Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI - 10.26479/2017.0205.17**

IN SILICO ANALYSIS OF SINGLE NUCLEOTIDE POLYMORPHISMS IN HUMAN PROLIFERATING CELL NUCLEAR ANTIGEN GENE

V. Chernapalli *[1], M. Komanduri [2], S. Sabbella[2]

1.Department of Bio-Chemistry, Government City College, Hyderabad, Telangana, India.

2.Synteny Life sciences Pvt. Ltd, Hyderabad, Telangana, India.

ABSTRACT: Aim: Proliferating cell nuclear antigen (PCNA) is an ancillary protein that assists in DNA replication and DNA repair process. Detection of mutations in PCNA gene or protein could be helpful to analyze many disorders relating to DNA replication and repair. In the present study, non-synonymous single nucleotide polymorphisms (nsSNPs) generated by missense mutation were studied using different computational methods to evaluate the nsSNPs that may be deleterious to PCNA function.**Method:** The missense nsSNPs were retrieved from the database of NCBI and subjected to functional prediction by the computational algorithms. The evolutionary conservation data of the PCNA protein was obtained from the ConSurf web server. The protein structural analysis for the variants was performed using I-Mutant, SPDB viewer and YASARA to check their structural variations and energy minimizations. **Results:** Out of the 42 nsSNPs, 5 nsSNPs namely rs780735449 (Q38R), rs1050525 (S39R), rs781573975 (E104G), rs772308650 (L182W) and rs753494859 (K248N) were identified as deleterious by using different computational algorithms. The evolutionary conservation data revealed that all the high risk nsSNPs positions were highly conserved and were either functional or structural residues in the protein. The I- Mutant tool had showed a decrease in the protein stability for the five high risk nsSNPs. A deviance in the energy minimization was observed for the variants with respect to the native protein. The RMSD (root mean square deviation) and TM (template modeling) values predicted that the mutants were structurally similar to the wild type protein. The PTM (post translational modifications) analysis using various in silico tools showed that S39R and K248N were putative PTM sites and through FTSite it was observed that these two variants were also involved in the ligand binding sites.**Conclusion:** Through the robust use of various in silico tools, five nsSNPs of PCNA protein were found to deleterious. Out of them two mutations at 39th and 248th positions (S39R & K248N) are to be further validated for their effect on the PCNA protein function.

Peer review under responsibility of Life Science Informatics Publications

KEYWORDS: PCNA, nsSNPs, in silico, protein stability, energy minimization

***Corresponding Author: Dr. V. Chernapalli** Ph.D.
Department of Bio-Chemistry, Government City College, Hyderabad, Telangana, India
* Email Address: sirisha.aroma@gmail.com

1.INTRODUCTION

In silico analysis of single nucleotide polymorphism (SNPs) are widely used in present years due to wide availability of computational tools. The analysis of SNPs by conventional methods is usually robust and time consuming. However, with the recent advances in computational tools and their ease of availability, the potential problems of conventional methods can be addressed. The application of computational tools in analysis not only makes the process easy but also reduces the cost of conventional methods by great fold as a number of tools are available online at a free of cost. The present study, the analysis of PCNA SNPs using computational tools is an attempt to understand the nature of different SNPs of PCNA using different algorithms which are available online. The study of single nucleotide polymorphism helps in understanding the nature of different diseases. About 90% of human polymorphisms are formed by single nucleotide variation. However, the nsSNPs with missense mutation usually lead to loss of function while in a few instances can cause a gain in protein function [1]. There are several instances where SNPs are proved to be involved in disorders like sickle cell anaemia, β thalassemia [2,3], rheumatoid arthritis [4] and even in cancer [5,6].PCNA is a cyclin protein which functions as a key factor in eukaryotic DNA polymerase [7]. PCNA is a ring-shaped protein complex that surrounds the DNA and increases the processivity of DNA replicases δ subunit and coordinates the various pathways in DNA replication [8, 9] by encircling and freely sliding along the DNA helix by forming a ring of homo trimer. DNA repair and replication is the key process in maintaining the integrity of a cell. Apart from DNA repair and replication, PCNA is also associated with remodeling of chromatin and in the process, interacts with a large number of accessory proteins and thus acts as a protein recruiting platform [10, 11]. PCNA is reported to be overexpressed in cancer cells and it is also very often used as a marker of proliferation [12]. In the base excision repair pathway, PCNA (A1876G) polymorphisms were found to be associated with increased risk of non-small cell lung cancer [13]. It is reported that two mutant forms of PCNA formed due to amino acid substitutions cause defects in mismatch repair system. The C22Y mutant PCNA protein was found to block MutSα-dependent MMR (mismatch repair) and C81R mutant

PCNA protein was found to partially block both MutSα-dependent and MutSβ-dependent MMR [14]. Though PCNA is one of the important cyclins in eukaryotic DNA polymerase complex, the SNP studies are only limited to Ser228Ile. It was reported that the Ser228Ile mutation caused a large conformational change in the protein thereby altering the binding site for PCNA interacting proteins [15, 16]. The missense mutation Ser228Ile in PCNA was found to be associated with a neurodegenerative phenotype, exhibiting manifestations common to other DNA repair disorders. Hence, the present study is an attempt to identify other PCNA SNPs which could be possibly detrimental to the functioning of the PCNA. This is done by using different computational algorithms. The resulting high risk nsSNPs can be further validated through wet lab experiments.

2. MATERIALS AND METHODS

Retrieval of SNPS for PCNA, human

The protein sequence data on PCNA was collected from National Center for Biological Information (NCBI) web site. The SNP data was retrieved from different web based data sources such as the NCBI dbSNP database (https://www.ncbi.nlm.nih.gov/SNP/) and the Ensembl genome browser (http://www.ensembl.org/index.html) [17, 18].

Evaluation of functional consequences of non-synonymous SNPs

The nsSNPs retrieved were further subjected to functional prediction using diffferent*in silico* algorithms namely SIFT, Polymorphism Phenotyping 2 (PolyPhen), nsSNPs analyzer, Predictor of Human Deleterious Single Nucleotide Polymorphisms (PhD-SNP), SNPs & GO and PMut. SIFT server was used to predict the deleterious effects of nsSNPs (http://sift.jcvi. org/). SIFT predicts tolerated and deleterious substitutions for every position of the query sequence by using multiple alignment information. The SIFT score ≤ 0.05 indicates the damaging effect of nsSNPs on protein function [19]. PolyPhen (http://genetics.bwh.harvard.edu/pph2) software utilizes the protein sequence as well as the amino acid modified position in protein sequence to predict the effect of nsSNP on protein structure and function. The tool assesses the position-specific independent count (PSIC) score for every variant and the score difference directly implies the functional consequences of nsSNPs on protein function. The mutation is evaluated as “possibly damaging” (probabilistic score > 0.15), “probably damaging” (probabilistic score > 0.85) and “benign” (remaining mutations) [20]. The nsSNPs analyzer (http://snpanalyzer.uthsc.edu/) uses information contained in the multiple sequence alignment and information embodied in the 3D structure to make predictions [21]. PhD-SNP (http://snps.biofold.org/phd-snp/phd-snp.html) uses support vector machine (SVM) based analyzing method and makes sequence and profile based prediction [22]. SNPs & GO is also a SVM-based nsSNPs classifier consisting of a single SVM that draws in input protein sequence,

profile and functional information [23]. And lastly PMut (http://mmb.pcb.ub.es/PMut/) is based on the use of different kinds of sequence information to label mutations, and neural networks to process this data [24]. FASTA sequence of PCNA was given as input and result was based on the differences among disease related and neutral variations of protein sequence. Probability score greater than 0.5 discloses the disease related effect of mutation on the protein function.On the whole six different SNP prediction programs were used. nsSNPs predicted to be deleterious by at least 4 different *in silico* algorithms were classified as high risk nsSNPs. Because each algorithm utilizes different parameters to assess the nsSNPs, hence,nsSNPs with more positive results in SNP algorithms are more likely to be deleterious.

Conservation Profile of PCNA gene

To define the conservation pattern, PCNA gene from *Homo sapiens* (Accession number-CAG46598) along with PCNA from two species belonging to Hominidae family such as *Pan troglodytes* (Accession number- XP_001165515), *Gorilla gorillagorilla* (XP_004061824), *Mesocricetusauratus* belonging to Cricetidae family (XP_012967724), *Caviaporcellus* from Caviidae family (XP_003476647), *Musmusculus* (NP_035175) from Muridae family, *Bostaurus* (NP_001029666) and *Bison bisonbison* (XP_010834739) belonging to Bovidae family were retrieved from NCBI sequence database. The protein sequences thus retrieved from NCBI were aligned using Clustal X software version 2.0 [25] and multiple sequence comparison by log-expectation (MUSCLE). MUSCLE analysis was performed by the online version (http://www.ebi.ac.uk/Tools/msa/muscle/) [26]. Evolutionary conservation of amino acids in the PCNA protein was determined by ConSurf server (http://consurf.tau.ac.il/2016/) [27]. The evolutionary conservation was performed by maximum likelihood in ConSurf server.

Prediction of Post - translational modification sites

Using the UbPred(www.ubpred.org) and BDM-PUB (bdmpub.biocuckoo.org) programs the putative ubiquitylation sites were predicted. The lysine residues with a score greater than 0.62 were considered ubiquitinated in UbPred. The balanced cut off option was selected for BDM-PUB [28]. Using the SUMO plot (http://www.abgent.com/sumoplot) and SUMOsp 2.0 (http://sumosp.biocuckoo.org/) programs the putative sumoylation sites were predicted. In SUMO plot, only high probability motifs with a score of 0.5 were considered sumoylated. For SUMOsp 2.0 Medium level threshold was selected [29, 30]. Putative phosphorylation sites were predicted using GPS 3.0 (http://gps.biocuckoo.org/) and NetPhos 3.1 (http://www.cbs.dtu.dk/services/NetPhos/). A high-level threshold was selected for GPS 3.0. For NetPhos 3.1, serine, threonine and tyrosine residues with a score of 0.5 were considered as phosphorylated [31, 32]. The web-server iPTM-mLys,

(http://www.jci-bioinfo.cn/iPTM-mLys) the first multi label PTM predictor was used to predict the identifying lysine PTM sites [33, 34, 35].

Protein Stability and Structural Analysis

I-Mutant 2.0 (http://folding.biofold.org/imutant/i-mutant2.0.html) is a support vector machine (SVM) based tool which predicts the protein stability changes upon nsSNPs. I-Mutant version2 predicts the Gibbs free energy change (DDG) by subtracting the unfold Gibbs free energy of mutated protein from the unfold Gibbs free energy of the wild type protein (DDG or $\Delta\Delta G = \Delta G$ mutant – ΔG wild type). Prediction of protein stability changes can be performed by use of either protein sequence or structure. I-Mutant version 2 also predicts the sign of decrease or increase in Gibbs free energy with Reliability Index (RI: 0–10, where 0 is the lowest reliability and 10 is the highest reliability) for amino acid change. The value of DDG <0 indicates decrease in protein stability and DDG>0 indicates increase in protein stability [36]. For all the nsSNPs submissions the pH and temperature were set as 7 and 25^0C respectively. To explore the structural deviances and stability variances between native and mutant forms of PCNA proteins the structural analysis was done. The Protein Data Bank (PDB) ID 3VKX corresponds to the crystal structure of human PCNA protein. In order to make the mutated models of the PCNA for corresponding amino acid substitutions, Swiss-PDB viewer was utilized [37]. Swiss-PDB "mutation tool" was utilized to substitute the wild type amino acid with a new amino acid. This tool enables the replacement of the native amino acid by the top rotamer of the new amino acid. The PDB files generated for all the models were saved. The TM scores and the root mean square deviations of the mutant models with respect to the native PCNA were calculated using TM-Align [38]. Further to improve the quality of the predicted models energy minimization studies were carried out using YASARA force field minimization tool [39]. To validate the structures produced, RAMPAGE server (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) was used which produces the Phi/Psi Ramachandran plot for the proteins [40].

Identification of Ligand Binding Sites by FTSite

Detection of binding site is important for elucidating the structure-function relationships among proteins, protein engineering and drug design. In the present study for the identification of binding site FTSite(http://ftsite.bu.edu/) was used. Based on the experimental evidences FTSite predicts the ligand binding sites of proteins with 94% accuracy [41].

Analysis of protein-protein interactions

To study the functions of the interactions of proteins at the cellular level, protein-protein interaction networks are vital. In the present study, online database resource Search Tool for the Retrieval of Interacting Genes (STRING) was used to identify the interactions of PCNA protein with other

corresponding proteins [42]. This tool offers exclusive coverage and ease of access to both experimental and theoretical interaction information of PCNA. To infer the most suitable interactions among the nodes in PCNA protein interaction network, various modes were used such as confidence view, evidence view, interactive mode and action view.

3. RESULTS AND DISCUSSION

SNP datasets

The polymorphism data for the PCNA, human gene retrieved from the NCBI dbSNP database and the Ensembl genome browser shows that the PCNA gene contains 48 SNPs in 5' UTR region, 27 SNPs in 3'UTR region, 745 SNPs in intronic region and 42 missense variants. The results of the function prediction of the 42 missense variants subjected to various *in silico* algorithms are summarized in Table 1.

Table 1. Predicted results for the nssnps in the pcna gene using different *in silico* algorithms

Prediction	Number of nsSNPs (%)					
	SIFT	PP-2	nsSNP AZ	PhD-SNP	SNP & GO	P Mut
Deleterious	20 (49)	08 (20)	-	-	-	-
PD	-	07(17)	-	-	-	-
Benign	21 (51)	26(63)	-	-	-	-
Disease	-	-	10 (24)	10 (24)	02 (5)	32 (78)
Neutral	-	-	31 (76)	31 (76)	39 (95)	09 (22)

Percentage of total nsSNPs shown in parentheses for each category. PD: probably deleterious; the in silico algorithms used are SIFT, PP-2: Polymorphism Phenotyping 2; nsSNP AZ: nsSNP Analyzer; PhD-SNP: Predictor of Human Deleterious Single Nucleotide Polymorphisms; SNP & GO; P Mut.

Non-synonymous SNP analysis

Functional consequences of the nsSNPs was analyzed using the following *in silico* algorithms: SIFT, PolyPhen2, nsSNP analyzer, PhD-SNP, SNP & GO and PMUT (see supplementary Table1). SIFT analysis predicted that 20 nsSNPs (49%) are deleterious and 21nsSNPs (51%) are tolerated. According to PolyPhen 2 results, 8 nsSNPs (20%) are deleterious, 26 nsSNPs (17%) are benign whereas the remaining 7 nsSNPs (63%) are possibly damaging. Both nsSNP analyzer and PhD-SNP predicted that 10 nsSNPs (24%) are disease causing and 31 nsSNPs (76%) are neutral. PMUT analysis predicted that 32 nsSNPs (78%) are pathological and 9 nsSNPs (22%) are neutral. On the

contrary SNP & GO predicted that 2 nsSNPs (5%) are deleterious and 39 nsSNPs (95%) are neutral. Here we classified the nsNSPs as high risk oriented if they were predicted to be deleterious by four or more *in silico* SNP prediction algorithms. Out of 41nsSNPs, 5nsSNPs met the mentioned criteria and were selected for further analysis. The deleterious predictions for the five nsSNPs are shown in Table 2. The nsSNPs of PCNA predicted to be deleterious byatleast four*in silico* algorithms were categorized as high risk nsSNPs. The decimal 0.5 indicates that the nsSNP is possibly deleterious rather than probably deleterious which is considered as 1.

Table 2. Prediction of deleterious nssnps in the pcna gene.

nsSNPs ID	Mutation	Number of deleterious predictions
rs780735449	Q38R	5
rs1050525	S39R	5.5
rs781573975	E104G	5
rs772308650	L182W	5
rs753494859	K248N	4

Conservation profiling

Conservation of amino acids in a protein play an important role in protein structure and function thereby in all the cellular metabolisms. Most of the conserved amino acids are buried in the protein structure while the most of the non-conserved amino acids are exposed [43]. Mutagenesis of conserved amino acids leads to lethal effect [44]. Due to this reason, conservation pattern of amino acids in the PCNA protein was studied using ConSurf web browser. Highly conserved residues are predicted to be either structural or functional based on their position relative to the protein surface [45]. We focused on amino acid sites that coincide in location with the 5 high-risk nsSNPs and they were found to be highly conserved (Table 3; Supplementary figure 1). To further evaluate the

sequence conservation of the protein, the multiple sequence alignment data was generated from Clustal X and MUSCLE. Both the servers generated same sequence alignment (Figure 1) The Percentage Identity Matrix for the 8 species was generated using MUSCLE (Table 4).

Table 3. Conservation analysis with consurf:

Amino acid position	Conservation Score	Conservation pattern	MSA result
Q38	9	Highly Conserved and Exposed (f)	Conserved across given species
S39	9	Highly Conserved and Buried (s)	Conserved across given species
E104	9	Highly conserved and exposed (f)	Conserved across given species
L182	9	Highly Conserved and Buried (s)	Conserved across given species
K248	9	Highly Conserved and Exposed (f)	Conserved across given species

It represents the conservation score of different amino acid residues calculated by ConSurf. Conservation score (1–4 = variable, 5 = average, 6–9 = conserved); (f): predicted functional site, (s): predicted structural site.MSA - Multiple Sequence Alignment

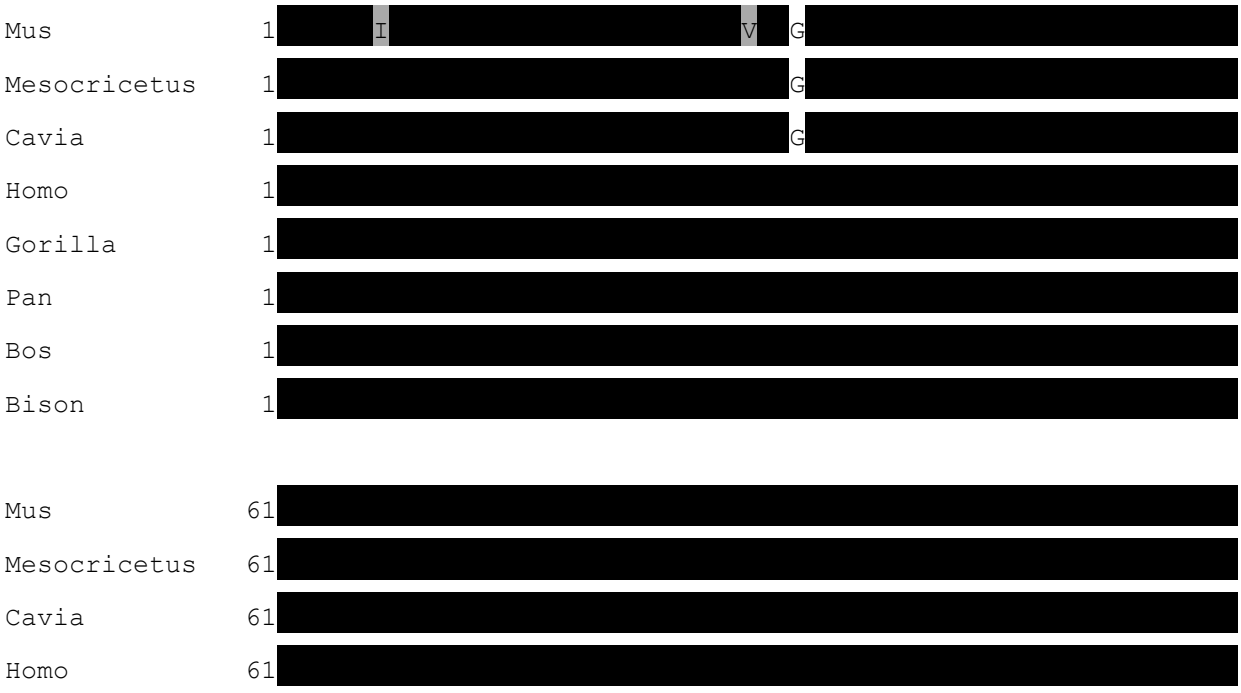


Figure1. **Multiple Sequence Alignment of PCNA generated by Clustal X and MUSCLE**

It represents the Multiple Sequence Alignment of PCNA protein of 8 different species. This was carried out by using multiple sequence comparison by log-expectation (MUSCLE) and Clustal X. Human PCNA sequence was take as reference. The regions highlighted in black represent the conserved regions.

Table 4. Percent identity matrix (pim) generated by muscle

	Mus	Mesocricetus	Cavia	Homo	Gorilla	Pan	Bos	Bison
Mus	100.00	96.93	97.32	96.93	96.93	96.93	96.93	96.93
Mesocricetus	96.93	100.00	99.62	98.47	98.47	98.47	99.23	99.23
Cavia	97.32	99.62	100.00	98.85	98.85	98.85	98.85	98.85
Homo	96.93	98.47	98.85	100.00	100.00	100.00	99.23	99.23
Gorilla	96.93	98.47	98.85	100.00	100.00	100.00	99.23	99.23
Pan	96.93	98.47	98.85	100.00	100.00	100.00	99.23	99.23
Bos	96.93	99.23	98.85	99.23	99.23	99.23	100.00	100.00
Bison	96.93	99.23	98.85	99.23	99.23	99.23	100.00	100.00

Prediction of putative post translational modification sites in PCNA

The Post Translational Modifications (PTMs) play a key role in regulating activity, localization, stability of proteins and also interaction of proteins with other cellular molecules [46]. The deregulation of PTMs also has a role in tumorigenesis [47]. Hence we investigated the PTMs in PCNA protein and for this purpose we used a variety of *in silico* tools to predict the putative PTM sites in the PCNA protein. To identify amino acid that might undergo phosphorylation we used GPS 3.0 and NetPhos 3.1 servers. The GPS 3.0 predicted that there were 7 serine specific phosphorylation sites and 1 threonine specific and no tyrosine specific sites in the PCNA protein. Conversely NetPhos 3.1 predicted that there were 13 serine specific phosphorylation sites and 9 threonine specific and 3 tyrosine specific sites in the PCNA protein. Among the 13 serine specific phosphorylation sites predicted by NetPhos 3.1 server Serine at position 39 occupies a place indicating that the mutant S39R will be devoid of this putative phosphorylation site (Table 5).In addition to phosphorylation we also screened the PCNA protein for the putative sumoylation, ubiquitylation and acetylation sites. To analyze residues in PCNA that may undergo sumoylation or ubiquitylation we used the SUMO-plot, SUMOsp 2.0, BDM-PUB and UbPred programs. SUMO-plot predicted that 4 lysine residues undergo sumoylation whereas SUMO sp 2.0 predicted that only 3 residues undergo sumoylation. BDM-PUB predicted that 6 lysine residues undergo ubiquitylation whereas UbPred predicted that 4 residues undergo ubiquitylation. The amino acid at position 164 was predicted to be a putative ubiquitylation site by both the servers. Similarly

iPTM-mLys server was used to predict lysine PTM sites and the analysis revealed that there were 15 acetylation sites. Among them lysine at position 248 occupies a place and the mutant K248N might be at loss of this PTM site (Table 6).

Table 5. Putative phosphorylation sites in pcna protein.

GPS 3.0			Net Phos 3.1		
Serine	Threonine	Tyrosine	Serine	Threonine	Tyrosine
134	196	-	10	51	60
170			32	59	114
183			39*	73	133
261			46	89	
112			112	98	
223			141	185	
261			152	196	
			183	219	
			186	224	
			222		
			228		
			230		
			261		

The * indicates the position of high risk nsSNPs.

Table 6. Putative sumoylation, ubiquitylation and acetylation sites in pcna protein.

Sumoylation		Ubiquitylation		Acetylation
SUMO plot	SUMOsp 2.0	BDM-PUB	UbPred	iPTM-mLys
139	13	14	164	13
166	190	20	181	14
257	254	80	190	20
192		138	254	77
		164		80
		168		110
				117
				164
				181
				168
				190
				217
				240
				248*
				254

The * indicates the position of high risk nsSNPs.

Protein stability and Structural analysis of high-risk non-synonymous SNPs

Protein Stability: I-mutant is a SVM based tool that predicts the protein stability changes upon single point mutations taking protein sequence or structure (PDB format) as an input. The prediction of stability changes of selected 5 nsSNPs by IMutant2.0 is given in Table 7. The results are predicted to either increase or decrease the free energy change upon amino acid substitutions. All the selected nsSNPs showed DDG value less than 0 suggesting decreased protein stability.

Table 7. Protein structural stability based on standard free energy change in pcna nssnps.

Mutation	pH	Temperature	Stability	DDG(kcal/mol)	RI	RSA
Q38R	7	25 °C	Decrease	-1.44	7	13.4
S39R	7	25 °C	Decrease	-1.22	5	18.8
E104G	7	25 °C	Decrease	-2.17	9	14.9
L182W	7	25 °C	Decrease	-0.26	8	17.5
K248N	7	25 °C	Decrease	-0.86	2	22.4

DDG: free energy change value between wild and mutant, sign of DDG indicates the direction of the change (increase or decrease); RI is the reliability where 0 is the lowest RI and 9 is the highest; RSA: relative surface accessibility. The five predicted deleterious and disease causing mutants were mapped to the PDB ID 3VKX native structure. Further, to generate mutated model structures for the five variants, substitution of corresponding amino acid residues was carried out using Swiss-PDB Viewer individually. We further calculated the RMSD and Tm score for the Q38R, S39R, E104G, L182W and K248N variants. RMSD measures the average distance between the alpha carbon backbones of the wild type and mutant proteins whereas TM score indicates the topological similarity between the two protein structures [38]. The greater is the RMSD value, the more is the discrepancy between the two structures. It can be seen from Table 8 that the RMSD values between the native structure and the mutant modeled structures are all similar suggesting that these mutations do not bring a significant variation in the mutant structures with regard to the native protein structure.

Table 8. RMSD Values And TM Score Of Mutant Modeled Structures Of PCNA Protein.

Variant	RMSD	TM score
Q38R	0.00	1.00
S39R	0.00	1.00
E204G	0.00	1.00
L182W	0.00	1.00
K248N	0.00	1.00

Tm-score value scales the structural similarity.TM-score value 1 indicates a perfect match between two structures. RMSD is the root mean square deviation. A higher RMSD value indicates greater deviation between wild type and mutant structure. Moreover, to this we have subjected the Mutated model structure of high-risk nsSNPs to RAMPAGE for model validation. The RAMPAGE analysis for the wild type PCNA protein showed that 243 (99.6%) of amino acid residues were found in the favored region, 0 (0%) were in the allowed region and 1 (0.4%) residue in the outlier region. The results of the Ramachandran plot analysis for each of the high risk nsSNPs model structures is given in Table 9. The energy minimizations were achieved by YASARA server for the native protein and the five mutant proteins. The total energy for the native type protein structure following energy minimization was -148368.7 kJ/mol (score: -0.12) whereas prior to energy minimization it was. 219601.1kJ/mol (score: -1.32). The total energy before and after energy minimization is given in

Table 10.

Table 9. Rampage analysis

Model	Amino acid residues in favored regions		Amino acid residues in allowed regions		Amino acid residues in outlier regions	
	No of residues	% of residues	No of residues	% of residues	No of residues	% of residues
Wild type	243	99.6	0	0	1	0.4
Q38R	240	92.7	16	6.2	3	1.2
S39R	236	93.3	13	5.1	4	1.6
E204G	234	90.3	20	7.7	5	1.9
L182W	241	93.1	18	6.9	0	0
K248N	240	92.7	16	6.2	3	1.2

Table 10. Total energy of native and mutant pcna structures before and after energy minimization.

Amino acid variants	Total energy before energy minimization (kj/mol)	Total energy after energy minimization (kj/mol)
Native	219601.1	-148368.7
Q38R	219678.8	-147934.5
S39R	220340.7	-148904.2
E104G	220275.7	-147361.4
L182W	238179.5	-146881.8
K248N	220499.3	-150222.4

Analysis of Ligand Binding Sites and Protein-Protein Interactions

FT Site recognizes 3 ligand binding sites on PCNA protein. The amino acids found in these 3 sites of PCNA protein are given in Table 11. By the results of FT Site, it is observed that out of 5 selected variants Q38 & S39 are involved in the ligand binding site 2 whereas K248 is found to be involved in ligand binding at site 3.

Table 11. Residues at ligand binding sites of PCNA protein.

Site 1	Site 2	Site 3
MET A 40	LEU A 22	GLN A 49
VAL A 45	ILE A 23	ILE A 128
SER A 46	ASN A 24	HIS A 246
LEU A 47	GLU A 25	LYS A 248*
LEU A 126	ALA A 26	TYR A 250
GLY A 127	CYS A 27	
ILE A 128	GLN A 38*	
PRO A 129	SER A 39*	
GLU A 130	MET A 40	
GLN A 131	ASP A 41	
TYR A 133	SER A 42	
PRO A 234	HIS A 44	
VAL A236	LEU A 121	
LYS A 248	VAL A 123	
TYR A 250		
LEU A 251		
ALA A 252		

The * indicates the position of high risk nsSNPs

STRING database predicted that the functional interaction pattern of PCNA protein to other proteins in a cell. Strong functional associations of the PCNA protein have been observed with FEN1, RFC3, RFC4, RFC5, RFC1, POLH, POLD1, LIG1, CDKN1A and MSH6 (Figure 2).

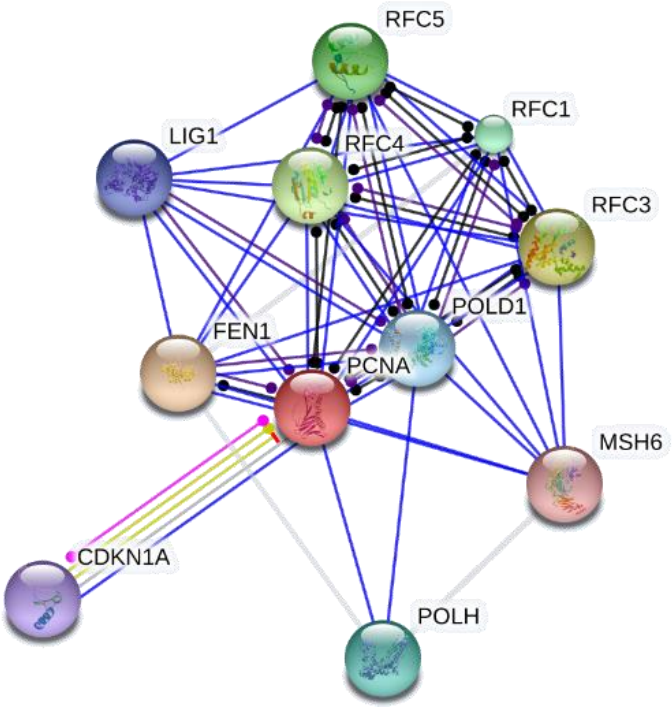


Figure2. PCNA protein-protein interactions in action view.

PCNA interacts with FEN1 (Flap structure specific endonuclease 1), RFC3 (Replication factor C 3), RFC4 (Replication factor C 4), RFC5 (Replication factor C 5), RFC1 (Replication factor C 1), POLH (Polymerase DNA, eta), POLD (Polymerase DNA, delta), LIG1(Ligase 1), CDKN1A (Cyclin-dependent kinase inhibitor 1A) and MSH6 (mutS homolog 6).Strong association pattern is shown by thick blue lines and weak association in the form of thin blue lines.

DISCUSSION

The Single Nucleotide Polymorphisms (SNPs) account for the major cause of variations in humans. Up till now millions of SNPs could be found on NCBI SNP database but due to degeneracy of amino acids and natural selection, many of them do not pose any significant change on protein structure or function. Therefore, it is necessary to distinguish between the functionally neutral and disease associated polymorphisms. As it is difficult to choose and study the SNPs which are more likely to contribute in disease development hence in this condition *in silico* approach is a convenient way to distinguish the damaging SNPs using specific algorithms that can discriminate between neutral and deleterious SNPs [48].Hence an effort was made to identify SNPs that can modify the structure, function and expression of the PCNA gene. As most of the disease associated SNPs are found in the exons or coding regions, also known as non-synonymous SNPs [49,50] in our study we

have submitted the 42 nsSNPs of PCNA to various *in silico* SNP characterizing tools and out of these 42 nsSNPs, 5 nsSNPs (Q38R, S39R, E104G, L182W and K248N) were found to be damaging. For evaluating the harmful effects of the high risk SNPs, we predicted the evolutionary conservation profile of the PCNA gene using the ConSurf server which also identifies the functional regions in the protein [51]. We found that 3 of the high risk nsSNPs (Q38R, E104G and K248N) were identified as important exposed, functional residues and are highly conserved whereas the other two nsSNPs (S39R and L182W) were identified as highly conserved structural, buried residues. This suggests that the high risk nsSNPs may alter the structure and function of the PCNA protein. To further validate we performed the multiple sequence alignments for the human PCNA protein with the PCNA from seven different species using CLUSTALX and MUSCLE software. The alignment generated for the PCNA protein was found to be similar throughout the eight species indicating that the PCNA protein is conserved throughout the selected species. PCNA is subjected to PTMs viz phosphorylation, sumoylation, acetylation and ubiquitilation. Posttranslational protein modifications (PTMs) basically alter the functions of their target proteins by blocking, creating or modifying the interaction areas. These modifications thereby translate the modification of the target protein into biological action by affecting the interaction between their targets and other cellular factors or proteins [52]. Mutations in the post translational target sites leading to gain or loss of the PTMs may be involved in the human diseases [53, 54]. We investigated the putative PTM sites in the PCNA protein and found that Serine at 39^{th} position was a recognizable phosphorylation site whereas lysine at 248^{th} position was the putative acetylation site. Thus, mutations at these positions lead to loss of PTM sites and may play a role in disease manifestation [55]. Amino acid substitutions can possibly disturb the ligand binding positions that are important in protein function and may lead to modifications in the protein stability and structure [56]. The DDG stability predictions and RSA calculations performed using the I Mutant 2.0 software indicate that the selected mutations decreased the protein stability. We further compared the RMSD values of the mutants and the wild type model and they were found to be similar. Therefore, it could be proposed that these mutations do not bring about any significant alteration in the mutant structures with regard to the native protein structure. We also performed energy minimization calculations and observed that the five mutant modeled structures exhibited deviation from native structures before and after energy minimization. Q38R, E104G and L182W mutants showed increase in free energy (less favorable change) after minimization whereas S39R and K248N showed decrease in free energy in comparison to the native structure. We have analyzed the ligand binding sites of the PCNA protein and found that out of the five high risk nsSNPs only three nsSNPs sites namely Q38,

S39 and K248 are involved in the ligand binding sites. In addition to this through STRING database we have observed that the PCNA protein shows strong associations with the FEN1, RFC1, RFC3, RFC4, RFC5, POLH, POLD1, LIG1, CDKN1A and MSH6. Since Q38, S39 and K248 are a part of different ligand binding sites and as PCNA protein shows interaction with a lot of other proteins, this implies that mutations at 38, 39 and 248 may alter the association of the PCNA with any of its interacting proteins. As mentioned earlier few polymorphisms of PCNA are associated with lung cancer, neurodegenerative disorders and DNA repair disorders. However, there was no study that reported the association between these deleterious nsSNPs (rs780735449, rs1050525, rs781573975, rs772308650 and rs753494859) and incidence of any disorders. Hence the validation of these high risk nsSNPs in any disease is required to complement the existing limited body of knowledge. The analysis of nsSNPs in the PCNA gene by using computational methods would help in the establishment of their effects on the protein functional characteristics. Precisely this *in silico* approach permits the estimation of probability of amino acid changes which can be detrimental for the PCNA protein functions.

4. CONCLUSION

By utilizing the publicly available databases we could sort out the nsSNPs which could be possibly dangerous to the functioning of the PCNA. This is evident through the PTM and ligand binding analysis. These high risk nsSNPs should be further validated by wet laboratory experimentation.

CONFLICT OF INTEREST

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

AUTHOR CONTRIBUTIONS

Experiment design and concept- MK, VC and SS. Performed the experiments and analyzed the data- VC & MK. Paper drafted and written- VC, MK and SS.

ACKNOWLEDGEMENTS

We are thankful to the Government city college, Hyderabad for providing platform for this research work. We are also thankful to Synteny Life sciences Pvt. Ltd, Hyderabad for providing facilities for our research work.

REFERENCES

1. Pal LR, Yu CH, Mount S, Moult J. Insights from GWAS: emerging landscape of mechanisms underlying complex trait disease. BMC Genomics. 2015; 16 (suppl 8):S4. doi: 10.1186/1471-2164-16-S8-S4.
2. Sebastiani P, Solovieff N, Hartley SW, Milton JN, Riva A, Daniel A, et al. Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study. Am J Hematol. 2010; 85:29-35. doi: 10.1002/ajh.21572.
3. Badens C, Joly P, Agouti I, Thuret I, Gonnet K, Fattoum S, et al. Variants in genetic modifiers of b-thalassemia can help to predict the major or intermedia type of the disease. Haematologica. 2011; 96(11):1712-1714. doi: 10.3324/haematol.2011.046748.
4. Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am. J. Hum. Genet. 2004; 75:330-337.
5. Figueroa JD, Malats N, Real FX, Silverman D,Kogevinas M, Chanock S, et al. Genetic variation in the base excision repair pathway and bladder cancer risk. Human Genetics. 2007; 121:233-242. doi: 10.1007/s00439-006-0294-y.
6. Cai Z, Han S, Li Z, He L, Zhou J, Huang W, et al. A genome-wide assessment of variations of primary colorectal cancer maintained in metastases. Gene. 2016; 595: 18-24. doi: 10.1016/j.gene.2016.09.023.
7. Matsumoto K, Moriuchi T, Koji T, Nakane PK. Molecular cloning of cDNA coding for rat proliferating cell nuclear antigen (PCNA)/cyclin. EMBO J. 1987; 3:637-642.
8. Gulbis JM, Kelman Z, Hurwitz J, O'Dennell M, Kuriyan J. Structure of the C-terminal region of p21 (WAF1/CIP1) complexed with human PCNA. Cell. 1996; 87(2):297-306.
9. Krishna TS, Kong XP, Gary S, Burgers PM, Kuriyan J. Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. Cell. 1994; 79:1233-1243.
10. Moldovan GL, Pfander B, Jentsch S. PCNA, the maestro of the replication fork. Cell. 2007; 129(4):665–679.
11. Mailand N, Gibbs-Seymour I, Bekker-Jensen S. Regulation of PCNA protein interactions for genome stability. Nat Rev Mol Cell Biol. 2013; 14: 269–282. doi:10.1038/nrm3562.
12. Stoimenov I, Helleday T. PCNA on the crossroad of cancer. BiochemSoc Trans. 2009; 37: 605–613. doi: 10.1042/BST0370605.

13. Zienolddiny S, Campa D, Lind H, Ryberg D, Skaug V, Stangeland L, et al. Polymorphisms of DNA repair genes and risk of non-small cell lung cancer. Carcinogenesis. 2006; 27:560–567.
14. Dieckman LM, Boehm EM, Hingorani MM, Washington MT. Distinct structural alterations in proliferating cell nuclear antigen block DNA mismatch repair. Biochemistry. 2013; 52 (33): 5611-9. doi: 10.1021/bi400378e.
15. Baple EL, Chambers H, Cross HE, Fawcett H, Nakazawa Y, Chioza BA, et al. Hypomorphic PCNA mutation underlies a human DNA repair disorder. J Clin Invest. 2014; 124(7):3137-46. doi: 10.1172/JCI74593.
16. Duffy CM, Hilbert BJ, Kelch BA.A Disease-Causing Variant in PCNA Disrupts a Promiscuous Protein Binding Site. J Mol Biol. 2016; 428(6):1023-40. doi: 10.1016/j.jmb.2015.11.029
17. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29: 308–311.
18. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, et al. Ensembl 2013. Nucleic Acids Res. 2013; 41: D48–55.
19. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001; 11: 863–874.
20. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7: 248–249. doi: 10.1038/nmeth0410-248.
21. Bao L, Zhou M, Cui Y. nsSNP Analyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005; 33: W480–482.
22. 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; 22: 2729–2734.
23. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat. 2009; 30: 1237–1244. doi: 10.1002/humu.21047.
24. Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease associated single amino acid polymorphisms in terms of sequence and structure properties. J MolBiol. 2002; 315: 771–786.
25. Larkin M.A, Blackshields G, Brown N.P, Chenna R, McGettigan P.A, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007; 23:2947-2948.

26. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Research. 2015; 43:W580-W584. doi: 10.1093/nar/gkv279.
27. 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; 44:W344-350.
28. Radivojac P, Vacic V, Haynes C, Cocklin R.R, Mohan A, Heyen J.W, et al. Identification, analysis and prediction of protein ubiquitination sites. Proteins. 2010; 78, 365–380. doi:10.1002/prot.22555.
29. Gill G. Post-translational modification by the small ubiquitin-related modifier SUMO has big effects on transcription factor activity. CurrOpin Genet Dev. 2003; 2: 108–113.
30. Zhao Q, Xie Y, Zheng Y, Jiang S, Liu W, Mu W, et al. GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acid Research.* 2014; 42:W325-W330. doi: 10.1093/nar/gku383.
31. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J MolBiol. 1999; 294: 1351– 1362.
32. Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, et al. GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Eng Des Sel. 2011; 24: 255–260. doi: 10.1093/protein/gzq094.
33. Chou KC. Some remarks on predicting multi-label attributes in molecular biosystems. Molecular Biosystems. 2013; 9(6):1092-1100. doi: 10.1039/c3mb25555g.
34. Chen W, Lin H, Chou K.C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. MolBiosyst, 11(10):2620-34. 2015; doi: 10.1039/c5mb00155b.
35. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016; 497:48-56. doi: 10.1016/j.ab.2015.12.009.
36. 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: W306– W310.
37. Johansson MU, Zoete V, Michielin O, Guex N. Defining and searching for structural motifs using DeepView/Swiss-PdbViewer. BMC Bioinformatics. 2012; 13:173. doi: 10.1186/1471-2105-13-173.

38. Zhang Y, Skolnick J, TM-align: A protein structure alignment algorithm based on TM-score. Nucleic Acids Res. 2005; 33: 2302-2309.
39. Krieger E, Joo K, Lee J, Raman S, Thompson J, Tyka M, et al. Improving physical realism, stereochemistry and side-chain accuracy in homology modeling: four approaches that performed well in CASP8. Proteins. 2009;77 Suppl 9:114-22. doi: 10.1002/prot.22570.
40. Lovell SC, Davis IW, Arendall WB, de Bakker PI, Word JM, Prisant MG, Richardson JS, Richardson DC. Structure validation by C alpha geometry: phi, psi and Cbeta deviation. Proteins. 2002; 50: 437-450.
41. Ngan CH, Hall DR, Zerbe B, Grove LE, Kozakov D, Vajda S. FTSite: high accuracy detection of ligand binding sites on unbound protein structures. Bioinformatics. 2012; 28: 286–287. doi: 10.1093/bioinformatics/btr651.
42. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins globally integrated and scored. Nucleic Acids Research. 2012; 39:D561-568. doi: 10.1093/nar/gkq973.
43. Reddy BV, Li WW, Shindyalov, Bourne PE. Conserved key amino acid positions (CKAAPs) derived from the analysis of common sub structures in proteins. Proteins. 2001; 42:148-163.
44. Carpenter BM, Gancz H, Benoit SL, Evans S, Olsen CH, et al. Mutagenesis of conserved amino acids of Helicobacter pylori fur reveals residues important for function. Journal of Bacteriology. 2010; p.5037-5052. doi: 10.1128/JB.00198-10.
45. 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 Application Note. 2004; 20: 1322-1324.
46. Liberman AC, Noguerol MA, Artz E. Modulation of the Glucocorticoid Receptor Activity by Post-Translational Modifications. Nuclear Receptor Research. 2014; Vol 1. doi: 10.11131/2014/101086.
47. Dai C, and Gu W. p53 post-translational modification: deregulated in tumorigenesis. Trends Mol Med. 2010; 16(11): 528-536. doi: 10.1016/j.molmed.2010.09.002.
48. Ali SK, Doss CG, Kumar DT,Hailong Zhu H. CoagVDb: a comprehensive database for coagulation factors and their associated SAPs. Biol Res. 2015; 48:35. doi: 10.1186/s40659-015-0028-5.
49. Singaraja RR, Brunham LR, Visscher H, Kastelein JJ, Hayden MR. Efflux and atherosclerosis: the clinical and biochemical impact of variations in the ABCA1 gene. ArterioscIerThrombVasc Biol. 2003; 23:1322-32.

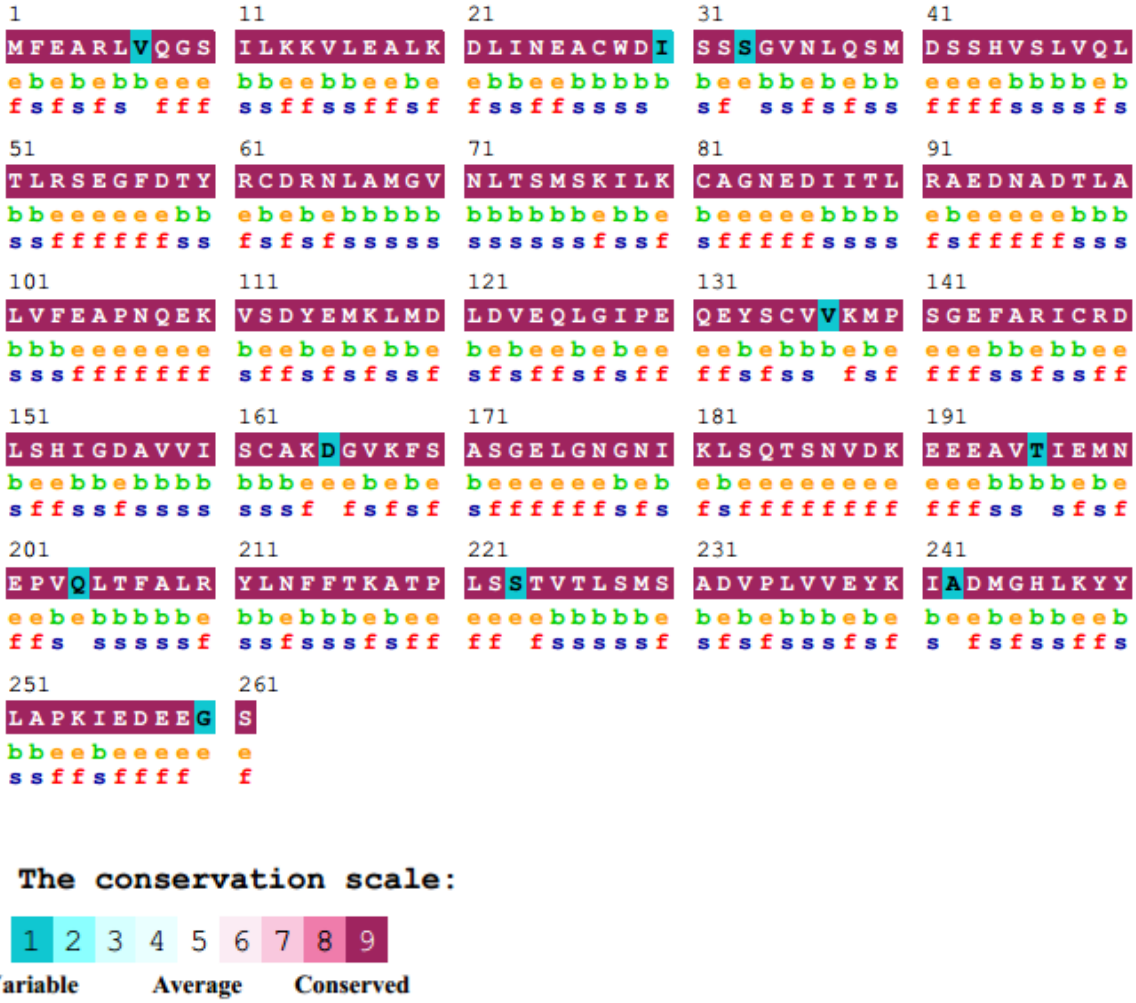
50. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD). Hum Mutat. 2003; 21:577-581.
51. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, et al. ConSurf: identification of functional regions in proteins by surface mapping of phylogenetic information. Bioinformatics. 2003; 1: 163–164.
52. Ulrich HD, Takahashi T. Readers of PCNA modifications. Chromosoma. 2013; 122(4):259-74. doi: 10.1007/s00412-013-0410-4.
53. Grasbon-Frodl E, Lorenz H, Mann U, Nitsch RM, Windl O,Kretzschmar HA. Loss of glycosylation associated with the T183A mutation in human prion disease. ActaNeuropathol. 2004; 108(6):476–84.
54. Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW, Mooney SD. Gain and loss of phosphorylation sites in human cancer. Bioinformatics. 2008; 24:241–247. doi: 10.1093/bioinformatics/btn267.
55. Narayan S, Bader GD, Reimand J. Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer. Genome Medicine. 2016; 8:55. doi: 10.1186/s13073-016-0311-2.
56. Doss CGP, Rajith B, Garwasis N, Mathew PR, Raju AS, Apoorva K, et al. Screening of mutations affecting protein stability and dynamics of FGFR1—A simulation analysis. Applied & Translational Genomics. 2012; 1: 37-43.

SUPPLEMENTARY FILES

Supplementary Table 1. SNP PREDICTIONS FOR ALL THE PCNA NSSNPS.

S.No	SNP	Amino acid position	Mutation	SIFT	PoplyPhen	nsSNP analyzer	Phd SNP	SNP & Go	P Mut
1	rs776734661	14	K14R	Tolerable	Benign	Neutral	Neutral	Neutral	Neutral
2	rs375496467	15	V15M	Deleterious	Possibly damaging	Neutral	Neutral	Neutral	Disease
3	rs749287681	24	N24S	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
4	rs779810002	24	N24K	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
5	rs769487182	36	N36I	Deleterious	Benign	Neutral	Disease	Neutral	Disease
6	rs780735449	38	Q38R	Deleterious	Probably damaging	Disease	Disease	Neutral	Disease
7	rs1050525	39	S39R	Deleterious	Possibly damaging	Disease	Disease	Disease	Disease
8	rs751096723	45	V45I	Deleterious	Probably damaging	Neutral	Neutral	Neutral	Neutral
9	rs759826699	55	E55A	Deleterious	Benign	Neutral	Neutral	Neutral	Disease
10	rs754051671	56	G56S	Deleterious	Possibly damaging	Disease	Neutral	Neutral	Disease
11	rs752756450	59	T59A	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
12	rs867201351	64	R64H	Deleterious	Possibly damaging	Neutral	Disease	Neutral	Disease
13	rs144468297	65	N65T	Deleterious	Benign	Neutral	Neutral	Neutral	Disease
14	rs141842220	67	A67T	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
15	rs769587124	68	M68I	Tolerable	Possibly damaging	Neutral	Neutral	Neutral	Disease
16	rs776541305	74	S74C	Tolerable	Benign	Disease	Disease	Neutral	Disease
17	rs760578899	76	S76F	Tolerable	Probably damaging	Disease	Neutral	Neutral	Disease
18	rs771828986	78	I78V	Tolerable	Benign	Neutral	Neutral	Neutral	Neutral
19	rs773719325	84	N84S	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
20	rs770091696	98	T98A	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
21	rs757200275	102	V102I	Tolerable	Benign	Neutral	Neutral	Neutral	Neutral
22	rs781573975	104	E104G	Deleterious	Probably damaging	Disease	Disease	Neutral	Disease
23	rs757469185	106	P106L	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
24	rs768007592	108	Q108H	Deleterious	Benign	Neutral	Neutral	Neutral	Disease
25	rs775027645	111	V111I	Tolerable	Benign	Neutral	Neutral	Neutral	Neutral
26	rs764789172	137	V137I	Tolerable	Benign	Neutral	Neutral	Neutral	Neutral
27	rs376351202	139	M139V	Deleterious	Possibly damaging	Neutral	Neutral	Neutral	Disease
28	rs777131504	147	I147M	Deleterious	Possibly damaging	Neutral	Neutral	Neutral	Neutral
29	rs771163574	151	L151I	Deleterious	Probably damaging	Neutral	Disease	Neutral	Neutral
30	rs772308650	182	L182W	Deleterious	Probably damaging	Disease	Disease	Neutral	Disease
31	rs769262502	189	D189N	Deleterious	Probably damaging	Neutral	Neutral	Neutral	Disease
32	rs779077153	194	A194S	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
33	rs768807139	199	M199I	Deleterious	Benign	Neutral	Neutral	Neutral	Disease
34	rs140522967	223	S223P	Tolerable	Benign	Neutral	Disease	Neutral	Disease
35	rs369958038	228	S228I	Deleterious	Possibly damaging	Disease	Neutral	Neutral	Disease
36	rs769780505	236	V236A	Deleterious	Benign	Disease	Neutral	Neutral	Disease
37	rs746900735	242	A242V	Tolerable	Benign	Neutral	Neutral	Disease	Disease
38	rs747916514	244	M244I	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
39	rs771809068	244	M244V	Tolerable	Benign	Neutral	Neutral	Neutral	Disease
40	rs754622048	246	H246Y	Tolerable	Benign	Neutral	Disease	Neutral	Neutral
41	rs753494859	248	K248N	Deleterious	Probably damaging	Disease	Neutral	Neutral	Disease
42	rs145675493	257	D257E	Tolerable	Benign	Neutral	Neutral	Neutral	Neutral

Supplementary Figure 1: Prediction of putative functional and structural residues by Consurf for the PCNA protein.



e - An exposed residue

b - A buried residue

f - A predicted functional residue (highly conserved and exposed)

s - A predicted structural residue (highly conserved and buried)