (7)	Effect (65)
rs1308364521	G55E	Deleterious (0.00)	Probably damaging (0.932)	Deleterious (−6.044)	Disease (4)	Disease (3)	Effect (61)
rs1434507661	G65V	Deleterious (0.00)	Probably damaging (0.996)	Deleterious (−5.743)	Disease (6)	Disease (8)	Effect (65)
rs1269727956	G71R	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−7.402)	Disease (6)	Disease (3)	Effect (72)
rs528301944	G71E	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−7.368)	Disease (7)	Disease (1)	Effect (73)
	G71V	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−8.278)	Disease (7)	Disease (7)	Effect (60)
rs1174724209	G86W	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−6.804)	Disease (6)	Disease (5)	Effect (85)
rs749595647	G89S	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−5.619)	Disease (6)	Disease (3)	Effect (83)
	G89C	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−8.34)	Disease (7)	Disease (5)	Effect (71)
rs771758729	G89V	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−8.21)	Disease (8)	Disease (6)	Effect (83)
rs902565316	G92R	Deleterious (0.00)	Probably damaging (0.995)	Deleterious (−7.241)	Disease (7)	Disease (4)	Effect (63)
rs1043270464	G149C	Deleterious (0.00)	Probably damaging (0.999)	Deleterious (−8.304)	Disease (6)	Disease (8)	Effect (52)
rs953707145	G157R	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−7.826)	Disease (8)	Disease (7)	Effect (81)
rs1570073403	G157D	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−6.848)	Disease (8)	Disease (6)	Effect (72)
rs754597784	Y159H	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−4.891)	Disease (7)	Disease (7)	Effect (86)
	Y159D	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−9.783)	Disease (8)	Disease (9)	Effect (90)
rs1570073417	T162P	Deleterious (0.00)	Probably damaging (0.995)	Deleterious (−3.606)	Disease (4)	Disease (9)	Effect (82)
rs755725663	W216R	Deleterious (0.00)	Probably damaging (0.97)	Deleterious (−11.526)	Disease (2)	Disease (5)	Effect (76)
rs146884691	I226N	Deleterious (0.01)	Probably damaging (0.999)	Deleterious (−2.823)	Disease (4)	Disease (5)	Effect (75)
rs1332792872	F236V	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−6.758)	Disease (7)	Disease (8)	Effect (78)
rs1213084266	G238S	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−5.797)	Disease (3)	Disease (5)	Effect (65)
rs1373684177	G238V	Deleterious (0.00)	Probably damaging (1.000)	Deleterious (−8.689)	Disease (4)	Disease (7)	Effect (82)

located on the collagen triple helix repeat while rs1043270464, rs953707145, rs1570073403, rs754597784, rs157007341, rs755725663, rs146884691, rs1332792872, rs1213084266, rs1373684177 are all located in the C1q domain.

Table 2: The impact of missense SNPs on protein stability.

SNP Id	AA change	I-mutant 2 prediction	Reliability index (RI)	DDG value (kcal/mol)
rs369062665	G31R	**Decrease**	7	−1.72
rs1250029890	G46A	Increase	8	0.4
	G46V	Increase	7	0.39
rs1308364521	G55E	Increase	6	0.67
rs1434507661	G65V	Increase	2	−1.33
rs1269727956	G71R	**Decrease**	1	−0.39
rs528301944	G71E	Increase	6	1.05
	G71V	Increase	4	−0.97
rs1174724209	G86W	Increase	0	−0.21
rs749595647	G89S	Increase	3	−0.53
	G89C	Increase	0	−0.62
rs771758729	G89V	Increase	6	−0.44
rs902565316	G92R	**Decrease**	2	−0.72
rs1043270464	G149C	**Decrease**	5	0.01
rs953707145	G157R	**Decrease**	8	−0.95
rs1570073403	G157D	**Decrease**	6	−0.65
rs754597784	Y159H	**Decrease**	8	−1.68
	Y159D	**Decrease**	3	−1.11
rs1570073417	T162P	**Decrease**	3	−0.93
rs755725663	W216R	**Decrease**	6	−1.7
rs146884691	I226N	**Decrease**	7	−1.22
rs1332792872	F236V	**Decrease**	7	−2.74
rs1213084266	G238S	**Decrease**	8	−1.02
rs1373684177	G238V	**Decrease**	2	−1.53

The predictions of decreasing protein stability are shown in bold font.

*Identification of the evolutionarily conserved **positions** in the protein*

The ConSurf tool was used to analyze the evolutionary conservation of protein residues, as shown in Figure 1. Ten SNPs showed high conservation scores; rs369062665, rs1269727956, rs902565316 and rs1043270464 are all located on functional residues. In contrast, rs953707145, rs1570073403, rs754597784, rs1332792872, rs1213084266 and rs1373684177 are located on structural residues. Only three SNPs (rs1570073417, rs755725663 and rs146884691) showed intermediate conservation scores (Table 3).

Predicting the secondary structures within the C1qA chain

The SOPMA tool was used to analyze the secondary structure of C1qA, as shown in Figure 2; 156 residues were associated with random coils (63.67%), 64 were associated with extended strands (26.12%), 15 with alpha helices (6.12%), and 10 with beta turns (4.08%). The alignment of SNPs within the secondary structure of the protein was also analyzed (Figure S2). Seven SNPs were identified in the extended strand (rs754597784, rs1570073417, rs755725663, rs146884691, rs1332792872, rs1213084266 and rs1373684177). Four SNPs were identified in the random coil (rs369062665, rs1269727956, rs902565316 and rs1043270464). Two SNPs were identified in beta turns (rs953707145 and rs1570073403). Analysis of the secondary structure with deleterious mutations is shown in Table S1;

all of these mutations led to changes in the predicted secondary structure of C1qA.

Gene–gene interactions

A predicted network for *C1qA* gene–gene interaction was generated by GeneMANIA to identify genes with strong interactions to the *C1qA* gene. Figure 3 shows the twenty genes with the strongest connections to the *C1qA gene*. Of these genes, the *C1qC* gene had the highest relatedness, followed by the *C1S gene* (complement C1s gene), *C1qB* gene and *CRP gene* (C-reactive protein gene).

Identification of the impact of SNPs on the 3D protein structure

Based on previous analysis, 10 SNPs that were located in highly conserved positions were selected for further analysis of their impact on protein structure and function using the HOPE server. For all 10 SNPs, there were differences in

Table 3: Locations of SNPs on protein domains and evolutionary conservation analysis.

SNP Id	AA change	Location on protein	ConSurf conservation score	Functional/structural	Buried/exposed
rs369062665	G31R	–	9/highly conserved	Functional	Exposed
rs1269727956	G71R	Collagen triple helix repeat	9/highly conserved	Functional	Exposed
rs902565316	G92R	Collagen triple helix repeat	9/highly conserved	Functional	Exposed
rs1043270464	G149C	C1q domain	8/highly conserved	Functional	Exposed
rs953707145	G157R	C1q domain	9/highly conserved	Structural	Buried
rs1570073403	G157D	C1q domain	9/highly conserved	Structural	Buried
rs754597784	Y159H	C1q domain	9/highly conserved	Structural	Buried
	Y159D	C1q domain	9/highly conserved	Structural	Buried
rs1570073417	T162P	C1q domain	6/intermediately conserved	–	Buried
rs755725663	W216R	C1q domain	6/intermediately conserved	–	Buried
rs146884691	I226N	C1q domain	5/intermediately conserved	–	Buried
rs1332792872	F236V	C1q domain	9/highly conserved	Structural	Buried
rs1213084266	G238S	C1q domain	9/highly conserved	Structural	Buried
rs1373684177	G238V	C1q domain	9/highly conserved	Structural	Buried

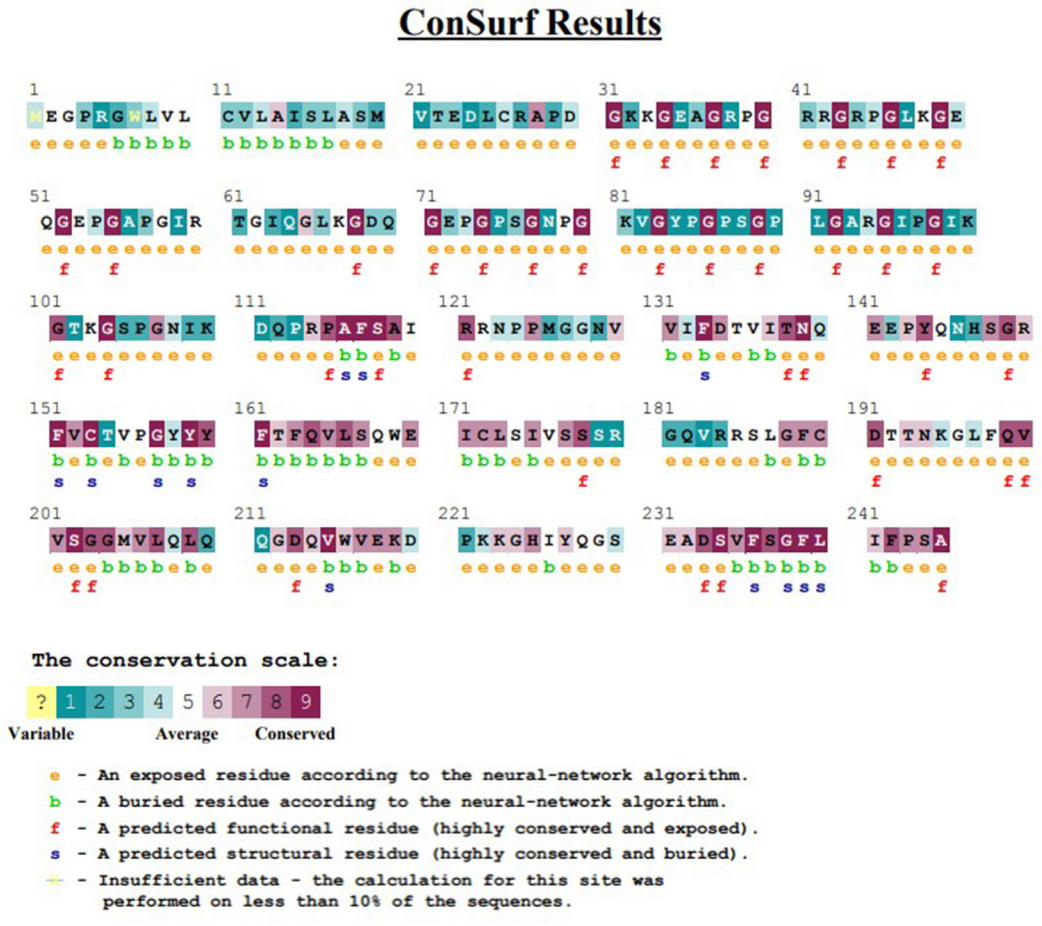


Figure 1: Evolutionary conservation analysis of C1qA using the Consurf tool.

Figure 2: Secondary structure analysis of C1qA using the SOPMA tool.

properties between mutant amino acids and wild amino acids , these differences caused different harmful impacts as bumps and repulsion with similar residues in addition to a disturbance in local structure. The G31R, G71R, G92R and G157D mutations led to differences in size and charge. G238V, G238S and G149C showed differences in size that could lead to bumps with a disturbance in the local structure. Y159H was associated with differences in size and hydrophobicity that could lead to the loss of interactions and hydrophobic interactions, respectively. F236V was associated with a difference in size that could lead to the loss of interactions. Y159D was associated with differences in charge, size and hydrophobicity that could lead to repulsion, the loss of interactions, and hydrophobic interactions, respectively.

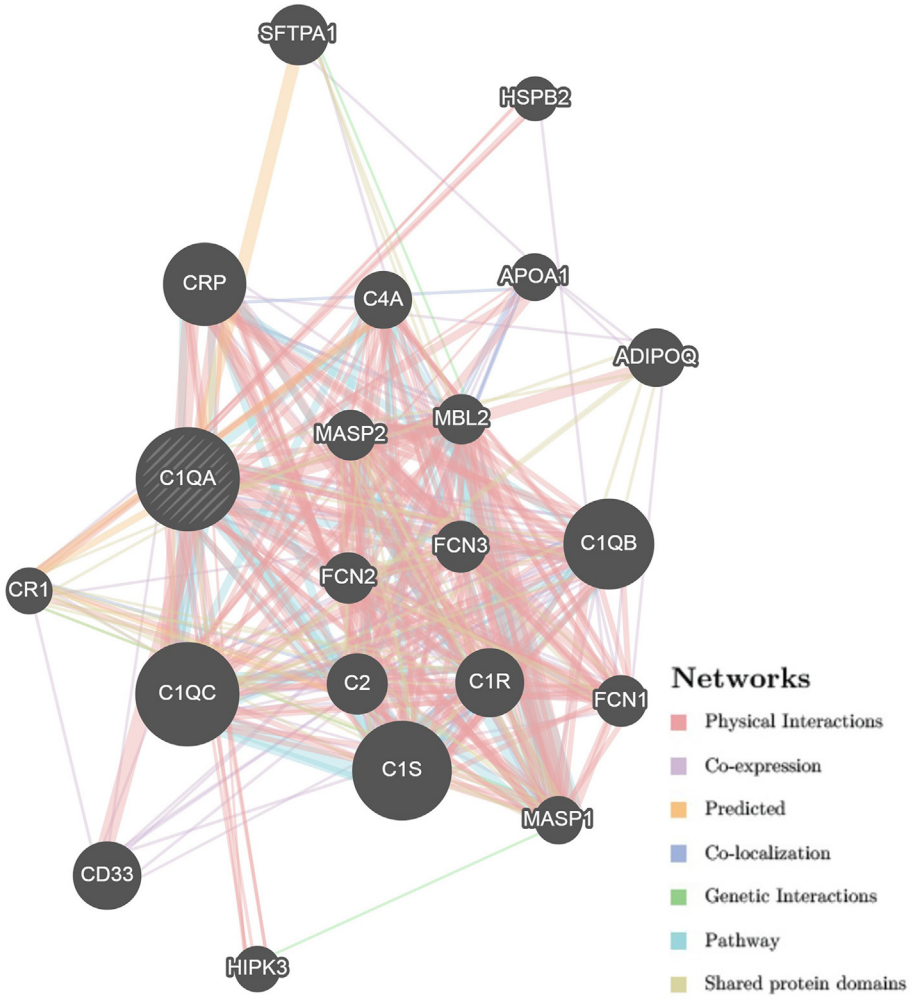


Figure 3: Network of gene–gene interactions for the C1qA gene, as generated by GeneMANIA tool.

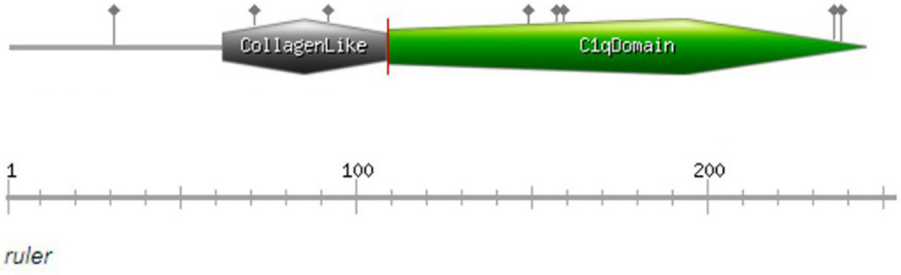


Figure 4: The positions of the 10 most deleterious mutations on their respective domains in C1qA (G157R and G157D are located at the same position; as with G238S and G238V), generated by (https://prosite.expasy.org/mydomains).

With regards to the locations of residues on the protein, all 10 SNPs resulted in different properties that could disturb their regional function. G31R was associated with an inability to compensate for the required glycine flexibility at this position for protein function. G71R and G92R were associated with different properties that could disturb the collagen triple helix repeat. The rest of the SNPs were located on the important C1q domain and exhibited different properties that could disturb this domain. Regarding the impact of SNPs on protein function, all positions showed high levels of conservation; the functionality of these positions could be damaged by the SNPs. Figure 4 demonstrates the positions of the 10 most deleterious mutations on their respective domains in C1qA.

Discussion

C1q is responsible for initiating the classical complement pathway that protects our bodies from invading pathogens and immune complexes. As a result, genetic defects in this key molecule, especially in *C1qA*, have been found to be associated with a liability to infection and sepsis.[3] Furthermore, approximately 90% of patients with a genetic deficiency in C1q eventually develop SLE, thus highlighting the role of C1q genetic defects in developing autoimmune disease and SLE.[21] Therefore, it is necessary to screen the clinically important missense SNPs in *C1qA* by *in silico* methods to identify the most deleterious missense SNPs.

To predict deleterious SNPs, we used different *in silico* tools (with different algorithms and approaches) to ensure the accuracy and strength of our results. All 184 missense SNPs found in *C1qA* by the NCBI database were analyzed by VEP, SIFT and PolyPhen-2 tools and then by four additional tools (PROVEAN, SNP&GO, PHD-SNP, SNAP); these analyses identified 20 SNPs that were designated as deleterious or disease-causing by all of the tools used. As the function and structure of proteins are critically dependent on their stability,[22] the effect of these 20 damaging SNPs on protein stability were analyzed. Thirteen missense SNPs were found to reduce protein stability and were selected for further analysis.

Then, we performed functional analysis to determine important regions and domains of the C1qA protein and the positions of the identified SNPs in these regions. We found that 10 SNPs were located on the important C1q domain of the A-chain that makes up the recognition head of the C1q molecule along with the C-terminal heads of the B and C chains; this recognition head is responsible for recognizing invading pathogens.[23,24] Moreover, two other SNPs were found to be located in the important collagen triple helix repeat which is believed to be crucial to C1q complex formation.[24] Mutations in these important regions are expected to affect protein function.

Functionally important amino acids show high degrees of evolutionary conservation.[16] Thus, mutations of these highly conserved residues are predicted to affect protein function. Ten SNPs were identified to be located on highly conserved functional or structural residues, thus indicating the high potentiality of a damaging effect on the protein. Then, the predicted secondary structure of C1qA was analyzed in the wild type and with different deleterious mutations. All of these mutations were found to result in changes in the predicted secondary structure of C1qA. The secondary structure of proteins has significant importance for their structure and protein folding[25] due to the important roles of the secondary structure in building the starting cores required for creating whole protein folding.[26,27]

The proven presence of an interaction between different genetic loci revealed the importance of analyzing gene−gene interactions when investigating the association of genes with disease.[28] In the present study, GeneMANIA analysis identified the genes that interacted most strongly; the genes that interacted most strongly were (in order): *C1qC*, *C1S*, *C1qB* and *CRP*. These genes encode important factors in the immune system and in the complement system; *C1qC* and *C1qB* genes encode C1q along with the *C1qA* gene.[1] The *C1S* gene encodes C1s that compose the C1 complex of the complement system along with C1r and C1q (https://www.ncbi.nlm.nih.gov/gene/716). *CRP* gene encodes C-reactive protein with its important roles in host defense including complement system activation and regulation (https://www.ncbi.nlm.nih.gov/gene/1401). These genes had the closest connection to *C1qA* and may be affected by its mutations, which could lead to greater impact on immune response and complement activation. Finally, the 10 highly damaging SNPs located on highly conserved residues were applied to further analysis using the HOPE server. All these SNPs were predicted to cause defects in protein structure and function.

As a result, 10 SNPs (rs369062665, rs1269727956, rs902565316, rs1043270464, rs953707145, rs1570073403,

rs754597784, rs1332792872, rs1213084266 and rs1373684177) were predicted to be the most deleterious missense SNPs in the *C1qA* gene and have the highest risk to induce infection, sepsis and SLE diseases.

Conclusion

In total, 184 missense SNPs were identified in the *C1qA* gene; 10 of these SNPs were predicted to have the most damaging effect on protein function, structure, and stability by *in silico* tools. Moreover, all these SNPs were located on highly conserved functional and structural positions. These SNPs are the best candidates for conducting further experimental studies to validate these results and allow the identification of people with a high risk of infection, sepsis and SLE.

Source of funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not for-profit sectors.

Conflict of interest

The authors have no conflict of interest to declare.

Ethical approval

This was an *in silico* bioinformatics study that did not include any human or animal participants. Hence, no IRB Ethical approval was needed.

Authors contributions

MYB (investigation, conceptualization, methodology, wrote the original draft of the manuscript); MMA, AAA and HYA (supervision, conceptualization, methodology, co-wrote the original draft of the manuscript), EAI and AAS (supervision, methodology, writing, reviewing and editing). All authors have critically reviewed and approved the final draft and are responsible for the content and similarity index of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jtumed.2022.04.014.

References

1. Bruiners N, Schurz H, Daya M, Salie M, van Helden PD, Kinnear CJ, et al. A regulatory variant in the C1Q gene cluster is associated with tuberculosis susceptibility and C1qA plasma levels in a South African population. **Immunogenetics 2020 Jul**; 72(5): 305–314. https://doi.org/10.1007/s00251-020-01167-5.
2. Bohlson SS, O'Conner SD, Hulsebus HJ, Ho MM, Fraser DA. Complement, c1q, and c1q-related molecules regulate macrophage polarization. **Front Immunol 2014 Aug 21**; 5: 402. https://doi.org/10.3389/fimmu.2014.00402.
3. Skattum L, van Deuren M, van der Poll T, Truedsson L. Complement deficiency states and associated infections. **Mol Immunol 2011 Aug**; 48(14): 1643–1655. https://doi.org/10.1016/j.molimm.2011.05.001.
4. Chasman D, Adams RM. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. **J Mol Biol 2001 Mar 23**; 307(2): 683–706. https://doi.org/10.1006/jmbi.2001.4510.
5. Brown JS, Hussell T, Gilliland SM, Holden DW, Paton JC, Ehrenstein MR, et al. The classical pathway is the dominant complement pathway required for innate immunity to *Streptococcus pneumoniae* infection in mice. **Proc Natl Acad Sci U S A 2002 Dec 24**; 99(26): 16969–16974. https://doi.org/10.1073/pnas.012669199.
6. Cao CW, Li P, Luan HX, Chen W, Li CH, Hu CJ, et al. Association study of C1qA polymorphisms with systemic lupus erythematosus in a Han population. **Lupus 2012 Apr**; 21(5): 502–507. https://doi.org/10.1177/0961203311430702.
7. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. **Nucleic Acids Res 2012 Jul**; 40(Web Server issue): W452–W457. https://doi.org/10.1093/nar/gks539.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. **Nat Methods 2010 Apr**; 7(4): 248–249. https://doi.org/10.1038/nmeth0410-248.
9. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. **Curr Protoc Hum Genet 2013 Jan**. https://doi.org/10.1002/0471142905.hg0720s76. Chapter 7:Unit7.20.
10. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. **Bioinformatics 2015 Aug 15**; 31(16): 2745–2747. https://doi.org/10.1093/bioinformatics/btv195.
11. Capriotti E, Calabrese R, Fariselli P, Martelli PL, Altman RB, Casadio R. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. **BMC Genom 2013**; 14(Suppl. 3): S6. https://doi.org/10.1186/1471-2164-14-S3-S6.
12. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. **Bioinformatics 2006 Nov 15**; 22(22): 2729–2734. https://doi.org/10.1093/bioinformatics/btl423.
13. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. **BMC Genom 2015**; 16(Suppl. 8): S1. https://doi.org/10.1186/1471-2164-16-S8-S1.
14. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. **Nucleic Acids Res 2005 Jul 1**; 33(Web Server issue): W306–W310. https://doi.org/10.1093/nar/gki375.
15. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. **Nucleic Acids Res 2021 Jan 8**; 49(D1): D344–D354. https://doi.org/10.1093/nar/gkaa977.
16. Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, et al. ConSeq: the identification of functionally and structurally important residues in protein sequences. **Bioinformatics 2004 May 22**; 20(8): 1322–1324. https://doi.org/10.1093/bioinformatics/bth070.
17. Ashkenazy H, Abadi S, Martz E, Chay O, Mayrose I, Pupko T, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. **Nucleic Acids Res 2016 Jul 8**; 44(W1): W344–W350. https://doi.org/10.1093/nar/gkw408.
18. Geourjon C, Deléage G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction

from multiple alignments. **Comput Appl Biosci 1995 Dec**; 11(6): 681–684. https://doi.org/10.1093/bioinformatics/11.6.681.
19. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. **Nucleic Acids Res 2010 Jul**; 38(Web Server issue): W214–W220. https://doi.org/10.1093/nar/gkq537.
20. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. **BMC Bioinf 2010 Nov 8**; 11: 548. https://doi.org/10.1186/1471-2105-11-548.
21. Son M, Diamond B, Santiago-Schwarz F. Fundamental role of C1q in autoimmunity and inflammation. **Immunol Res 2015 Dec**; 63(1–3): 101–106. https://doi.org/10.1007/s12026-015-8705-6.
22. Deller MC, Kong L, Rupp B. Protein stability: a crystallographer's perspective. **Acta Crystallogr F Struct Biol Commun 2016 Feb**; 72(Pt 2): 72–95. https://doi.org/10.1107/S2053230X15024619.
23. Eggleton P, Tenner AJ, Reid KB. C1q receptors. **Clin Exp Immunol 2000 Jun**; 120(3): 406–412. https://doi.org/10.1046/j.1365-2249.2000.01218.x.
24. Reid KBM. Complement component C1q: historical perspective of a functionally versatile, and structurally unusual, serum protein. **Front Immunol 2018 Apr 10**; 9: 764. https://doi.org/10.3389/fimmu.2018.00764.
25. Ji YY, Li YQ. The role of secondary structure in protein structure selection. **Eur Phys J E Soft Matter 2010 May**; 32(1): 103–107. https://doi.org/10.1140/epje/i2010-10591-5.
26. Colubri A. Prediction of protein structure by simulating coarse-grained folding pathways: a preliminary report. **J Biomol Struct Dyn 2004 Apr**; 21(5): 625–638. https://doi.org/10.1080/07391102.2004.10506953.
27. Cao M, Shi L, Peng P, Han B, Liu L, Lv X, et al. Determination of genetic effects and functional SNPs of bovine HTR1B gene on milk fatty acid traits. **BMC Genom 2021 Jul 27**; 22(1): 575. https://doi.org/10.1186/s12864-021-07893-8.
28. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. **Nat Rev Genet 2009 Jun**; 10(6): 392–404. https://doi.org/10.1038/nrg2579.

How to cite this article: Behairy MY, Abdelrahman AA, Abdallah HY, Ibrahim EEl-DeenA, Sayed AA, Azab MM. *In silico* analysis of missense variants of the *C1qA* gene related to infection and autoimmune diseases. J Taibah Univ Med Sc 2022;17(6):1074–1082.