American Journal of Bioinformatics Research 2018, 8(1): 1-11
DOI: 10.5923/j.bioinformatics.20180801.01

Computational Analysis of Deleterious Single Nucleotide Polymorphisms (SNPs) in Human CALR Gene

Sahar G. Elbager[1,*], Hadeel F. Gad Mahgoob[2], Mohamed Y. Basher[3], Asia M. Elrashid[4], Hazem Abdo[3], Shazalia K. Babiker[1], Ali I. Alsaid[5], Mohamed A. I. Alfaki[6], Safinaz I. Khalil[7], Amar A. Dowd[1], Magdi A. Bayoumi[1]

[1]Faculty of Medical Laboratory Sciences, University of Medical Sciences and Technology (UMST), Sudan
[2]Faculty of Science, University of Princess Nourah bint Abdul Rahman, Saudi Arabia
[3]Faculty of Veterinary Medicine, University of Khartoum, Sudan
[4]Faculty of Science, University of Khartoum, Sudan
[5]Faculty of Science and Technology, Omdurman Islamic University, Sudan
[6]Faculty of Computer Science, Neelain University, Sudan
[7]Faculty of Medicine, University of Medical Sciences and Technology (UMST), Sudan

Abstract **Background:** The human Calreticulin (CRT) is a multifunctional protein encoded by CALR gene (CALR) located on chromosome 19. Calreticulin plays an important role in protein folding and calcium homeostasis. It's also has been associated with cell proliferation, cell cycle progression and immunogenic cell death. High CRT serum levels were reported in various cancer like prostate cancer and breast carcinoma. Being observed as an important molecule in biological responses and its association with various diseases, we aimed to systematically explore the probable effects of CALR genetic variants on functions and structure of calreticulin using in silico prediction softwares. **Methods:** The data on human CALR gene was retrieved from dbSNP/NCBI. Eleven different prediction algorithms; SIFT, Polyphen, PROVEAN, SNAP2, Condel, Pmut, nsSNPs Analyzer, PhD-SNP, I-Mutant, Mutpred and Project Hope were used to analyzing the effect of nsSNPs on functions and structure of the CRT protein. STRING and KEGG database were used for CRT protein-protein interaction. **Results:** As per dbSNP database, the human CALR gene investigated in this work contained a total of 682SNPs:53 SNPs in 3′ UTR region, 25 SNPs in 5′ UTR region, 343 SNPs in intron region, 103 SNPs in coding synonymous regions and 154 non-synonymous SNPs (nsSNPs) which comprises of 150 missense mutations, 3 frameshift mutations and one nonsense mutation. We selected missense nsSNPs for our investigation. A total of 4 nsSNPs P216L, R73C, W261G and Y128C were predicted to have the most damaging effects on CRT protein's structure and function. STRING and KEGG revealed that CRT protein had strong interactions with proteins that involved in protein processing and presentation networks. Therefore, any structural alterations in the CRT protein that interfere or harm these networks interactions would probably increase susceptibility to diseases. **Conclusion:** Based on these analyses, the present study suggested that the reported functional SNPs may act as potential targets in genetic association studies.

Keywords CALR gene, Calreticulin, Computational analyses, Deleterious SNPs, Regulatory SNPs, Single nucleotide polymorphism (SNP)

1. Introduction

The human Calreticulin (CRT) is a 46 KDa multifunctional protein predominantly located in endoplasmic reticulum (ER) [1] and it is also known as calregulin, CaBP3, CRP55, calsequestrin-like protein, and endoplasmic reticulum resident protein 60 (ERp60). It is encoded in humans by CALR gene (CALR) located on chromosome 19p13.13 [1].

CRT molecule consists of three domains, including N-domain, P-domain, and C-domain. The N-terminal region of CRT is encoded by a highly conserved sequence that is folded in a stable globular domain containing eight antiparallel β-strands [2]. The N-terminal is an important functional domain that includes polypeptide and carbohydrate-binding sites, Zn^{2+} binding site and DNA-binding site of steroid receptor [3]. The disulfide bond formed by cysteine residues within this domain may interact with P-domain to produce important chaperone function of calreticulin [4]. The Proline-rich P-domain contains two sets of three repetitive amino acid sequence

* Corresponding author:
saharelbagir@ymil.com (Sahar G. Elbager)

Published online at http://journal.sapub.org/bioinformatics

regions. These repeated amino acid sequences form the lectin-like chaperone structures which are responsible for protein-folding function of CRT [4]. In addition, the P-domain comprises a region with a high-affinity and low-capacity Ca^{2+}-binding region [5]. The C-domain of CRT is a highly acidic region [4]. It binds to Ca^{2+} with high capacity and low affinity [4]. This terminal contains a KDEL sequence [lysine, aspartate, glutamate and leucine] ER retention motif which prevents CRT from being secreted from the ER [5]. These three domains are illustrated in Figure 1.

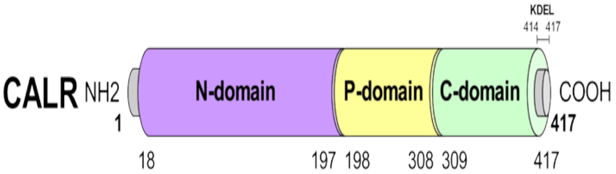

Figure 1. Schematic primary structure of CALR protein. The conserved P-domain, N-domain and C-domain and KDEL motif are illustrated. Amino acid (aa) positions are indicated [6]

Since calreticulin has neither a transmembrane domain nor a GPI-anchor attachment site, it is anchored to the membranes through specific adaptors such as the receptors for thrombopoietin, erythropoietin or granulocyte colony stimulating factor (G-CSF) in hematopoietic cells [7, 8] or CD59 in neutrophils [9].

The biological functions of CRT protein are related to its localization. It is found either intracellularly in the ER or cytoplasm and extracellularly on cell surface. Intracellularly CRT was reported to be involved in protein folding [10], calcium homeostasis and regulation of p53 expression and it interacted with other chaperones as CANX (calnexin) [5]. Extracellularly, CRT has been found to be involved in modulation of gene expression, nuclear export, cardiogenesis and immunogenic cell death [10] cell adhesion [11] and RNA stability [12]. This suggested an involvement of CRT in various biological and pathological processes.

Many studies revealed that CRT expression might play a crucial role during cancer progression. Over expression has been reported in oral cancer [13], breast carcinoma [14], colorectal cancer [15], prostate cancer [16] and vaginal carcinoma [17]. Reduced CRT expression was observed in malignant effusions of high-grade ovarian carcinoma along disease progression [18].

Recently, somatic frameshift mutations in CALR gene were detected in essential thrombocythemia and the primary myelofibrosis lacking JAK2 or MPL [19, 20]. These CRT mutations included 52 bp deletion and 5 bp insertion of certain base pairs, lead to frameshift mutations [21]. Both types of mutations resulted in the formation of a novel mutant C-terminal which lack the KDEL motif. The mutant CALR protein, through its interaction with the thrombopoietin receptor MPL, activates the JAK-STAT pathway which is the hallmark of these disorders [7, 8]. Furthermore, normal Ca^{2+} binding and cell growth might be affected [20, 22]. In addition to MPNs, a CALR mutations were described in patient's other hematopoietic diseases such as leukemia and MDS [23-25].

Few studies addressing CALR mutations from a bioinformatics point of view have been published [26]. Nonetheless, out of them do not specifically focus on a single nucleotide polymorphisms (SNPs).

Recently, bioinformatics resources (data bases and software) that facilitate the extraction of useful results from large amounts of raw data like analysis of gene and protein expression, comparison of genetic data, modeling of DNA and protein structures and aids prediction of deleterious SNPs and its association with diseases.

In the present study we aimed to determine the influence of various SNPs in CALR gene using in silico prediction software and assessing the effect of these SNPs on the structure and function of CRT protein that may have an important role in disease susceptibility.

2. Material and Method

2.1. Data Set

The data on human CALR gene (accession ID: NP_004334) was retrieved from the Entrez Gene databases from National Center for Biological Information (NCBI) database on 13 March 2018. The CRT protein sequence (accession ID: P27797) and SNPs information of the CALR gene were obtained from UniProtKB databases (http://www.uniprot.org) and NCBI dbSNP (http://www.ncbi.nlm.nih.gov/snp/) respectively.

2.2. Identification of Deleterious nsSNPs

To determine the functional impact (deleterious, damaging or natural), coding nsSNPs were analyzed using five different tools (SIFT, Polyphen -2, PROVEAN, SNAP2 and Condel). nsSNPs predicted to be deleterious by these five tools that were categorized as high-risk nsSNPs were subjected for further analysis like association with disease, stability analysis and structural effect using different tools. STRING and KEGG database were used for CRT protein-protein interaction.

2.2.1 SIFT (Sorting intolerant from tolerant; http://siftdna.org/www/SIFT_dbSNP.html) server was used to identify the tolerated and deleterious SNPs. The effect of amino acid substitution on protein structure was assessed on the basis of degree of conservation of amino acids using sequence homology Substitution of an amino acid at each position with probability < 0.05 is predicted to be deleterious and intolerant, while probability ≥ 0.05 is considered as tolerant [27].

2.2.2 Polyphen -2 (polymorphism and phenotype; http://genetics.bwh.harvard.edu/pph2/) server was used to predict the functional impact of amino acid substitution on protein structure and function based on sequence based characterization. The Prediction outcome was obtained in the form of probability score which classifies the variations as 'probably damaging', 'possibly damaging' and 'benign' [28]. UniProt protein accession ID P27797 along with position and name of wild type and variant amino acids of screened nsSNPs were submitted as query.

2.2.3 PROVEAN (Protein Variation Effect Analyzer; http://provean.jcvi.org/index.php) is used to predict the possible impact of a substituted amino acid and indels on protein structure and biological function. It analyses the nsSNPs as deleterious or natural, if the final score was below the threshold score of −2.5 were considered deleterious; scores above this threshold were considered neutral [29]. The input query is a protein FASTA sequence along with amino acid substitutions.

2.2.4 SNAP2 (Screening of Non-Acceptable Polymorphism 2; https://rostlab.org/services/snap2web/) is a tool, developed based on a neural network classification method which is freely available. It predicts the effect of nsSNPs on protein function [30].The input query submitted is the protein FASTA sequence and lists of mutants which provided scores of each substitution that can then be translated into binary predictions neutral or non-neutral effect natural.

2.2.5 Condel (CONsensus DELeteriousness score of missense SNVs; http://bbglab.irbbarcelona.org/fannsdb/query/condel). Condel is a method, used assess the outcome of non-synonymous SNVs, via applying a CONsensus DELeteriousness score that combines various tools. It computes a weighted approach of missense mutations from the complementary cumulative distributions of scores of deleterious and neutral mutations [28]. The input query submitted is rhe UniProt protein accession ID P27797 along with the position and name of wild type and variant amino acids of screened nsSNPs.

2.3. Prediction of Disease Associated SNPs

The SNPs occurring in the protein coding region may lead to the deleterious consequences in its 3D structure and thus may lead to disease-associated phenomena. Here nsSNPAnalyzer, Pmut and PhD- SNP were used as tools to examine the disease-associated nsSNP occurring in the CALR protein coding region.

2.3.1 nsSNPAnalyzer (http://snpanalyzer.uthsc.edu) is a tool, employed to predict the phenotypic effect (disease-associated vs. neutral) of a nsSNP by using a machine learning method called Random Forest, and extracting structural and evolutionary information from a query nsSNP [31]. The input was provided in form of protein FASTA sequence along with amino acid substitution.

2.3.2 Pmut (http://mmb2.pcb.ub.es:8080/PMut) web server was used to predict disease-associated mutations. PMUT method is based on the use of neural networks (NNs) which is trained on large database of neutral mutations (NEMUs) and pathological mutations of mutational hot spots, which are obtained by alanine scanning, massive mutation, and genetically accessible mutations. The final output is displayed as a pathogenicity index ranging from 0 to 1 (indexes > 0.5 single pathological mutations) [32]. UniProt accession ID of CRT (P27797) along with name and position of wild type and mutant amino acid was submitted as an input for this server.

2.3.3 PhD-SNP (Predictor of human Deleterious Single Nucleotide Polymorphisms; http://snps.biofold.org/phd-snp/phd-snp.html) web server was used to predict if a given single point protein mutation can be classified as a disease-related or as neutral polymorphism. This server was mainly based on the support vector machines which can corroborate all the information regarding variations from the existing databases [33]. The input FASTA sequences of protein along with the residues change were submitted to PhD-SNP server for the analysis.

2.4. Prediction of nsSNPs Impact on the Protein Stability by I-Mutant2.0

I-Mutant2.0 is a tool used for prediction of changes in protein stability due to single site mutations under different conditions (http://gpcr.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi). It is a web server based on support vector machine which worked on dataset derived from Protherm, a database of experimental records on protein mutations. It can predict the stability changes in protein with 80% accuracy based on its structure and with 77% of accuracy based on its sequence [34]. The input can be submitted as either in the form of protein sequence or on a structure basis. For the present study, input was submitted in the form of protein FASTA sequence.

Analyzing the effect of nsSNPs on physiochemical properties of the proteins.

2.5. Analyzing the Effect of nsSNPs on 3D Structure of the Proteins and Physiochemical Properties

2.5.1 Project Hope server (http://www.cmbi.ru.nl/hope/) was used to search protein 3D structures by collecting structural information from a series of sources, including calculations on the 3D coordinates of the protein, sequence annotations from the UniProt database, and predictions by DAS services [35]. Furthermore, it described the reaction and physiochemical properties of these candidates. Protein sequence and mutant variants were submitted to project hope server in order to analyze the structural and conformational variations that have resulted from single amino acid substitution.

2.5.2 Mutpred was used to predict structural and functional changes as a consequence of amino acid substitution. (http://mutpred.mutdb.org/). These changes were expressed as probabilities of gain or loss of structure and function. In addition, it predicts molecular cause of disease. The MutPred output contains a general score (g), i.e., the probability that the AAS is deleterious/disease-associated and top five property scores (p), where p is the P-value that certain structural and functional properties are impacted. A missense mutation with a MutPred (g) score > 0.5 could be considered as "harmful," while a (g) score > 0.75 should be considered a high confidence "harmful" prediction [36]. The input was submitted in the form of protein FASTA sequence along with amino acid substitution.

2.6. Prediction of Genetic and Protein Interactions

Protein-protein interactions are important to assess all functional interactions among cell proteins.

2.6.1 STRING (Search Tool for the Retrieval of Interacting Genes/Proteins; https://string-db.org/) [37]. The STRING database gives a protein- protein interaction either it is direct or indirect associations. The input option we use is the protein name and the organism.

2.6.2 KEGG (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/) [38] is a knowledge base for systematic analysis of gene functions in terms of the networks of genes and molecules, including metabolic pathways, regulatory pathways, and molecular complexes for biological systems.

3. Results

3.1. SNP Dataset from dbSNP

As per dbSNP database, the human CALR gene investigated in this work contained a total of 682SNPs: 53 SNPs in 3′ UTR region, 25 SNPs in 5′ UTR region, 343 SNPs in intron region, 103 SNPs in coding synonymous regions and 154 non-synonymous SNPs (nsSNPs) that comprises of 150 missense mutations, 3 frameshift mutations and one nonsense mutation. The missense nsSNPs were selected for the investigation.

3.2. Prediction of Functional Mutations

A total number of 150 nsSNPs CALR gene were submitted as batch to the SIFT program. According to our SIFT analysis predicted, 26 SNPs were predicted to be deleterious, 48 SNPs were predicted to be tolerated and 76 nsSNPs were not found. Deleterious SNPs were submitted to Polyphen-2, PROVEAN, SNAP and Condel.

Polyphen-2 results analysis predicted that 9 nsSNPs were probably damaging, whereas 2 nsSNPs were predicted to be 'possibly damaging' and the remaining one nsSNPs was categorized as benign (Table 1). Our PROVEAN analysis predicted that 10 nsSNPs were deleterious and 2 nsSNPs were neutral. The SNAP2 Analyzer predicted that 9 nsSNPs were affected the function of protein and 3 nsSNPs were neutral (Table 1). Out of 26 nsSNPs predicted to be deleterious by SIFT, 12 nsSNPs were classified by the other computational tools used and 14 nsSNPs were unclassified.

Table 1. List of nsSNP analysis by SIFT, PolyPhen-2, PROVEAN, SNAP2, CONDEL

SNP ID	AA CHANGE	SIFT PREDICTION	SCORE	POLY-2 PREDICTION	SCORE	PROV	SCORE	SNAP2	SCORE	CONDEL LABEL	Score
rs11547563	**W261G**	**Deleterious**	**0**	**Probably Damaging**	**1**	**Deleterious**	**-12.24**	**Effect**	**82**	**Deleterious**	**0.81**
rs11547569	V176M	Deleterious	0.03	Probably Damaging	0.99	Neutral	-2.33	Neutral	-6	Neutral	0.59
rs138503495	**Y57C**	**Deleterious**	**0**	**Probably Damaging**	**1**	**Deleterious**	**-7.8**	**Effect**	**29**	**Deleterious**	**0.69**
rs149740908	P228S	Deleterious	0.01	Possibly Damaging	0.79	Deleterious	-7.17	Neutral	-8	Deleterious	0.71
re199565419	D199N	Deleterious	0.05	Probably Damaging	1	Deleterious	-4.81	Neutral	-3	Deleterious	0.62
rs200173883	**P216L**	**Deleterious**	**0**	**Probably Damaging**	**1**	**Deleterious**	**-9.83**	**Effect**	**20**	**Deleterious**	**0.81**
rs200421584	**Y128C**	**Deleterious**	**0**	**Probably Damaging**	**1**	**Deleterious**	**-8.8**	**Effect**	**82**	**Deleterious**	**0.77**
rs367922867	**G273R**	**Deleterious**	**0.01**	**Probably Damaging**	**1**	**Deleterious**	**-7.91**	**Effect**	**22**	**Deleterious**	**0.8**
rs370585408	**R73C**	**Deleterious**	**0**	**Probably Damaging**	**1**	**Deleterious**	**-6.73**	**Effect**	**79**	**Deleterious**	**0.7**
rs373993748	**T90M**	**Deleterious**	**0.03**	**Probably Damaging**	**1**	**Deleterious**	**-3.85**	**Effect**	**33**	**Deleterious**	**0.68**
rs375135016	D302N	Deleterious	0.01	Possibly Damaging	0.93	Deleterious	-4.78	Effect	14	Neutral	0.65
rs375318260	E127D	Deleterious	0.04	Benign	0	Neutral	-0.9	Effect	23	Neutral	0.45

Because each computational method uses different parameters to evaluate the nsSNPs, the accuracy of the deleterious SNPs can be increased by combining different computational methods. 7 nsSNPs (W261G, Y57C, P216L, Y128C, G273R, R73C, T90M) were filtered and they were combinedly predicted to be deleterious and damaging by SIFT, Polyphen-2, PROVEAN, SNAP2 and CONDEL. Out of 26 nsSNPs, 7 nsSNPs were predicted as deleterious by those five servers (Table 1).

3.3. Prediction of Disease Associated SNPs

Pmut, nsSNPAnalyzer and PhD-SNP were used to validate the results obtained from five previous tools. Out of 7 nsSNPs that predicted to be deleterious with SIFT, Polyphen, SNAP2, PROVEAN and CONDEL; Pmut predicted all 7 nsSNP to be associated with disease, nsSNPAnalyzer predicted 6 nsSNP while nsSNPAnalyzer predicted 5 nsSNP to be associated with disease (Table 2).

Finally, out of 26 nsSNP, 4 nsSNPs namely rs11547563 (W261G), rs200173883 (P216L), rs200421584 (Y128C) and rs370585408 (R73C) were found to be highly significant deleterious by (SIFT, Polyphen, PROVEAN, SNAP2, CONDEL, Pmut, nsSNPAnalyzer and PhD-SNP) servers.

Table 2. List of nsSNP predicted as disease associated by Pmut, nsSNPA alyzer and PhD-SNP

SNP ID	AA Change	Pmut	nsSNP Analyzer	PhD-SNP
rs11547563	**W261G**	**Disease**	**Disease**	**Disease**
rs138503495	Y57C	Disease	Neutral	Disease
rs200173883	**P216L**	**Disease**	**Disease**	**Disease**
rs200421584	**Y128C**	**Disease**	**Disease**	**Disease**
rs367922867	G273R	Disease	Disease	Neutral
rs370585408	**R73C**	**Disease**	**Disease**	**Disease**
rs373993748	T90M	Disease	Disease	Neutral

3.4. Prediction of nsSNPs Impact on the Protein Stability

All SNPs predicted to be associated with disease were submitted to I- mutant server to predict the impact of this SNPs on the protein Stability. The results were as follows: 3 SNPs (W261G, Y128C, R73C) were predicted to decrease effective stability of the CRT protein and one SNP (P216L) was predicted to increase the stability of CRT protein (Table 3).

Table 3. Prediction result of I-Mutant software

SNP ID	AA Change	Stability Prediction	Reliability Index (RI)	Ph	Temp
rs11547563	W261G	Decrease	8	7	25
rs200173883	P216L	Increase	1	7	25
rs200421584	Y128C	Decrease	1	7	25
rs370585408	R73C	Decrease	0	7	25

3.5. Analysing the Effect of nsSNPs on 3D Structure of the Proteins and Physiochemical Properties

3.5.1 3D structural modelling by Project Hope server

rs11547563 (W261G) The residue is located in the P-domain of the protein; this domain is important for binding of other molecules and in contact with residues in N-terminal region domain that is also important for binding. Mutation of this residue might disturb the interaction between these two domains and as such affect the function of the protein. The mutation introduces a glycine at this position. Glycine are very flexible and can disturb the required rigidity of the protein at this position. The mutant residue is smaller and less hydrophobic than the wild-type residue. The wild-type residue forms a hydrogen bond with Valine at position 50. The size difference between wild-type and mutant residue makes that the new residue that is not in the correct position to make the same hydrogen bond as the original wild-type residue did. Moreover, this difference in hydrophobicity will affect hydrogen bond formation. (Figure 2).

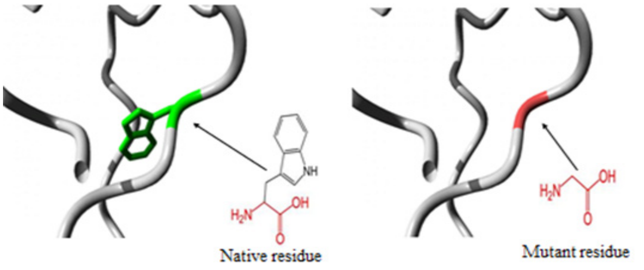

Figure 2. Close-up of the mutation. The protein is colored grey, the side chains of both the wild-type and the mutant residue are shown and colored green and red respectively. SNP ID: rs28989182, protein position 261changed from Tryptophan to Glycine

rs200173883 (P216L) The residue is located on the surface of the protein, the mutation of this residue can disturb interactions with other molecules or other parts of the protein. The wild-type residue is a proline. Prolines are known to be very rigid and therefore induce a special backbone conformation which might be required at this position. The mutation can disturb this special conformation. The wild-type residue is very conserved, neither the mutant residue nor another residue type with similar properties was observed at this position in other homologous sequences. Based on conservation scores this mutation is probably damaging to the protein (Figure 3).

rs200421584 (Y128C) The mutation is located in the N-domain of the protein; the differences in amino acid properties can disturb this region and disturb its function. The mutant residue is smaller and more hydrophobic than the wild-type residue than the wild-type residue. The mutation will cause an empty space in the core of the protein, in addition to the loss of hydrogen bonds in the core of the protein and as a result disturb correct folding. Only this residue type was found at this position. Mutation of a

100% conserved residue is usually damaging for the protein. The mutant and wild-type residue are not very similar. This means that this mutation is probably damaging to the protein (Figure 4).

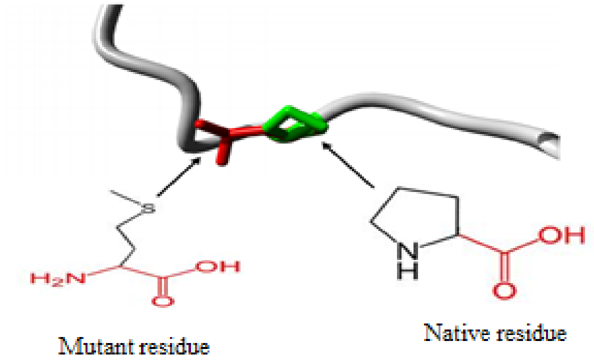

Figure 3. Close-up of the mutation. The protein is coloured grey, the side chains of both the wild-type and the mutant residue are shown and coloured green and red respectively. SNP ID: rs200173883, protein position 216 changed from Proline to Leucine

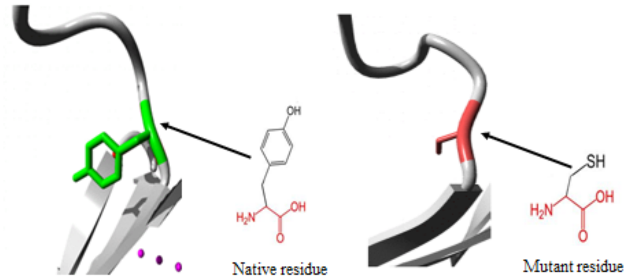

Figure 4. Close-up of the mutation. The protein is colored grey, the side chains of both the wild-type and the mutant residue are shown and colored green and red respectively. SNP ID: rs200421584, protein position 128 changed from Tyrosine to Cysteine

rs370585408 (R73C) The residue is located in the N-domain of the protein; the differences in amino acid properties can disturb this region and disturb its function. The wild-type residue charge was positive and the mutant residue charge is neutral. The wild-type residue forms a salt bridge with Aspartic Acid at position 45. The charge of the wild-type residue is lost by this mutation, and it consequently disturbs the ionic interaction made by the original, wild-type residue. The mutant residue is smaller and more hydrophobic than the wild-type residue. The wild-type residue forms a hydrogen bond with Aspartic Acid at position 45. The difference in size and hydrophobicity and between wild-type and mutant residue will affect the hydrogen bond formation (Figure 5).

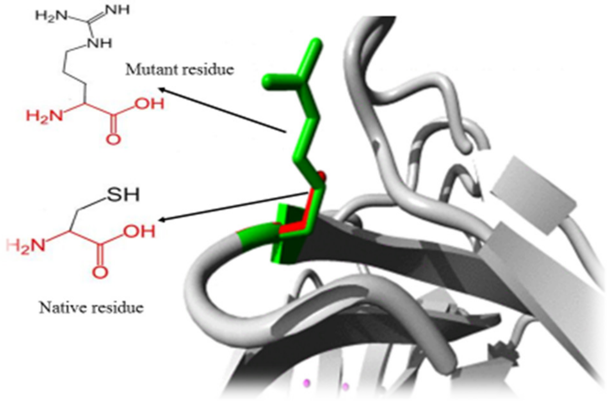

Figure 5. Close-up of the mutation. The protein is colored grey, the side chains of both the wild-type and the mutant residue are shown and colored green and red respectively. SNP ID: rs370585408, protein position 73 changed from Arginine to Cysteine

3.5.2 Analysing the effect of nsSNPs on the protein physiochemical properties by MutPred

MutPred result showed (W261G, Y128C, R73C) were highly harmful and P216L was harmful. The possible molecular mechanism disrupted were predicted; loss of catalytic residue at P264 (P = 0.0417) for the mutation W261G and loss of phosporylation at Y128 (P = 0.0194) for the mutation Y128C. Table 4 summarizes the result obtained from MutPred server.

3.6. Protein-Protein Interactions Analysis

STRING interaction analysis revealed that CALR gene involved in many molecular and biological process and has high-confidence interactions with Heat shock protein 90kDa beta member 1 (HSP90B1), Protein disulfide isomerase family A member 3 (PDIA3), Heat shock 70kDa protein 5 (HSPA5), Prolyl 4-hydroxylase, beta polypeptide (P4HB), Glucosidase, alpha; neutral AB (GANAB), Protein kinase C substrate 80K-H (PRKCSH), TAP binding protein (TAPBP), Apolipoprotein B (APOB), Activating transcription factor 6 (ATF6) and Beta-2-microglobulin (B2M) (Figure 6).

KEGG results demonstrate that CRT protein involved in the antigen processing and presentation network by MHC class I molecules (Figure 7) as well as the protein processing pathways in endoplasmic reticulum (Figure 8).

Table 4. Prediction of functional effects of nsSNPs using MutPred

Amino acid change	Probability of deleterious mutation	Top 5 features	Hypotheses of molecular mechanism disrupted
W261G	0.810 (Highly harmful)	Loss of catalytic residue at P264 (P = 0.0417) Loss of helix (P = 0.3949) Loss of stability (P = 0.4169) Loss of sheet (P = 0.437) Gain of loop (P = 0.6993)	Loss of catalytic residue at P264 (P = 0.0417)
P216L	0.674 (harmful)	Loss of methylation at K215 (P = 0.0778) Loss of glycosylation at S214 (P = 0.1091) Loss of ubiquitination at K215 (P =0.1172) Gain of helix (P = 0.2684) Loss of loop (P = 0.4412)	
Y128C	0.832 (Highly harmful)	Loss of phosphorylation at Y128 (P = 0.0194) Loss of disorder (P = 0.1147) Loss of sheet (P = 0.3635) Gain of catalytic residue at M131 (P = 0.4513) Loss of loop (P = 0.4786)	Loss of phosphorylation at Y128 (P = 0.0194)
R73C	0.764 (Highly harmful)	Loss of disorder (P = 0.0509) Loss of phosphorylation at Y75 (P = 0.151) Gain of sheet (P = 0.1945) Gain of helix (P = 0.2059) Loss of methylation at R73 (P = 0.2158)	

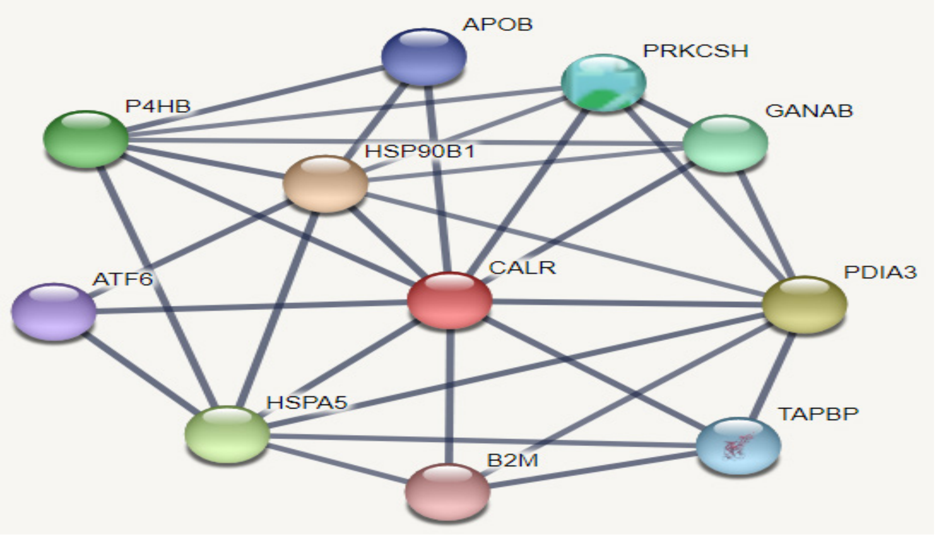

Figure 6. Functional interaction between CALR and its related genes

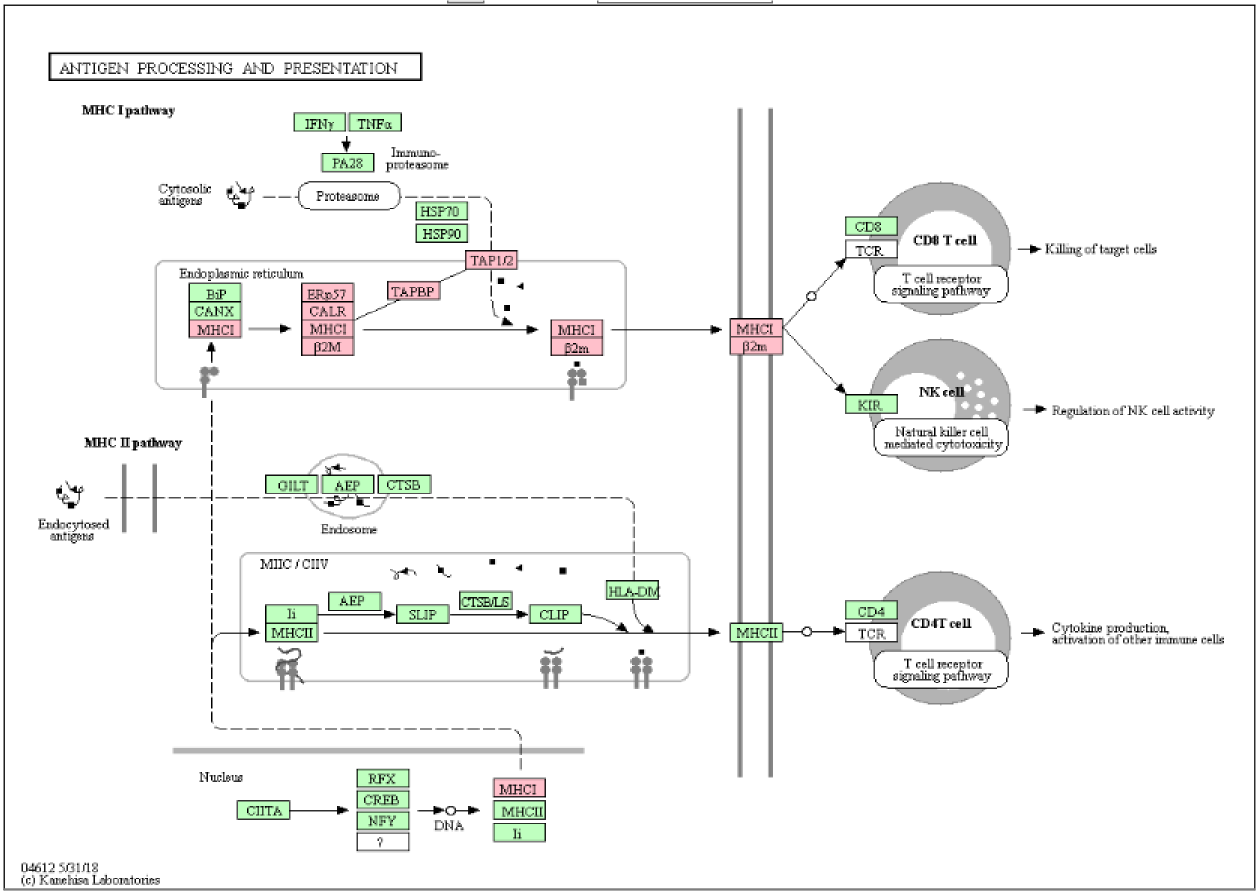

Figure 7. Antigen processing and presentation network by MHC class I (KEGG ID: hsa04612)

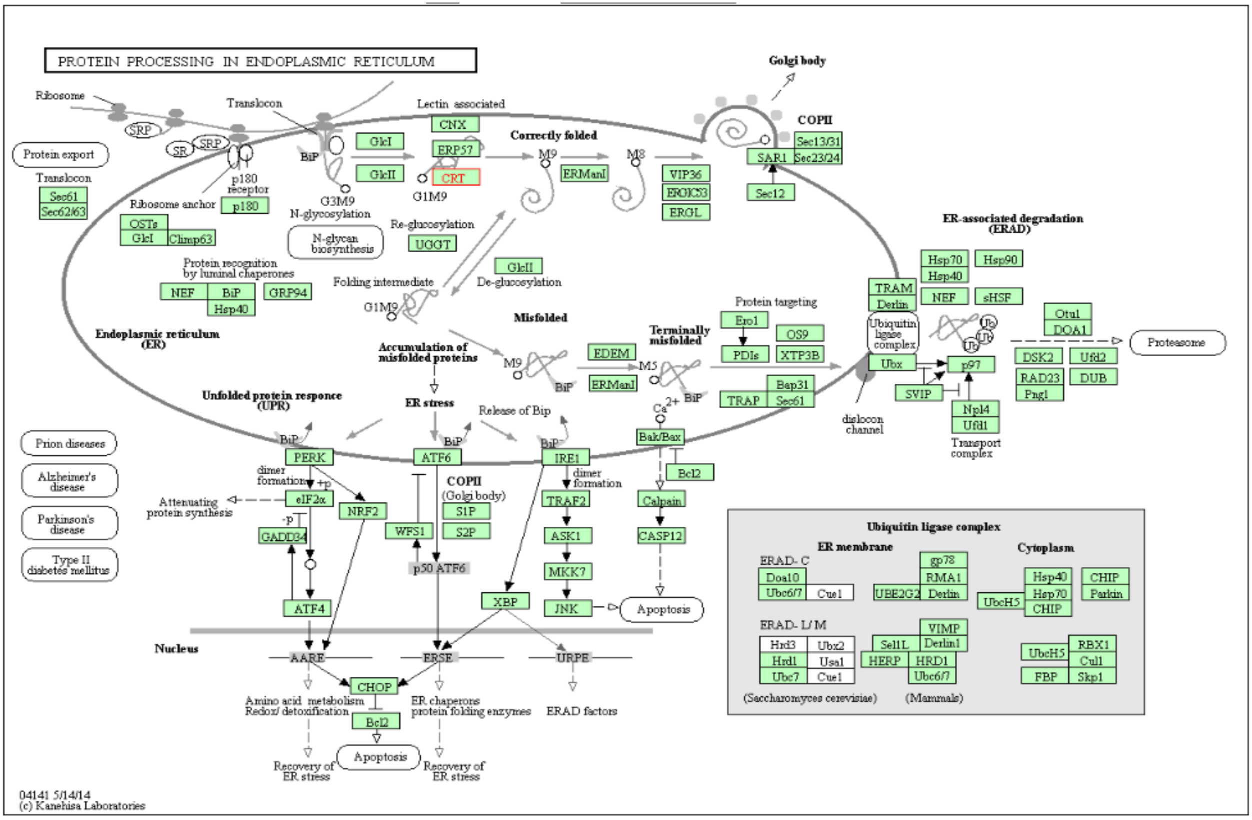

Figure 8. Protein processing in endoplasmic reticulum (KEGG ID: HSA04141)

4. Discussion

The present study undertakes a systematic approach to identify functionally important nsSNPs in human CALR gene in an attempt to understand how do these mutations affect the protein structure and function and hence promote a disease.

To sort out tolerant from intolerant nsSNPs eleven different prediction algorithm were used; SIFT, Polyphen, PROVEAN, SNAP2, CONDEL, Pmut, SNP Analyzer and PhD-SNP. By comparing the scores of all 8 methods, 4 nsSNPs with positions W261G, P216L, Y128C and R73C, W261G were found to be highly damaging nsSNPs and those were selected for further analysis by other in silico tools.

Decreased stability of protein changes was observed in all deleterious SNPs except P216L. Most of the diseases causing and cancer- related mutations are found to destabilize the corresponding protein. Decreasing in protein stability was reported in VWF A2 domain causing von Willebrand disease type 2 [39] and it was found that N214D mutation could cause a defect in tumor suppressor protein ING4 (N214D) associated with lung carcinoma [40]. Meanwhile, increased protein stability was reported to be accompanied by an elevated protein levels and protein dysregulation. Such cases have included mutations in the CLIC2 protein (H101Q) that is associated with a mental disorder [41] and in DNMT1 protein associated with breast cancer [42].

The possible molecular mechanism by which a deleterious variation associated with disease state was predicted by Mutpred. The majority of predicted effects were alteration in phosphorylation, loss of methylation, loss of catalytic residue and altered disorderness in protein structure. There is an evidence in the literature that loss of phosphorylation is to be associated with esophageal cancer [43] and loss of methylation sites are associated with hepatocellular carcinoma [44] and colorectal tumors [45].

Furthermore, mutation 3D was used to investigate the location of nsSNPs in CRT domains which resulted in variations in amino acid properties (charge, size, flexibility and hydrophobicity value). Due to these polymorphisms, new version of mutant protein might be produced which then may affect functions and structure of the original protein.

From STRING protein-protein interaction analysis, CALR was predicted with the strong interactions with HSP90B1, PDIA3, HSPA5, P4HB, GANAB, PRKCSH, TAPBP, APOB, ATF6 and B2M. The STRING interaction result was further validated by using KEGG pathways for CALR. Interestingly, same set of proteins were involved in protein processing in endoplasmic reticulum and antigen processing and presentation network by MHC class I molecules. Our literature search demonstrated that HSP90B1, HSPA5, PRKCSH, TAPBP, APOB, ATF6, B2M are involved in pathways of prostate cancer (46), Prion diseases (47), Polycystic liver disease (48), bare lymphocyte syndrome (BLS) type1 (49), familial hypercholesterolaemia (50), Achromatopsia (51) and familial hypercatabolic hypoproteinemia (52) respectively.

Therefore, any changes in the CRT protein function would have an impact on many pathways, involved in diseases.

5. Conclusions

Nowadays, in silico analysis has got major concern to screen diseases related nsSNP at molecular level. In this study a systematic and extensive in silico analysis of functional nsSNPs in the CALR gene has been performed to investigate the effect of nsSNPs on structure and function of CRT protein. Four nsSNPs with positions P216L, R73C, W261G and Y128C were predicted to be the most damaging mutations for CRT protein altering physiochemical properties of the protein; size, charge, hydrophobicity and stability leading to loss or disturbance of the protein internal and external interactions and eventually loss of the protein's function as well as disease association. Further, 16 SNPs at the 3UTR were predicted to introduce a change in the micro RNA binding site at the 3UTR which might result in deregulation of the gene function. This is computational approach, thus, population genetics based studies with clinical evidences are necessary to accompaniment the findings of this study.

ACKNOWLEDGEMENTS

The authors acknowledge Dr. Ahmed Abdelbagi Hammad for language editing and proofreading of the manuscript.

REFERENCES

[1] Michalak M, Corbett EF, Mesaeli N, Nakamura K, Opas M. Calreticulin: one protein, one gene, many functions. The Biochemical journal. 1999; 344 Pt 2:281-92.

[2] Michalak M, Robert Parker JM, Opas M. Ca^{2+} signaling and calcium binding chaperones of the endoplasmic reticulum. Cell calcium. 2002; 32(5-6): 269-78.

[3] K. Burns BD, E. A. Atkinson et al. Modulation of gene expression by calreticulin binding to the glucocorticoid receptor. Nature. 1994; 367(6462): 476–80.

[4] Lu CY WW, Lee H. Functional Role of Calreticulin in cancer biology. Biomedical Research Journal. 2015; 2015 (526524): 1-9.

[5] Varricchio L MA. Calreticulin in Myeloproliferative neoplasms: The other side of the Alice Mirro. European Medical Journal Haematology. 2014; 1:114-22.

[6] Machado-Neto JA dMCP, Traina F. CALR (calreticulin). Atlas Genet Cytogenet Oncol Haematol in press. 2016.

[7] Chachoua I, Pecquet C, El-Khoury M, Nivarthi H, Albu RI, Marty C, et al. Thrombopoietin receptor activation by myeloproliferative neoplasm associated calreticulin mutants. Blood. 2016; 127(10): 1325-35.

[8] Araki M, Yang Y, Masubuchi N, Hironaka Y, Takei H, Morishita S, et al. Activation of the thrombopoietin receptor by mutant calreticulin in CALR-mutant myeloproliferative neoplasms. Blood. 2016; 127(10): 1307-16.

[9] Ghiran I, Klickstein LB, Nicholson-Weller A. Calreticulin is at the surface of circulating neutrophils and uses CD59 as an adaptor molecule. J Biol Chem. 2003; 278(23): 21024-31.

[10] Qui Y MM. Transcription Control of the Calreticulin gene in health and disease. The International Journal of Biochemistry and Cell Biology. 2009; 41 (3): 531-8.

[11] M. Villagomez ES, A. Podcheko, T. Feng, S. Papp, and M. Opas. Calreticulin and focal-contact-dependent adhesion. Biochemistry and Cell Biology. 2009; 87(4): 545–56.

[12] H. Totary-Jain TN-M, Y. Riahi, N. Kaiser, J. Eckel, and S. Sasson. Calreticulin destabilizes glucose transporter-1 mRNA in vascular endothelial and smooth muscle cells under high-glucose conditions. Circulation Research. 2005; 97(10): 1001–8.

[13] W.-F. Chiang T-ZH, T.-C. Hour et al. Calreticulin, an endoplasmic reticulum-resident protein, is highly expressed and essential for cell proliferation and migration in oral squamous cell carcinoma. Oral Oncology. 2013; 9(6): 534–41.

[14] K. Chahed MK, L. Ehret-Sabatier et al. Expression of fibrinogen E-fragment and fibrin E-fragment is inhibited in the human infiltrating ductal carcinoma of the breast: the two-dimensional electrophoresis and MALDI-TOF-mass spectrometry analyses. International Journal of Oncology. 2005; 27(5): 1425–31.

[15] Milone MR, Pucci B, Colangelo T, Lombardi R, Iannelli F, Colantuoni V, et al. Proteomic characterization of peroxisome proliferator-activated receptor-γ (PPARγ) overexpressing or silenced colorectal cancer cells unveils a novel protein network associated with an aggressive phenotype. Molecular Oncology. 2016; 10(8): 1344-62.

[16] M. Alur MMN, S. E. Eggener et al. Suppressive roles of calreticulin in prostate cancer growth and metastasis. The American Journal of Pathology. 2009; 175(2): 882–890.

[17] K. Hellman AAA, K. Schedvins, W. Steinberg, A.-C. Hellström, and G. Auer. Protein expression patterns in primary carcinoma of the vagina. British Journal of Cancer. 2004; 91 (2): 319–326.

[18] O. Vaksman BD, C. Tropé, and R. Reich. Calreticulin expression is reduced in high-grade ovarian serous carcinoma effusions compared with primary tumors and solid metastases. Human Pathology. 2013; 44 (12): 2677–2683.

[19] J. Nangalia CEM, E. J. Baxter et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. The New England Journal of Medicine. 2013; 369(25): 2391–2405.

[20] T. Klampfl HG, A. S. Harutyunyan et al. Somatic mutations of calreticulin in myeloproliferative neoplasms. The New England Journal of Medicine. 2013; 369(25): 2379–2390.

[21] Lavi N. Calreticulin Mutations in Myeloproliferative Neoplasms. Rambam Maimonides Medical Journal. 2014; 5(4): e0035.

[22] Luo W YZ. Calreticulin (CALR) mutation in myeloproliferative neoplasms (MPNs). Stem Cell Investigation 201. 2015; 2: 16.

[23] Helbling D, Mueller BU, Timchenko NA, Schardt J, Eyer M, Betts DR, et al. CBFB-SMMHC is correlated with increased calreticulin expression and suppresses the granulocytic differentiation factor CEBPA in AML with inv(16). Blood. 2005; 106(4): 1369-75.

[24] Hou HA, Kuo YY, Chou WC, Chen PH, Tien HF. Calreticulin mutation was rarely detected in patients with myelodysplastic syndrome. Leukemia. 2014; 28: 1555.

[25] Heuser M, Panagiota V, Koenecke C, Fehse B, Alchalby H, Badbaran A, et al. Low frequency of calreticulin mutations in MDS patients. Leukemia. 2014; 28: 1933.

[26] Eder-Azanza L, Navarro D, Aranaz P, Novo FJ, Cross NC, Vizmanos JL. Bioinformatic analyses of CALR mutations in myeloproliferative neoplasms support a role in signaling. Leukemia. 2014; 28(10): 2106-9.

[27] P.C. Ng SH. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006;7 61-80.

[28] I.A. Adzhubei SS, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7: 248-9.

[29] Hoi Y and Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics. 2015; 31(16): 2745-2747.

[30] Hecht L, Wass J, Kelly L, Clevenger-Firley E, Dunn C. SNAP-Ed Steps to Health Inspires Third Graders to Eat Smart and Move More. Journal of Nutrition Education and Behavior. 2013; 45(6): 800-2.

[31] L. Bao MZ, Y. CuiNsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005; 33: W480-W4802.

[32] Orozco CF-CJLGLZIPXdlCM. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005; 21(14): 3176–8.

[33] Capriotti E CRaCR. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006; 22: 2729-34.

[34] E. Capriotti PF, R. CasadioI-Mutant 2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res: 2005; 33: W306-W310.

[35] 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 bioinformatics. 2010; 11: 548.

[36] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. 2009; 25(21): 2744-50.

[37] Mering V. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res 2005; 33(Suppl. 1): D433-D7.

[38] Kanehisa, Minoru and Susumu Goto. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research. 1999; 27 (1): 29-34.

[39] Xu AJ ST. Mechanisms by which von Willebrand disease mutations destabilize the A2 domain. J Biol Chem 2013; 288: 6317–24.

[40] Moreno A, Palacios A, Orgaz JL, Jimenez B, Blanco FJ, Palmero I. Functional impact of cancer-associated mutations in the tumor suppressor protein ING4. Carcinogenesis. 2010; 31(11): 1932-8.

[41] Witham S, Takano K, Schwartz C, Alexov E. A Missense Mutation in CLIC2 Associated with Intellectual Disability is Predicted by In Silico Modeling to Affect Protein Stability and Dynamics. Proteins. 2011; 79(8): 2444-54.

[42] Agoston AT, Argani P, Yegnasubramanian S, De Marzo AM, Ansari-Lari MA, Hicks JL, et al. Increased Protein Stability Causes DNA Methyltransferase 1 Dysregulation in Breast Cancer. Journal of Biological Chemistry. 2005; 280(18): 18302-10.

[43] Al-Kharashi LA, Al-Mohanna FH, Tulbah A, Aboussekhra A. The DNA methyl-transferase protein DNMT1 enhances tumor-promoting properties of breast stromal fibroblasts. Oncotarget. 2018; 9(2): 2329-43.

[44] Anwar SL, Krech T, Hasemeier B, Schipper E, Schweitzer N, Vogel A, et al. Loss of DNA methylation at imprinted loci is a frequent event in hepatocellular carcinoma and identifies patients with shortened survival. Clinical Epigenetics. 2015; 7(1): 110.

[45] Ashktorab H, Daremipouran M, Goel A, Varma S, Leavitt R, Sun X, et al. DNA methylome profiling identifies novel methylated genes in African American patients with colorectal neoplasia. Epigenetics. 2014; 9(4): 503-12.

[46] Cornfold PA, Dodson AR, Parsons KF, Desmond AD, Woolfenden A, Fordham M, et al. Heat-shock protein expression independently predicts clinical outcome in prostate cancer. Cancer Res. 2000; 60: 7099-7105.

[47] Park KW, Eun Kim G, Morales R, Moda F, Moreno Gonzalez I, Concha-Marambio L, et al. The Endoplasmic Reticulum Chaperone GRP78/BiP Modulates Prion Propagation in vitro and in vivo. Scientific Reports. 2017; 7.

[48] Gao H, Wang Y, Wegierski T, Skouloudaki K, Putz M, Fu X, et al. PRKCSH/80K-H, the protein mutated in polycystic liver disease, protects polycystin-2/TRPP2 against HERP-mediated degradation. Hum Mol Genet. 2010; 19(1): 16-24.

[49] Yabe T, Kawamura S, Sato M, Kashiwase K, Tanaka H, Ishikawa Y, et al. A subject with a novel type I bare lymphocyte syndrome has tapasin deficiency due to deletion of 4 exons by Alu-mediated recombination. Blood. 2002; 100(4): 1496-8.

[50] Bassani Borges J, Medeiros Bastos G, Almendros Afonso TK, da Silva Rodrigues E, Strelow Thurow H, Dominguez Crespo Hirata T, et al. Study of APOB mutations and familial hypercholesterolemia phenotypes. Atherosclerosis. 2016; 252: e43.

[51] Ansar M1 S-CR, Saqib MA, Zulfiqar F, Lee K, Ashraf NM, et al. Mutation of ATF6 causes autosomal recessive achromatopsia. Hum Genet. 2015; 34(9): 941-50.

[52] Wani MA, Haynes LD, Kim J, Bronson CL, Chaudhury C, Mohanty S, Waldmann TA, Robinson JM, Anderson CL. Familial hypercatabolic hypoproteinemia caused by deficiency of the neonatal Fc receptor, FcRn, due to a mutant beta2-microglobulin gene. Proc Natl Acad Sci USA. 2006; 103(13): 5084-5089.