Pathogenic Genetic Variants From Highly Connected Cancer Susceptibility Genes Confer the Loss of Structural Stability

Mahjerin Nasrin Reza
Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

Nadim Ferdous
Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

Md. Tabassum Hossain Emon
Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

Md. Shariful Islam (sharifbge@gmail.com)
University of Kentucky

A. K.M. Mohiuddin
Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh

Mohammad Uzzal Hossain
National Institute of Biotechnology

Research Article

Keywords: Genetic polymorphisms, single nucleotide polymorphisms, mechanics/Poisson Boltzmann surface area

Posted Date: May 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-480522/v1

Abstract

Genetic polymorphisms in DNA damage repair and tumor suppressor genes have been associated with increasing the risk of several types of cancer. Analyses of putative functional single nucleotide polymorphisms (SNP) in such genes can greatly improve human health by guiding choice of therapeutics. In this study, we selected nine genes responsible for various cancer types for gene enrichment analysis and found that BRCA1, ATM, and TP53 were more enriched in connectivity. Therefore, we used different computational algorithms to classify the nonsynonymous SNPs which are deleterious to the structure and/or function of these three proteins. Our study demonstrated that V1687G and V1736G variants of BRCA1, I2865T and V2906A variants of ATM, V216G and L194H variants of TP53 are major mutations with pathogenic impact and are likely to have a greater impact on destabilizing the proteins. To stabilize the high-risk SNPs, we performed mutation site-specific molecular docking analysis and validated using molecular dynamics (MD) simulation and molecular mechanics/Poisson Boltzmann surface area (MM/PBSA) studies. Additionally, SNPs of untranslated regions of these genes affecting miRNA binding were characterized. Hence, this study will assist in developing precision medicines for cancer types related to these polymorphisms.

Introduction

The burden of non-communicable diseases is growing fast across the world and cancer is one of them. Genetic factors play a significant role in the development of a majority of cancers as in these cases, genes regulating cell division, apoptosis, invasiveness, or metastasis undergo mutation. Several important cancers such as breast, ovarian, esophageal, lung, colon, colorectal, melanoma, pancreatic cancers, etc. develop due to the accumulation of mutations in BRCA1[1], BRCA2[1], ATM[2], TP53[3], MSH2[4], MLH1[4], MSH6[5], CDKN2A[6], and PALB2[7] genes. Mutations in the BRCA1 and the BRCA2 genes are responsible for 90% of hereditary breast cancer and the majority of hereditary ovarian cancer cases[1,8–10]. The ATM, BRCA1, BRCA2, and TP53 are known as “caretaker” genes involved in DNA repair and function in the maintenance of genomic stability[2]. Mutations in ATM lead to T-cell prolymphocytic leukemia, B-cell chronic lymphocytic leukemia and it was also observed at increased frequency in breast cancer cases[2]. Somatic TP53 mutations occur in ovarian, esophageal, colorectal, head and neck, larynx, and lung cancers at rates ranging from 38%–50%[3]. MLH1 and MSH2 mutations are known to be associated with lifetime ovarian and endometrial cancer[4]. Increased risk of colorectal carcinoma occurs in the MSH6 mutation carriers[5]. Germline alterations in CDKN2A have been identified in melanoma cases[6]. Apart from mutations in major genes responsible for breast cancer, germline mutations in PALB2 have been identified in familial breast cancer and familial pancreatic cancer cases[7]. Insights into the molecular mechanisms underlying the effects of mutations that result in cancer susceptibility are indispensable considering their crucial roles in cell cycle regulation, metabolism, DNA mismatch repair, and immunity.

Single nucleotide polymorphisms (SNPs) are the most frequent type of genetic alteration in humans consisting of about 90% of sequence variants[11]. They serve as indicators in the detection of part of the

genome involved in disease[12]. They are dispersed throughout the genome in both coding and regulatory regions of genes and the most important SNPs are the ones that are in the coding region of the human genome consisting of around 500,000 SNPs[13]. Among these, missense SNPs, also known as non-synonymous SNPs (nsSNPs) are especially significant as they are responsible for amino acid substitutions resulting in structural and functional variations of proteins[14]. Genetic tools and comparative genomics have been used for large-scale extraction of SNPs over the years. More than 10 million SNPs have been identified covering the most common polymorphisms[15]. Currently, the most common polymorphisms that have been identified covers more than 10 million SNPs[15]. DNA sequencing is the first step towards understanding the nature of a variation[16]. With the advancement of computational algorithms, cost-effective, robust, and refined methods are being created, helping in the development of high-throughput tools for the identification of structural and dynamic changes of protein products as a result of SNPs. Although most of the SNPs cause damages to the protein, some of them are neutral[17]. Therefore, a proper selection of computational methods and algorithms are required for the prediction of structural and functional consequences of the target proteins to distinguish the deleterious nsSNPs from the neutral ones.

Most of the proteins carry out functions primarily through their integral domains. These are independent units having potentially different biological functions and they can be gained by proteins in order to acquire novel functions[18]. Domains are, therefore, defined to be the functional units through which proteins evolve. Numerous computational studies have been conducted to analyze the nsSNPs of cancer susceptibility genes. A study analyzed genetic variations in the BRCA1-associated RING domain protein encoded by the BARD1 gene and predicted their deleterious effects causing breast, ovarian and uterine cancers[19]. Similar study identified missense SNPs in the homeobox domain protein encoded by the human HOXB13 gene which is responsible for hereditary prostate cancer[20]. The nsSNPs in the RASSF5 gene that plays a crucial role as a tumor suppressor were found to reduce the binding affinity with H-Ras protein[21]. As nsSNPs result in amino acid substitutions of a protein, presence of these polymorphisms in domains alters protein interactions, functions and post-translational modifications[22]. But majority of these studies focus on nsSNPs located on regions other than the significant protein domains. All the analyses were performed on a single gene and there was no connection shown with other cancer susceptibility genes in these studies. Also, there were no further experiments conducted to stabilize these deleterious nsSNPs keeping a gap in identifying the choice of probable therapeutics.

In this study, we conducted our investigation to identify and evaluate the pathogenic nsSNPs in highly connected cancer susceptibility genes, their associations in causing diseases, and the effect of these deleterious nsSNPs in the structural behavior of proteins. We carried out gene enrichment analysis using the Enrichr[23] web server. Further we employed SIFT[24], PolyPhen-2[25], PROVEAN[26], PhD-SNP[27], PANTHER[28], SNPs&GO[29], SNAP2[30], PredictSNP[31], I-Mutant[32] to postulate the nsSNPs and to evaluate their pathogenic effects on their respective proteins. NCBI-CD search was used to find out the protein domains while ConSurf[33] was employed for the identification of amino acid conservations. We modeled

the structure of wild-type (WT) and mutant proteins by MODELLER 9.22[34]. Secondary structures and physico-chemical properties of proteins were analyzed using the Stride[35] and ProtParam[36] servers. In order to investigate the rescue mechanisms of the damaging substitutions, we performed mutation site targeted protein-ligand docking with PhiKan083 (PK083), a carbazole derivative using Molegro Virtual Docker[37]. For a clear depiction of the dynamic behavior of WT and mutants over time, Molecular Dynamics simulation was performed by GROMACS 5.1.4[38]. Furthermore, binding free energies were calculated using Molecular mechanic/Poisson–Boltzmann Surface Area (MM/PBSA) approach to validate the favorable binding of PK083 to all the mutants. Finally, we utilized the PolymiRTS v3.0 database to identify the functional impact of 3' and 5' SNPs in the binding of miRNA.

Results

Gene enrichment

From the Enrichr result, we found that BRCA1 and ATM seem to have high likelihood and greater degree of interaction than other genes (Figure 2). The p-value was computed from the fisher's exact test, a proportion test assuming binomial distribution and independence of probability of any gene belonging to any set (Table 1). Based on the enrichment analysis and p-value, we selected BRCA1, ATM, and TP53 gene for analysis of SNPs.

SNP datasets retrieved from dbSNP database

We retrieved the SNPs of BRCA1, ATM and TP53 genes from dbSNP database. We found a total number of 25,754 SNPs of BRCA1, 41,948 SNPs of ATM and 6,977 SNPs of TP53 in the dbSNP database. The percentage of different types of SNPs of each gene are shown in Figure 3. We selected only the nsSNPs of the three genes for our investigation. Finally, including the multiple allele changes, we have analyzed a total number of 3,467 nsSNPs of BRCA1, 4,650 nsSNPs of ATM and 1,106 nsSNPs of TP53 for our investigation.

Deleterious and damaging nsSNPs in BRCA1, ATM and TP53

We subjected all the nsSNPs of the three genes to SIFT, PolyPhen-2 and PROVEAN tools to investigate the effect of amino acid substitution on the respective protein function. We shortlisted those nsSNPs as highly deleterious that were predicted damaging/probably damaging or deleterious by at least two of these tools. Total 1,013 nsSNPs of BRCA1, 1815 nsSNPs of ATM and 528 nsSNPs of met the criteria and we classified them as highly deleterious nsSNPs. Prediction results by SIFT, PolyPhen2 and PROVEAN are shown in Supplementary file 1-3.

Disease-associated nsSNPs in BRCA1, ATM, and TP53

The above shortlisted nsSNPs were submitted to PhD-SNP, PANTHER, SNPs&GO, SNAP2, and PredictSNP tools to identify the disease-associated nsSNPs. We shortlisted the nsSNPs those were found deleterious

by all the five tools. A total number of 250 nsSNPs of BRCA1, 796 nsSNPs of ATM, and 341 nsSNPs of TP53 were predicted deleterious by all the five tools and were considered for further investigation (Supplementary file 4-6).

Identified nsSNPs in conserved domains

NCBI's conserved domain search tool revealed that each of the BRCA1, ATM, and TP53 proteins were found to have four domains shown in Figure 4. Results showed that the BRCA1 protein had a serine-rich domain associated with BRCT, the first BRCT domain, the second (C-terminal) BRCT domain and a RING finger domain. Originally BRCT domain was identified in the tumor suppressor protein and missense mutations in this region correspond to a high risk for breast and ovarian cancers[9,57]. The ATM protein was found to have a catalytic domain, a FAT domain, a TAN domain, and a FATC domain. The catalytic domain of ATM is pivotal for phosphorylation of dozens of substrates that are involved in repair of DNA double strand breaks[58]. The FAT domain is located adjacent and upstream of the kinase domain and the name of this domain is derived from FRAP (mTOR), ATM, and TRAPP, all of which members of the phosphoinositide 3-kinase-like family[59]. ATM protein also contains a Tel1/ATM N-Terminal Motif (TAN) that is essential for telomere length maintenance and DNA damage response[60]. Search results of p53 protein showed that it had a p53 DNA-binding domain (DBD), a p53 tetramerization motif, the transactivation domain 2 (TAD2), and a p53 transactivation motif. The DNA-binding domain is absent in p63, p73 and other p53 homologues in primitive organisms thus making it a unique feature for the vertebrate p53[61]. A flexible linker region connects the structured DNA-binding and tetramerization domains of p53[62]. There are 2 distinct TADs found in p53 and it was found that TAD2 also play important part in tumor suppression[63]. As domains play significant role in proteins, we shortlisted the disease-associated nsSNPs that fall on the domain sequences of the three proteins. From CD-search results, we found that, 171 nsSNPs of BRCA1, 313 nsSNPs of ATM and all the shortlisted 340 nsSNPs of TP53 occur on the predicted domains and were considered for further analysis (Supplementary file 7-9).

Conservation profile of deleterious nsSNPs in BRCA1, ATM and TP53

We calculated the evolutionary conservation of amino acid residues of the three proteins using ConSurf server to further explore the possible effects of shortlisted nsSNPs. Results were collected in the form of structural representation of the protein sequence. Based on the location either on protein surface or inside its core, the highly conserved residues are predicted as either functional or structural. Amino acids involved in various vital biological processes appear to be more conserved than others. Considering this, the nsSNPs located at these conserved regions are highly damaging to proteins as compared to those located at non-conserved sites[64,65]. Hence, we focused only on the residues of domains matching their positions with the shortlisted high risk nsSNPs. The results predicted that out of the shortlisted nsSNPs, 89 nsSNPs of BRCA1, 218 nsSNPs of ATM and 282 nsSNPs of TP53 were highly conserved (either exposed or buried). Supplementary file 10-12 contains the graphical representation of ConSurf results of the three proteins.

Predicted stability modification

We analyzed the stability alterations in the three proteins using I-Mutant server which completed this task by considering the amino acid substitutions. I-mutant 2.0 results revealed that all the 89 nsSNPs of BRCA1, 192 nsSNPs of ATM, 245 nsSNPs of TP53 decrease stability of the respective proteins (Supplementary file 13-15). Thus, these polymorphisms in the protein domains might cause supreme damage to the protein affecting their stability. According to some studies, phenomenon such as increase in degradation, misfolding and aggregation of proteins are caused by decreased protein stability[66–68]. So, considering the lower ΔΔG values, we finally selected top five nsSNPs (Table 2) from each protein and considered them for 3D structure modeling.

Characterized functional effects of SNPs in 3’ and 5’ untranslated regions

PolymiRTS database predicted the list of miRNAs disrupted and created by SNPs of ATM and TP53 genes. Interestingly, no miRNAs were found to be disrupted or created by BRCA1 gene from the database. Supplementary file 16 contains the tables (Supplementary table S1-S4) showing the effect of SNPs in the 3’ and 5’ region of ATM and TP53 genes. Higher conservation score indicates greater effect of the SNPs. In addition, higher context+score denotes higher likelihood of disruption or creation that occurs in the miRNA target site. The miRNAs greatly affected by the SNPs of ATM and TP53 genes based on highest conservation score are shown in Table 3.

Structural insights into BRCA1, ATM, p53 domains and their mutants

The shortlisted 15 nsSNPs of BRCA1, ATM and TP53 were taken in order to the 3D structures of mutated domains. All the nsSNPs of BRCA1 fall on the first BRCT domain of BRCA1 protein. The shortlisted nsSNPs of ATM and TP53 fall on the catalytic domain and the DNA-binding domain of the two proteins respectively. The crystal structure of BRCA1 BRCT (PDB ID: 4JLU) and p53 DNA binding domain (PDB ID: 2PCX) were available in the RCSB PDB database. These two structures were retrieved and cleaned by removing the ligands and inhibitors (Figure 5). The “Mutagenesis” tool of Pymol was utilized to carry out mutation with the selected five nsSNPs of each protein. The 3D structure of the catalytic domain of ATM was not available in PDB database. So, the structure was modeled using MODELLER 9.22 and was later refined using GalaxyRefine server. The Ramachandran plot analysis, ERRAT server and ProSA-Web analysis results of the refined structure are shown in Supplementary file 16 (Supplementary figure S5). Later the structural analysis was extended by calculating the RMSD values for each mutant model using Maestro 11.8 tool of the Schrodinger suite. The average distance among all atoms, α-carbon atoms and backbones of WT and mutant models were measured from RMSD values. Greater RMSD value indicates greater deviation of mutant structures from that of the WT proteins. The mutant models for V1687G and V1736G of BRCT, I2865T and V2906A of Catalytic domain, V216G and L194H of p53 exhibited the maximum RMSD values shown in Table 4. From the Stride server results, it was found that V1687G and V216G mutations were in β-strand region, V1736G, V2906A and L194H mutations were in coil region and I2865T mutation was in α-helix region of the respected proteins (Supplementary Figure S6). Also, majority

of the mutations caused significant increase in solvent accessible area (Supplementary Table S7). Analysis from ProtParam server revealed that all the mutants had lower instability index and aliphatic index than the WT proteins (Supplementary Table S8).

Stabilization of high-risk mutants by PK083

PK083 belongs to a class of organic compounds known as carbazoles and they contain a three-ring system containing a pyrrole ring fused on either side to a benzene ring. Molecular docking of PK083 with six mutants exhibited nearly same score in binding affinity. The binding affinity of this molecule with the mutants ranges from -32.8 kcal/mol to -72.3 kcal/mol shown in Table 5. PK083 binds at the binding pockets of mutation positions as the defined binding sites of three proteins. Interaction of PK08 with the mutants were found to be largely hydrophobic (Figure 6). Pi-Sigma, Pi-Alkyl, Pi-Lone Pair and conventional hydrogen bonding interactions were observed in the binding of PK083 with the mutant residues.

Molecular dynamics (MD) simulation

As physiological conditions are not considered in evaluating the damaging nature of the mutants using computational tools, we performed MD simulation of both the WT protein domains and the six mutants to view the various conformations that they might acquire in the solvated state. Their dynamic behavior was analyzed by RMSD, RMSF, Rg, and SASA analysis shown in Figure 7 while, Table 6 contains the average values obtained from trajectory analyses.

Configuration changes of all the WT and mutant proteins were analyzed in terms of RMSD during the simulation period. Figure 7(A1-A2) depicts that RMSD values from the mutant structures are quite unstable comparing with the WT-BRCT and WT-DBD. The WT BRCT shows steady fluctuation throughout the 100 ns in the WT structure. V1687G and V1736G structures showed similar way of deviation till 45 ns from their starting structure, after that, fluctuated up to ~0.3 nm for V1687G. RMSD values of V216G and L194H were higher than the WT protein's RMSD at major points demonstrating that the mutations have considerable destabilizing effects on DBD. Interestingly, RMSD values of I2865T were lower than the WT-catalytic domain whereas significant fluctuation was observed in the RMSD values of V2906A with several spikes of ~1 nm from 25 ns to 32 ns period. We have monitored the RMSF to calculate the average fluctuation of amino acid residue in order to determine the mutation's effect on the protein residues dynamic behavior. From Figure 7(B1-B3), it can be inferred that residue level fluctuations for V1736G, V2906A and V216G were quite high, up to ~0.5 nm, ~0.6 nm and ~1 nm respectively when compared with native proteins and other mutations. Analysis of the fluctuations also revealed that the greatest degree of flexibility was shown by the V2906A mutant. We have also analyzed the radius of gyration (R_g) for the WT proteins along with its associated mutations contributing to their compactness shown in Figure 7(C1-C5). From Table 6, it can be concluded that V1687G and V216G had approximately similar compactness as of their respective WT proteins. Rather than these, all the mutations possessed higher Rg values than their WT proteins suggesting their structural destabilizing effects caused by the

mutations, ultimately leading to the loss of protein compactness. Finally, the solvent-accessible surface areas (SASAs) were analyzed to understand the changes in the protein volume upon mutation. V1687G, I2865T and V2906A mutants showed increased SASA values compared to their WT proteins. The decreased SASA value in the remaining mutants denotes their relatively shrunken nature as compared to the WT structures. The change of SASA value of WT and mutant proteins with time is shown in Figure 7(D1-D3).

Stability of the docked protein-ligand complexes

We assessed the stability of the docked complexes during simulation period analyzing RMSD of the protein backbone and ligand structure as well as the hydrogen bonds analysis formed by PK083 with the mutants (Figure 8).

From the RMSD graph, it was found that rather than the V2906A complex, all the remaining protein–ligand complexes were stable. For the five stable complexes, the RMSD was less than 0.1 nm indicating the initial ligand-backbone contacts remained intact during the simulation period. In case of V2906A, the PK083 showed multiple binding orientations and it re-equilibrated several times during the 100 ns simulation (Figure 8B2). It was also observed that the protein backbone RMSD of I2865T and have significantly decreased upon binding of PK083 than the apo form shown in Figure 8(A2). Further, we calculated the number of hydrogen bonds formed during the simulations period for the mutant complexes, as presented in Figure 8(A3-C3) as hydrogen bonding is one of the principal components responsible for the molecular interactions in biological systems. In the V1736G-PK083 complex, highest number of conformations formed up to three hydrogen bonds during the simulation. A very few conformations showed less than two hydrogen bonds. Except the V1687G-PK083 complex, the conformations of the rest of the complexes formed up to two hydrogen bonds throughout the simulation. The simulation trajectories of all the complexes were further exploited to study the interaction between the mutants and PK083.

Post molecular dynamics binding free energy calculation

We calculated the binding free energy of the last 20 ns with an interval of 50 ps (picoseconds) from MD trajectories using MM/PBSA method. We also utilized the MmPbSaStat.py script included in g_mmpbsa package calculating the average free binding energy and its standard deviation/error from the simulation output files (Table 7). The interaction between a ligand and protein is shown in the form of binding energy where lesser the binding energy, the better is the binding of the ligand and protein. The cumulative sum of van der Wall, electrostatic, polar solvation, and SASA energy is the final binding energy. PK083 showed the least binding free energy (-113.211 kJ/mol) with the V216G variant of p53 among the mutants. The carbazole derivative showed almost similar binding free energy with the two variants of BRCT. By plotting the binding energy versus time graphs, a comparison of the binding free energies of all the six complexes were made, shown in Figure 9. These results verify that PK083 might possess stabilizing effect on majority of the deleterious mutations of BRCA1, ATM and TP53 effectively.

Further, we identified the contribution of each residue of the six mutants in terms of binding free energy to the interaction with PK083. By decomposing the total binding free energy of the system into per residue contribution energy, the contribution of each residue was calculated, shown in Figure 9. This gives us an insight into the 'hotspot' residues that contributes favorably to the binding of this molecule to the mutants. It was found that except the V2906A variant of ATM, more than five residues of the remaining mutants contributed higher than -1 KJ/mol binding energy. These identified key residues from our analysis will facilitate the study of mutation sites stabilization of three significant domains of these proteins.

Discussion

In human genome, non-synonymous single nucleotide polymorphisms account for about 50% of allele variation of all hereditary diseases[69]. It can help in improving medication strategies by facilitating more tailored personalized treatment to patients[70]. Also, new compounds can be tested to correct the effects of those mutations studying the effects generated by nsSNPs in disease-associated proteins. Identification of such nsSNPs responsible for specific phenotypes using molecular approaches is time-consuming and expensive[71]. Bioinformatics predicting approaches can help in narrowing down the number of high-risk pathogenic nsSNPs to be screened in genetic association studies, and in a better understanding of the function and structure of protein products.

In this study, we performed an intensive in silico evaluation to identify pathogenic nsSNPs of BRCA1, ATM and TP53 genes using a wide variety of computational tools. We selected these genes based on gene enrichment analysis from a list of nine genes. Only two studies have been carried out to evaluate the nsSNPs of human BRCA1 and ATM gene previously. Previously, a mutation of proline to serin at position 1812 as a main target mutation in the BRCA1 gene was reported analyzing only 65 nsSNPs of this gene in 2007[72]. Also, upon analyzing the functional impact of 168 nsSNPs of ATM gene using two computational tools, SIFT and PolyPhen in 2012, six nsSNPs were classified as highly damaging substitutions[73]. We expanded our study to include all the nsSNPs currently available in dbSNP database and hypothesized that a more reliable and precise estimation of a substitution consequence could be provided by using a variety of computational methods based on different algorithms to filter the pathogenic and neutral variants.

In this study, we retrieved all the available nsSNPs of the corresponding three genes and annotated them using nine computational tools to distinguish between the functional and neutral variants. Assessing the pathogenicity of functional nsSNPs, we filtered those that occur in the conserved domains of respective proteins. Further, evolutionary conservation analysis revealed that majority of the pathogenic nsSNPs occupy conserved amino acid positions whether they decrease or increase protein stability. The nsSNPs greatly decreasing the stability of proteins were finally selected as the high-risk ones. Combining these results, we found that V1687G and V1736G variants of BRCA1, I2865T and V2906A variants of ATM, V216G and L194H variants of TP53 were highly damaging mutations that greatly decrease protein

stability and might alter their respective protein functions (Table 2). All these six variants were found to occur in BRCT domain, catalytic domain and DNA-binding domain of BRCA1, ATM and p53 respectively (Figure 4). The missense mutations in these domains might cause severe consequences disrupting their ability of functioning. The harmful polymorphic mutations are mainly found to be located in helixes and coil regions of a protein structure[74]. Our secondary structure analysis also revealed that three of these six mutations occur in coil regions, two in β-strand region and one in α-helix region which is in accordance with the previous recognition. Therefore, these mutations might result in significant distortion of the backbone over a turn leading to the likelihood of impaired molecular assembly.

Hence, these novel findings encouraged us to study how the dynamic properties differ from WT and mutant amino acids using MD simulation analysis. The RMSD and RMSF analysis in agreement to each other revealed that all the variants except the I2865T, decreased stability and increased the flexibility of protein (Figure 7A1-7B3). The radius of gyration revealed that the WT proteins showed higher level of compactness throughout the time, whereas all the variants showed differential level of compactness (Figure 7C1-7C3). The SASA analysis showed both increased and reduced volumes gained by the mutants and thereby might be responsible for change in function of protein (Figure 7D1-7D3). Based on the simulation study, we demonstrated that the variants V1687G, V1736G, V2906A, V216G and L194H imparted changes in the native conformation or structure of the BRCA1, ATM and TP53 proteins in any sense of behavior and hence speculated to affect the respective protein function and structure in damaging manner.

Further, we carried out our investigation to stabilize the high-risk mutations using a small molecule inhibitor PK083. Mutation site specific molecular docking analysis revealed that PK083 had a strong binding affinity towards all the six mutation sites of three proteins (Table 5 and Figure 6). For the validation of docking process, we performed MD simulation of six mutant-PK083 complexes over 100 ns and findings demonstrated that PK083 showed RMSD of less than 0.1 nm and no disassociation of bound PK083 is observed throughout the 100 ns simulations (Figure 8A1-8C2). Stability of PK083 during MD simulation were also supported by several H-bonds estimations (Figure 8A3-8C3). The results of MM/PBSA indicated that PK083 binds to all the six mutants efficiently as they exhibit good binding free energies (Table 7). We also identified several key residues essential for binding of PK083 to the mutation sites that will provide valuable insight in drug development against these deleterious mutations (Figure 9).

Our further analysis from PolymiRTS database identified both the target sites disrupted and created by SNPs and INDELs in miRNA seeds. Four miRNAs, hsa-miR-6796-3p, hsa-miR-1233-5p, hsa-miR-525-3p and hsa-miR-548i were affected by the SNPs rs3745198, rs71309450, rs190453265, rs201549145 respectively having high conservation score (Table 3). As a result, the areas affected by those SNPs might have evolutionary important function.

Our study reports that six nsSNPs (V1687G and V1736G of BRCA1, I2865T and V2906A of ATM, V216G and L194H of TP53) in the BRCT, Catalytic and DNA-binding domains are highly damaging to the

structure and function of three proteins. We also found that these mutants are drug targets for PK083 as revealed by molecular interaction results. Several miRNAs were affected by SNPs in the 3' and 5' untranslated regions of ATM and TP53 genes and hsa-miR-6796-3p, hsa-miR-1233-5p, hsa-miR-525-3p and hsa-miR-548i due to their higher likelihood of causing disruption or creation in the miRNA target site. Therefore, our findings could provide a cornerstone to the study of potential therapeutic inventions upon clinical-trial and experimental mutational studies.

Methods

A step-wise protocol was followed to identify the pathogenic nsSNPs of the selected cancer susceptibility genes. The work flow is depicted in Figure 1.

Gene enrichment analysis

Gene symbols of nine selected genes were uploaded to Enrichr[23] web-server that evaluates the biological properties of genes based on enrichment analysis. The z-score method was used for computation of enrichment and a combined scoring method was used to compute a combined P value from Fisher's exact test[23].

Retrieval of nsSNPs

Information of nsSNPs (SNP rs IDs, position, and residue changes) of the selected genes was retrieved from the NCBI dbSNP database[39]. The "Missense" filter was used in the function class and all the nsSNPs were retrieved for analysis.

Prediction of deleterious nsSNPs

Three computational tools, SIFT[24], PolyPhen-2[25], and PROVEAN[26] were used for the identification of deleterious nsSNPs. The SIFT (Sorting Intolerant from Tolerant) webserver predicts tolerated or deleterious substitution for every position of the query sequence based on multiple alignment information. PolyPhen-2 (Polymorphism Phenotyping-2) calculates the functional significance of an allele change by a set of supervised learning algorithms called the Naive Bayes classifier. PROVEAN (Protein Variation Effect Analyzer) also classifies the effect of an amino acid substitution, in-frame insertions and deletions on the biological function of a query protein.

Prediction of disease-associated nsSNPs

Five different web tools, PhD-SNP[27] (Predictor of human Deleterious Single Nucleotide Polymorphisms), PANTHER[28] (Protein ANalysis THrough Evolutionary Relationships), SNPs&GO[29], SNAP2[30] (Screening for Nonacceptable Polymorphisms 2) and PredictSNP[31] were used to detect the disease-associated nsSNPs from the selected genes. These tools use various algorithms such as support vector machines (PhD-SNP and SNPs&GO), Hidden Markov-Model based statistical modeling (PANTHER), and neural network

(SNAP2) to predict the SNPs with functional effects upon utilizing the user-provided sequence information. The PredictSNP tool classifies nsSNPs based on consensus method that combines the output of six different prediction webtools (MAPP, nsSNP Analyzer, PANTHER, PhD-SNP, PolyPhen-1, PolyPhen-2, SIFT, and SNAP) to analyze the effect of nsSNPs on protein function.

Identification of nsSNPs in conserved protein domains

Protein domains are distinct functional and structural units in a protein. They are responsible for a particular function or interaction which contributes to the overall role of a protein[18]. Protein domains can be highly altered by the presence of SNPs and proteins with these domain-altering SNPs contain highly connected nodes in various cellular pathways[40]. So, we intended to find out the nsSNPs that occur on the domains of the proteins encoded by the selected genes. Domain search was carried out at NCBI's CD-Search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) to find out the conserved domains of the selected proteins[41].

Evolutionary conservation analysis

The evolutionary conservation of amino acid substitution was analyzed using ConSurf web server[33]. This server uses an empirical Bayesian inference to automatically analyze evolutionary conservation of amino acid substitutions in protein. The corresponding ConSurf conservation score ranges from 1 to 9, where 1 designates rapidly evolving (variable) regions, 5 designates mildly evolving regions, and 9 indicates conserved regions[42]. The exposed residues having high scores are thought to be functional residues, whereas the buried residues having high scores are considered structural.

Prediction of changes in protein stability

I-Mutant 2.0 was used to analyze the protein stability changes upon nsSNPs. The tool is based on support vector machine (SVM) that provides a free energy change value (ΔΔG) of protein after and before mutation as output and uses data derived from ProTherm database which is one of the most comprehensive of experimental data on protein mutations[32,43]. ΔΔG value of less than '0' indicates that the variant decreases the protein stability and ΔΔG value of greater than 0 indicates that the variant increases the protein stability.

Functional effect analysis of SNPs in 3' and 5' untranslated regions (UTR)

PolymiRTS v3.0 database was used to characterize the SNPs in 3' and 5' regions and to analyze the functional impact of genetic polymorphisms in miRNA seed regions and miRNA target sites. This is an integrated platform for analyzing SNPs that affect miRNA. We entered the gene symbols of the selected genes and acquired a list of the miRNAs affected by these mutations that might lead to a decrease/increase of the expression of genes.

Homology modeling and structural analysis of variants

To evaluate whether the high risk nsSNPs alter the WT structure of protein domains, we analyzed the three-dimensional (3D) structures of both WT and mutant domains. The PDB database (https://www.rcsb.org/) was used to retrieve the available protein structures whereas MODELLER 9.22[34] was used to generate the 3D structures that were not available in the database. The DOPE and GA341 objective functions were used to choose the best structure from MODELLER, where higher GA341 and/or lower DOPE indicates higher quality of a generated model. The best modelled structures were refined through the GalaxyRefine[44] server. The resultant structures were verified by PROCHECK[45] and ERRAT[46] tool from SAVES 6.0 server and ProSA-web[47] analysis program. Later, the Maestro 11.8 tool of Schrödinger suite[48] was used to compare the WT protein structures with the mutants computing the root mean square deviation (RMSD). Higher RMSD indicates greater variation between WT and mutant structures[49]. All the structures were visualized by Pymol and Maestro 11.8. Further, the Stride[35] web-server was employed to view the mutation location in secondary structures while the ProtParam[36] server was used for the analysis physicochemical properties of WT and mutant proteins.

Molecular docking analysis

We used Molegro Virtual Docker[37] (MVD) for molecular docking study to view the binding affinity of a small molecule stabilizer with the mutant domains. The software is unified with high potential Piece Wise Linear Potential (PLP) and MVD scoring function. A carbazole derivative, PK083 (1-(9-ethylcarbazol-3-yl)-N-methylmethanamine) was used as ligand. Carbazole based small-molecules were tested to act as stabilizers in restoring the function of several mutants of p53 DBD[50,51]. So, we tested the stabilizing capability of PK083 to the highly damaging mutants of our study. All the mutant proteins were subjected to energy minimization using SwissPdb viewer[52] and the ligand structure was optimized with MMFF94 force field using steepest descent algorithm prior to docking. The binding site was defined covering the mutant residue to assess the binding affinity of PK083 to each mutation region. The docking processes were composed of maximum iteration of 1,500, maximum population size of 50 and Grid solution of 0.3. Further, we carried out post docking processes by hydrogen bonds optimization and energy minimization, simplex evolution at max steps 300 and neighbor distance technical setting fast at 1.00. The energy of the receptor-ligand complexes was minimized using Nelder Mead Simplex Minimization. Later on, the interactions of mutants with ligand were visualized in Discovery Studio 4.1.

Molecular dynamics (MD) simulations and MM/PBSA analysis

Molecular dynamics simulation of the WT and mutant protein structures was performed using GROMACS 5.1.4 version[38] and Linux 5.4 package. The GROMOS96 54a7[53] forcefield was selected as the force field for proteins and the ligand topologies were generated from the Automated Topology Builder version 3.0[54] (ATB) server. The proteins and mutants-ligand complexes were solvated using simple point charge (SPC) water molecules in a rectangular box where every structure was placed in the center at least 1.0 nm from the box edges. Required number of Na+ and Cl− ions were added to make the simulation system electrically neutral. The salt concentrations were set to 0.15 mol/L in all the systems. The solvated

systems were subjected to energy minimization for 5000 steps using the steepest descent method. Afterwards, three steps were conducted in the MD simulation: NVT (constant number of particles, volume, and temperature) series, NPT (constant number of particles, pressure, and temperature) series, and the production run. The NVT and the NPT series were conducted at a 300 K temperature and 1 atm pressure for the duration of 100 ps. V-rescale and Parrinello-Rahman were selected as the thermostat and barostat respectively of the performed simulation. Finally, the production run of nine proteins (WT and mutants) and six protein-ligand complexes were performed at 300 K for a duration of 100 ns (nanoseconds) in a supercomputing system provided by the Bioinformatics Division of National Institute of Biotechnology (NIB), Bangladesh. Thereafter, a comparative analysis was performed between WT and mutants measuring root mean square deviation (RMSD), root mean square fluctuation (RMSF), radius of gyration (Rg), solvent accessible surface area (SASA) and hydrogen bonds. Qtgrace program was used to represent all these analyses in the form of plots[55]. Further, the g_mmpbsa[56] package of GROMACS was used to calculate the MM/PBSA (Molecular Mechanics/Poisson Boltzmann Surface Area) binding free energies followed by final MD run to get a more detailed overview of the biomolecular interactions between the mutated proteins and ligand. The free solvation energy (polar and nonpolar solvation energies) and potential energy (electrostatic and Van der Waals interactions) of each protein-ligand complex were analyzed to determine the total ΔG_{bind}. The binding energies were calculated using the following equation in this method:

$$\Delta G_{binding} = G_{complex} - (G_{protein} + G_{ligand})$$

Here, the $\Delta G_{binding}$ = the total binding energy of the protein-ligand complex, $G_{protein}$ = the binding energy of free protein, and G_{ligand} = the binding energy of unbounded ligand.

Conclusions

BRCA1, ATM and TP53 protein plays a significant role as tumor suppressor in several cancer types. The structural conformation of the functional domains is very crucial for exerting their functional role. The present study showed that some plausible genetic variants could destabilize the protein structure and ultimately might play a critical role in altering the biological function of the respective gene. Hence, the explored mutations from our study might offer a great interest to cancer research as well as a methodological approach for further possible structural aberrant genetic variants findings.

Declarations

Author Contribution

Conceptualization and Methodology was performed by MNR, NF, THE, MSI, MUH; Formal analysis and Data curation was performed by THE, MNR, MUH, and AKMM; Writing-original draft prepared by MUH, NF, THE, MNR and MSI; Writing-review and editing performed by MUH, AKMM and MSI; Supervised by MUH, MSI. All authors have read and agreed to submit the final version of the manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data sets

All data generated and analyzed during this study are included in this article.

References

1. Calderón-Garcidueñas, A. L., Ruiz-Flores, P., Cerda-Flores, R. M. & Barrera-Saldaña, H. A. Clinical follow up of Mexican women with early onset of breast cancer and mutations in the BRCA1 and BRCA2 genes. *Salud Publica Mex.***47**, 110–115 (2005).
2. Ahmed, M. & Rahman, N. ATM and breast cancer susceptibility. *Oncogene* vol. 25 5906–5911 (2006).
3. Olivier, M., Hollstein, M. & Hainaut, P. TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology* vol. 2 (2010).
4. Bonadona, V. *et al.* Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in lynch syndrome. *JAMA - J. Am. Med. Assoc.***305**, 2304–2310 (2011).
5. Hendriks, Y. M. C. *et al.* Cancer risk in hereditary nonpolyposis colorectal cancer due to MSH6 mutations: Impact on counseling and surveillance. *Gastroenterology***127**, 17–25 (2004).
6. Foulkes, W. D., Flanders, T. Y., Pollock, P. M. & Hayward, N. K. The CDKN2A (p16) gene and human cancer. *Molecular Medicine* vol. 3 5–20 (1997).
7. Hofstatter, E. W. *et al.* PALB2 mutations in familial breast and pancreatic cancer. *Fam. Cancer***10**, 225–231 (2011).
8. Martin, S. E. *et al.* BRCA1 E1644X: A deleterious mutation in an African American individual with early onset breast cancer. *Breast Cancer Res. Treat.***113**, 393–395 (2009).
9. Futreal, P. A. *et al.* BRCA1 mutations in primary breast and ovarian carcinomas. *Science (80-.).***266**, 120–122 (1994).
10. Malone, K. E. *et al.* Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in White and Black American women ages 35 to 64 years. *Cancer Res.***66**, 8297–8308 (2006).
11. Rozman, V. & Kunej, T. Harnessing Omics Big Data in Nine Vertebrate Species by Genome-Wide Prioritization of Sequence Variants with the Highest Predicted Deleterious Effect on Protein Function. *Omi. A J. Integr. Biol.***22**, 410–421 (2018).
12. Krawczak, M. *et al.* Human Gene Mutation Database - A biomedical information and research resource. *Hum. Mutat.***15**, 45–51 (2000).

13. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.***8**, 1229–1231 (1998).
14. Ng, P. C. & Henikoff, S. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.***12**, 436–446 (2002).
15. Ronaghi, M. & Langaee, T. Single nucleotide polymorphisms: discovery, detection and analysis. *Per. Med.***2**, 111–125 (2005).
16. Kwok, P.-Y. *Single Nucleotide Polymorphisms*. vol. 212 (Humana Press, 2002).
17. Capriotti, E. & Altman, R. B. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics***12**, (2011).
18. Basu, M. K., Poliakov, E. & Rogozin, I. B. Domain mobility in proteins: Functional and evolutionary implications. *Brief. Bioinform.***10**, 205–216 (2009).
19. Alshatwi, A. A., Hasan, T. N., Syed, N. A., Shafi, G. & Grace, B. L. Identification of Functional SNPs in BARD1 Gene and In Silico Analysis of Damaging SNPs: Based on Data Procured from dbSNP Database. *PLoS One***7**, e43939 (2012).
20. Chandrasekaran, G. *et al.* In silico analysis of the deleterious nsSNPs (missense) in the homeobox domain of human HOXB13 gene responsible for hereditary prostate cancer. *Chem. Biol. Drug Des.***90**, 188–199 (2017).
21. Hossain, M. S., Roy, A. S. & Islam, M. S. In silico analysis predicting effects of deleterious SNPs of human RASSF5 gene on its structure and functions. *Sci. Rep.***10**, 14542 (2020).
22. Deng, N., Zhou, H., Fan, H. & Yuan, Y. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget* vol. 8 110635–110649 (2017).
23. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics***14**, (2013).
24. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.***31**, 3812–3814 (2003).
25. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.***Chapter 7**, (2013).
26. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics***31**, 2745–2747 (2015).
27. 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***22**, 2729–2734 (2006).
28. Thomas, P. D. *et al.* Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res.***34**, W645–W650 (2006).
29. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.***30**, 1237–1244 (2009).

30. Bromberg, Y. & Rost, B. SNAP: Predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.***35**, 3823–3835 (2007).
31. Bendl, J. *et al.* PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. *PLoS Comput. Biol.***10**, e1003440 (2014).
32. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.***33**, (2005).
33. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.***44**, W344–W350 (2016).
34. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. *Curr. Protoc. Bioinforma.***15**, (2006).
35. Frishman, D. & Argos, P. Knowledge-based protein secondary structure assignment. *Proteins Struct. Funct. Bioinforma.***23**, 566–579 (1995).
36. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* 571–607 (Humana Press, 2005). doi:10.1385/1-59259-890-0:571.
37. Bitencourt-Ferreira, G. & de Azevedo, W. F. Molegro virtual docker for docking. in *Methods in Molecular Biology* vol. 2053 149–167 (Humana Press Inc., 2019).
38. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX***1–2**, 19–25 (2015).
39. Bhagwat, M. Searching NCBI's dbSNP database. *Curr. Protoc. Bioinforma.***Chapter 1**, (2010).
40. Liu, Y. & Tozeren, A. Domain Altering SNPs in the Human Proteome and Their Impact on Signaling Pathways. *PLoS One***5**, e12890 (2010).
41. Yang, M., Derbyshire, M. K., Yamashita, R. A. & Marchler-Bauer, A. NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. *Curr. Protoc. Bioinforma.***69**, (2020).
42. Zhang, M., Huang, C., Wang, Z., Lv, H. & Li, X. In silico analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in the human GJA3 gene associated with congenital cataract. *BMC Mol. Cell Biol.***21**, 12 (2020).
43. Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K. & Sarai, A. ProTherm, version 4.0: Thermodynamic database for proteins and mutants. *Nucleic Acids Res.***32**, (2004).
44. Heo, L., Park, H. & Seok, C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.***41**, (2013).
45. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.***26**, 283–291 (1993).
46. Colovos, C. & Yeates, T. O. Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Sci.***2**, 1511–1519 (1993).
47. Wiederstein, M. & Sippl, M. J. ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.***35**, 407–410 (2007).

48. Dagan-Wiener, A. *et al.* Bitter or not? BitterPredict, a tool for predicting taste from chemical structure. *Sci. Rep.***7**, (2017).
49. Kuzmanic, A. & Zagrovic, B. Determination of ensemble-average pairwise root mean-square deviation from experimental B-factors. *Biophys. J.***98**, 861–871 (2010).
50. Bauer, M. R. *et al.* Targeting Cavity-Creating p53 Cancer Mutations with Small-Molecule Stabilizers: The Y220X Paradigm. *ACS Chem. Biol.***15**, 657–668 (2020).
51. Raghavan, V., Agrahari, M. & Gowda, D. K. Virtual screening of p53 mutants reveals Y220S as an additional rescue drug target for PhiKan083 with higher binding characteristics. *Comput. Biol. Chem.***80**, 398–408 (2019).
52. Guex, N. & Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis***18**, 2714–2723 (1997).
53. Wei, G. & Baker, N. Differential Geometry-Based Solvation and Electrolyte Transport Models for Biomolecular Modeling: A Review. in *Many-Body Effects and Electrostatics in Biomolecules* 417–461 (Pan Stanford, 2016). doi:10.1201/b21343-15.
54. Stroet, M. *et al.* Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane. *J. Chem. Theory Comput.***14**, 5834–5845 (2018).
55. Mohammad, T. *et al.* Virtual screening approach to identify high-affinity inhibitors of serum and glucocorticoid-regulated kinase 1 among bioactive natural products: Combined molecular docking and simulation studies. *Molecules***25**, (2020).
56. Kumari, R., Kumar, R. & Lynn, A. G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. *J. Chem. Inf. Model.***54**, 1951–1962 (2014).
57. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science (80-.).***266**, 66–71 (1994).
58. Nissenkorn, A. & Ben-Zeev, B. Ataxia telangiectasia. in *Handbook of Clinical Neurology* vol. 132 199–214 (Elsevier B.V., 2015).
59. Lavin, M. F. *et al.* Functional consequences of sequence alterations in the ATM gene. *DNA Repair* vol. 3 1197–1205 (2004).
60. Seidel, J. J., Anderson, C. M. & Blackburn, E. H. A Novel Tel1/ATM N-Terminal Motif, TAN, Is Essential for Telomere Length Maintenance and a DNA Damage Response. *Mol. Cell. Biol.***28**, 5736–5746 (2008).
61. Tokino, T. Dual role of p53 in DNA binding. *Cancer Biol. Ther.***3**, 1322–1323 (2004).
62. Joerger, A. C. & Fersht, A. R. The Tumor Suppressor p53: From Structures to Drug Discovery. *Cold Spring Harb. Perspect. Biol.***2**, (2010).
63. Raj, N. & Attardi, L. D. The transactivation domains of the p53 protein. *Cold Spring Harb. Perspect. Med.***7**, a026047 (2017).
64. Miller, M. P. & Kumar, S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.***10**, 2319–2328 (2001).

65. Doniger, S. W. *et al.* A Catalog of Neutral and Deleterious Polymorphism in Yeast. *PLoS Genet.***4**, e1000183 (2008).
66. Du, K., Sharma, M. & Lukacs, G. L. The ΔF508 cystic fibrosis mutation impairs domain-domain interactions and arrests post-translational folding of CFTR. *Nat. Struct. Mol. Biol.***12**, 17–25 (2005).
67. Mayer, S., Rüdiger, S., Ang, H. C., Joerger, A. C. & Fersht, A. R. Correlation of Levels of Folded Recombinant p53 in Escherichia coli with Thermodynamic Stability in Vitro. *J. Mol. Biol.***372**, 268–276 (2007).
68. Singh, S. M., Kongari, N., Cabello-Villegas, J. & Mallela, K. M. G. Missense mutations in dystrophin that trigger muscular dystrophy decrease protein stability and lead to cross-β aggregates. *Proc. Natl. Acad. Sci. U. S. A.***107**, 15069–15074 (2010).
69. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.***33**, (2005).
70. Venkata Subbiah, H., Ramesh Babu, P. & Subbiah, U. In silico analysis of non-synonymous single nucleotide polymorphisms of human DEFB1 gene. *Egypt. J. Med. Hum. Genet.***21**, 66 (2020).
71. Chen, X. & Sullivan, P. F. Single nucleotide polymorphism genotyping: Biochemistry, protocol, cost and throughput. *Pharmacogenomics Journal* vol. 3 77–96 (2003).
72. Rajasekaran, R., Sudandiradoss, C., Doss, C. G. P. & Sethumadhavan, R. Identification and in silico analysis of functional SNPs of the BRCA1 gene. *Genomics***90**, 447–452 (2007).
73. Doss, C. G. P. & Rajith, B. Computational refinement of functional single nucleotide polymorphisms associated with ATM gene. *PLoS One***7**, (2012).
74. Kucukkal, T. G., Petukh, M., Li, L. & Alexov, E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Current Opinion in Structural Biology* vol. 32 18–24

Tables

Table 1. The p-value of genes computed from the fisher's exact test.

Gene term	P-value	Adjusted P-value	Odds Ratio
BRCA1	1.45E-15	5.58E-13	82.30452675
ATM	6.67E-13	1.28E-10	70.07007007
TP53	1.91E-08	1.84E-06	26.56042497

Table 2. I-MUTANT 2.0 predictions for nsSNPs with lower DDG values in BRCA1, ATM and TP53.

Gene	nsSNP ID	Amino Acid Change	Stability	DDG
BRCA1	rs45553935	V1736G	Decrease	-2.7
	rs80357107	V1838G	Decrease	-2.37
	rs80357451	V1832G	Decrease	-2.32
	rs1555579627	V1687G	Decrease	-2.3
	rs730881496	L1729Q	Decrease	-2.25
ATM	rs201216427	V2757G	Decrease	-2.81
	rs587779873	I2865T	Decrease	-2.47
	rs876659516	I2948T	Decrease	-2.29
	rs1591264625	V2830G	Decrease	-2.12
	rs730881328	V2906A	Decrease	-2.1
TP53	rs1057520004	V216G	Decrease	-2.56
	rs1057519998	L194H	Decrease	-2.51
	rs1330865474	I254S	Decrease	-2.48
	rs730882027	I251T	Decrease	-2.34
	rs760043106	I195T	Decrease	-2.28

Table 3. Target sites disrupted and created by single nucleotide polymorphisms (SNPs) in miRNA seeds.

Target sites disrupted by SNPs and INDELs in miRNA seeds					
Gene	**SNP ID**	**miR ID**	**miRSite**	**Conservation**	**context+ score change**
ATM	rs3745198	hsa-miR-6796-3p	AGAGCUU	5	-0.025
TP53	rs71309450	hsa-miR-1233-5p	UCCCACA	5	-0.192
Target sites created by SNPs and INDELs in miRNA seeds					
ATM	rs190453265	hsa-miR-525-3p	GCACCUUA	7	-0.438
TP53	rs201549145	hsa-miR-548i	UUAUUUUA	11	0.014

Table 4. Predicted RMSD (All atoms, Cα atoms and Backbone) values of WT and mutants of BRCA1, ATM and TP53.

Gene	SNP ID	Amino Acid Substitutions	RMSD (All atoms)	RMSD (Cα atoms)	RMSD (Backbone)
BRCA1	rs45553935	V1736G	3.3315	0.0019	0.0012
	rs80357107	V1838G	1.4924	0.0014	0.0015
	rs80357451	V1832G	1.6726	0.0005	0.0015
	rs1555579627	V1687G	3.9798	0.0017	0.0016
	rs730881496	L1729Q	2.0206	0.0018	0.0015
ATM	rs201216427	V2757G	2.6360	0.0015	2.0706
	rs587779873	I2865T	3.7960	0.0011	1.9320
	rs876659516	I2948T	3.1163	0.0012	2.0694
	rs1591264625	V2830G	2.7638	0.0020	2.0559
	rs1555053927	V2906A	3.1687	0.0010	1.9769
TP53	rs1057520004	V216G	2.6572	0.0020	0.0017
	rs1057519998	L194H	2.5904	0.0022	0.0019
	rs1330865474	I254S	1.6255	0.0023	0.0014
	rs730882027	I251T	1.2445	0.0018	0.0018
	rs760043106	I195T	1.8478	0.0014	0.0017

Table 5. Interactions of PK083 with six mutant proteins of BRCA1 (BRCT domain), ATM (Catalytic domain) and p53 (DNA-Binding domain).

Protein	Mutations	Binding Score (kcal/mol)	Interacting bond with mutant residue	Distance (Å)
BRCA1	V1687G	-68.2	Pi-Sigma	2.02
	V1736A	-65.7	Pi-Sigma	2.09
ATM	I183T	-72.3	Hydrogen	1.75
	V224A	-66.3	Pi-Alkyl	4.44
TP53	L194H	-32.2	Pi-Alkyl	5.36
	V216G	-32.8	Pi-Lone pair	2.83

Table 6. The average values of RMSD, RMSF, Rg and SASA for WT proteins and protein-drug complexes.

Protein/Complex	RMSD (nm)	RMSF (nm)	Rg (nm)	SASA (nm^2)
WT-BRCT	~0.10	~0.34	~1.94	~116
V1687G	~0.11	~0.47	~1.94	~119
V1736G	~0.14	~0.45	~1.96	~114
WT-Catalytic domain	~0.45	~0.50	~1.96	~160
I2865T	~0.33	~0.40	~2.02	~164
V2906A	~0.45	~0.68	~2.07	~168
WT-DBD	~0.12	~0.43	~1.70	~115
V216G	~0.16	~0.62	~1.70	~112
L194H	~0.16	~0.51	~1.71	~114

Table 7: MM/PBSA calculations of binding free energy for six mutant-PK083 complexes.

Complexes	Van der Waal energy (KJ mol^{-1})	Electrostatic energy (KJ mol^{-1})	Polar solvation energy (KJ mol^{-1})	SASA energy (KJ mol^{-1})	Binding energy (KJ mol^{-1})
V1687G-PK083	-162.942 +/- 8.328	-4.006 +/- 4.198	72.501 +/- 10.932	-15.973 +/- 0.812	-110.419 +/- 11.896
V1736G-PK083	-176.793 +/- 9.540	-29.386 +/- 4.914	111.041 +/- 11.923	-15.701 +/- 0.803	-110.839 +/- 12.909
V216G-PK083	-163.826 +/- 9.659	-14.571 +/- 4.866	81.793 +/- 11.348	-16.607 +/- 0.861	-113.211 +/- 10.362
L194H-PK083	-168.406 +/- 8.744	-29.885 +/- 8.523	121.305 +/- 13.577	-15.397 +/- 0.843	-92.383 +/- 11.060
I2865T-PK083	-128.608 +/- 9.752	-11.997 +/- 5.278	55.265 +/- 7.757	-14.316 +/- 1.079	-99.656 +/- 9.611
V2906A-PK083	-86.402 +/- 9.708	-27.661 +/- 11.869	80.415 +/- 20.936	-9.403 +/- 1.008	-43.051 +/- 12.340

Figures

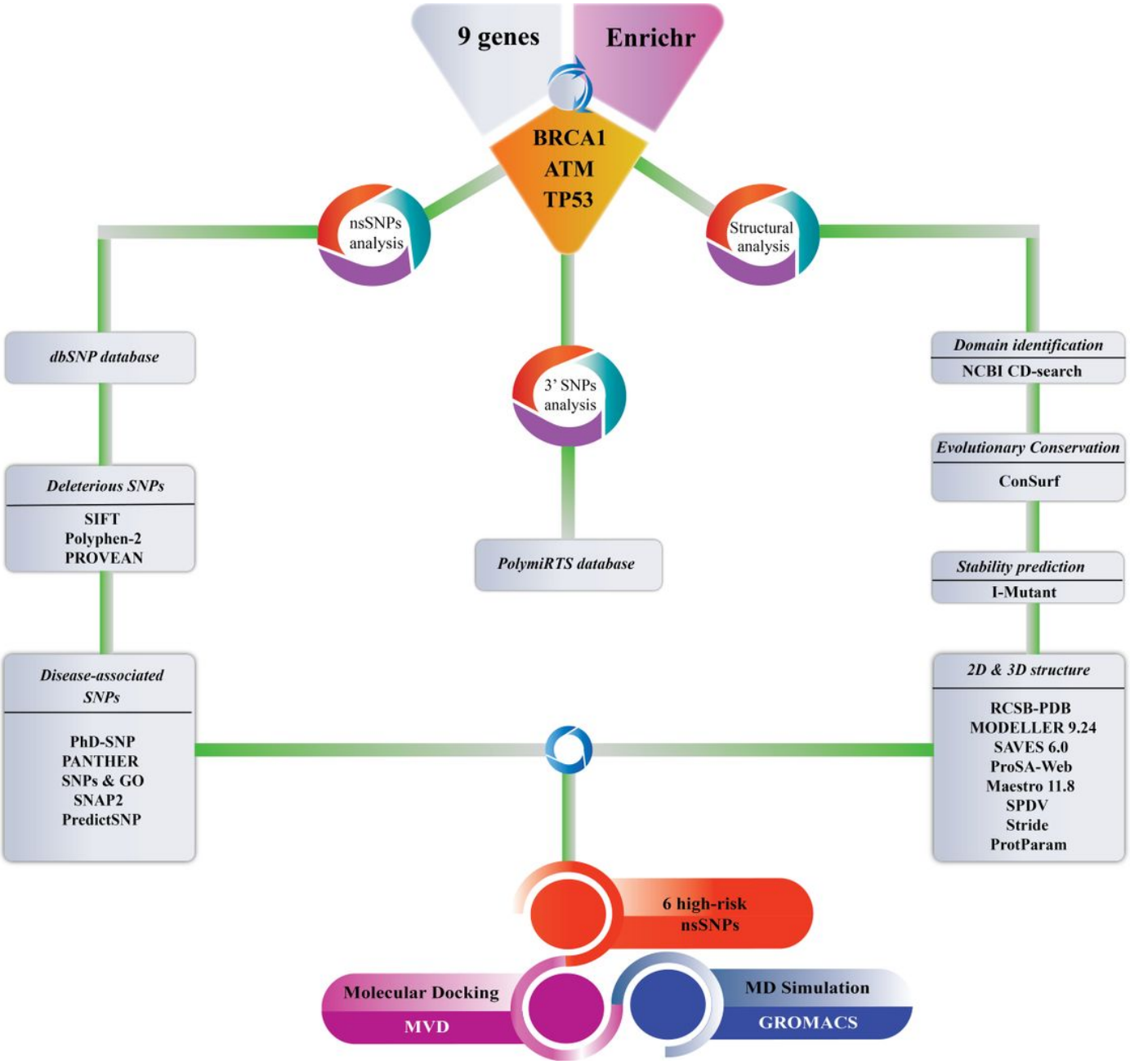

Figure 1

Schematic workflow of identifying the deleterious SNPs of cancer susceptibility genes.

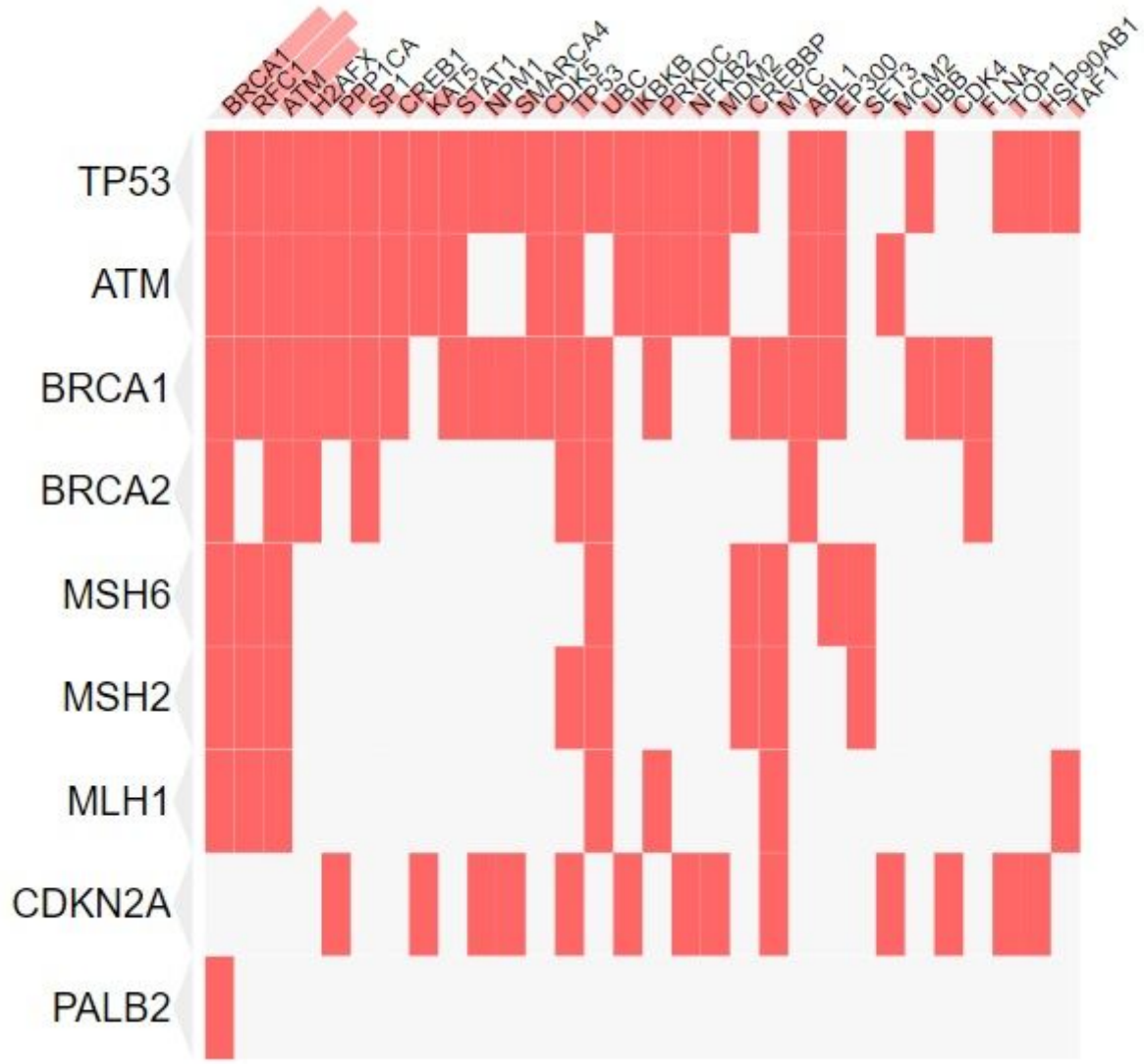

Figure 2

Gene enrichment result from the Enrichr server. From the curated PPI interactions, BRCA1, ATM and TP53 is more enriched in connectivity on a clustegram view.

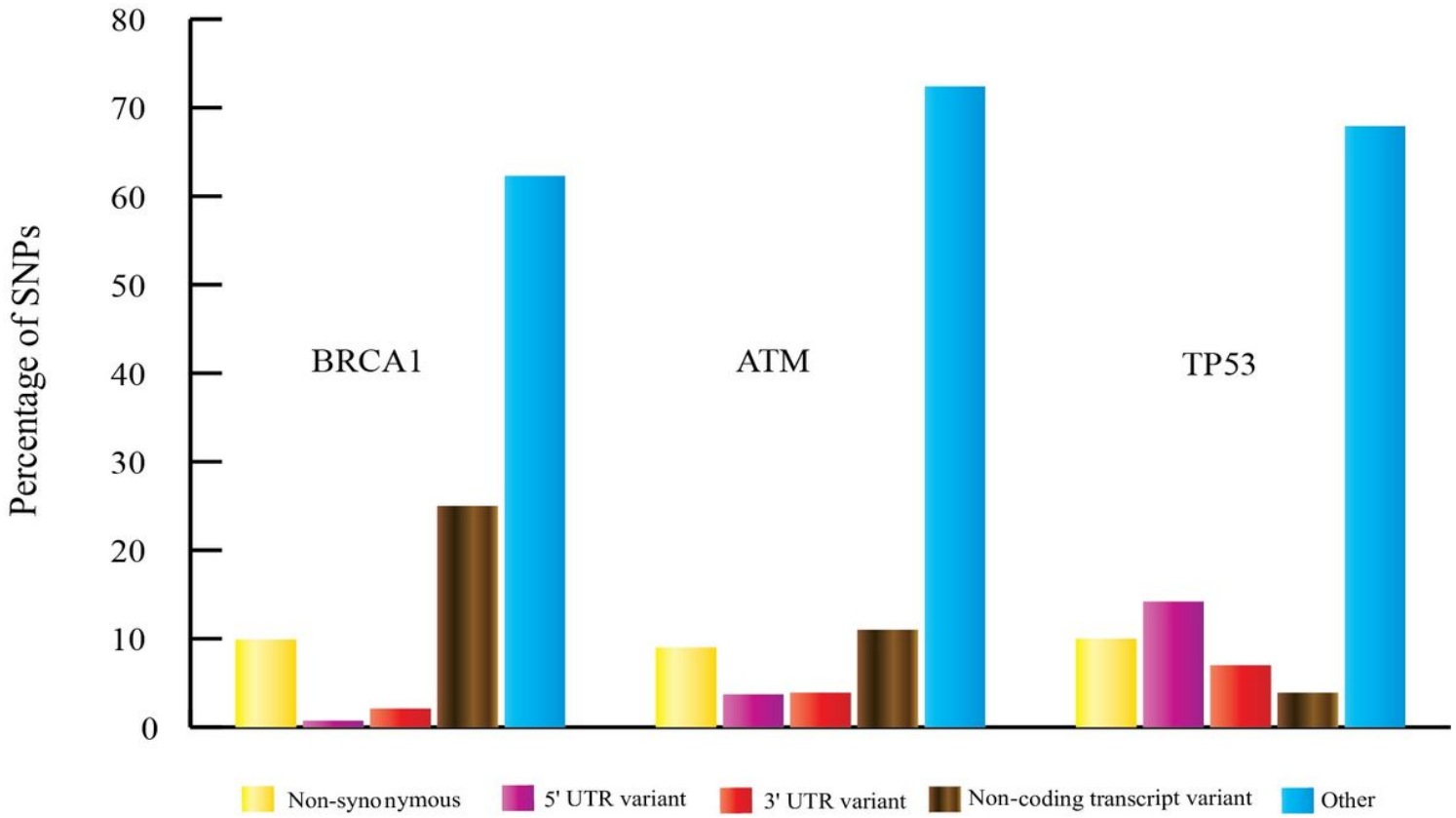

Figure 3

Bar diagram showing the percentages of the SNPs in BRCA1, ATM and TP53 genes. BRCA1 gene had 9.9% nsSNPs, 0.72% 5'UTR SNPs, 2.1% 3'UTR SNPs, 9% non-coding transcript variants and 62.28% other SNPs of total SNPs; ATM gene had 9% nsSNPs, 3.7% 5'UTR SNPs, 3.9% 3'UTR SNPs, 11% non-coding transcript variants and 72.4% other SNPs of total SNPs; TP53 gene had 10% nsSNPs, 14.2% 5'UTR SNPs, 7% 3'UTR SNPs, 3.9% non-coding transcript variants and 67.9% other SNPs of total SNPs.

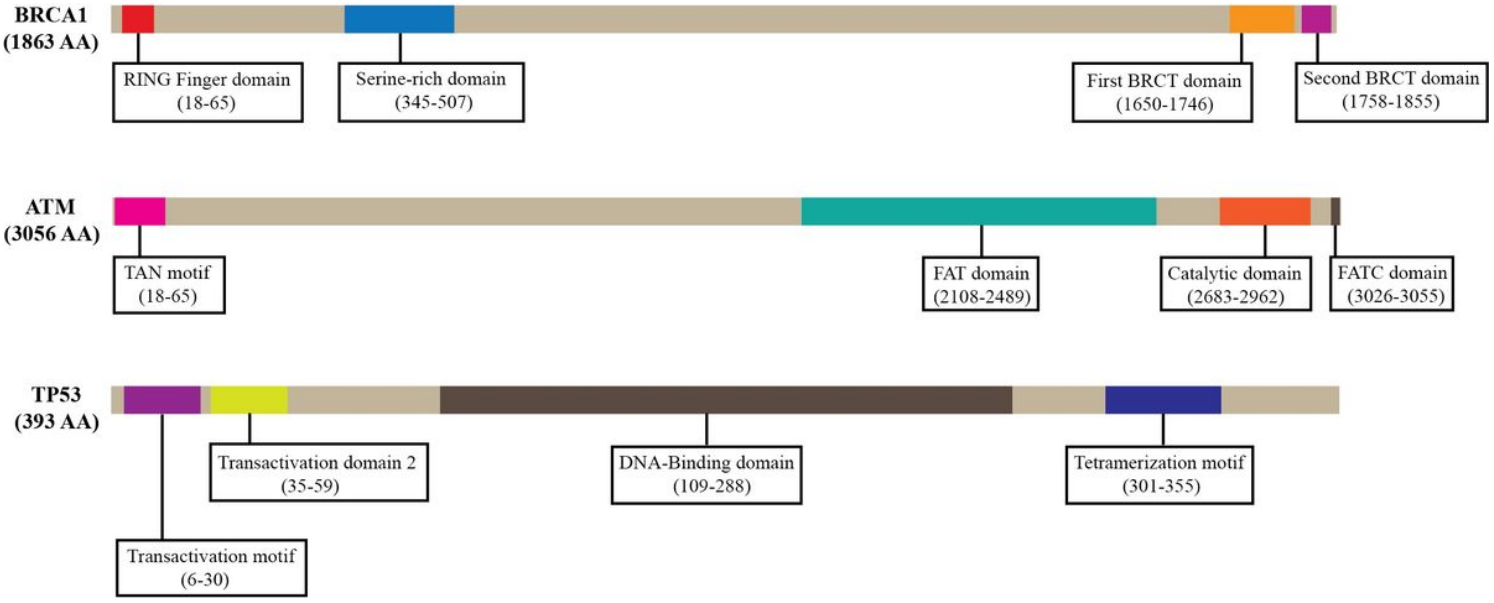

Figure 4

Domains and sequence intervals of BRCA1, ATM and TP53 proteins.

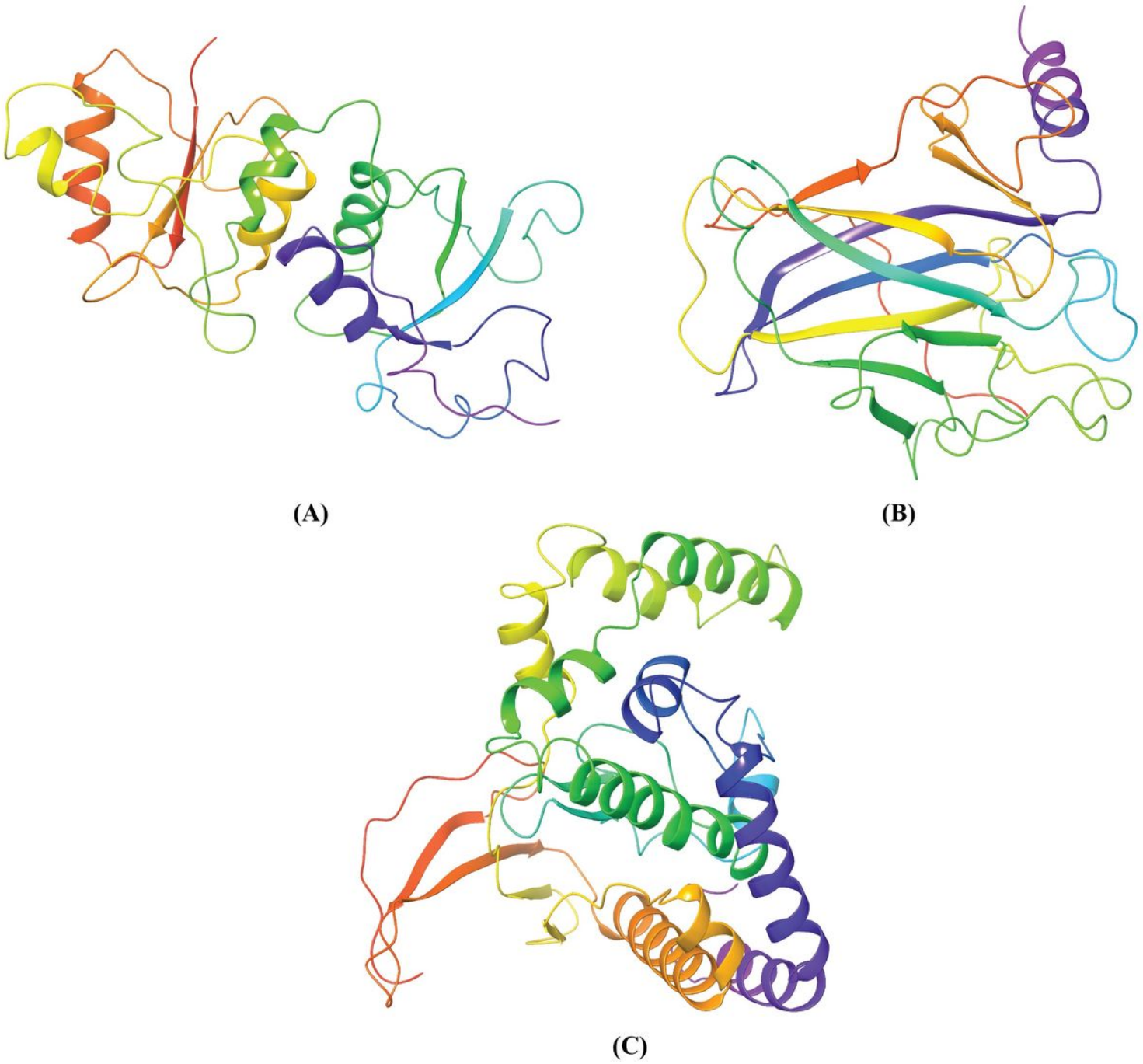

Figure 5

3D structures of three domains from BRCA1, ATM and p53 protein. (A) Crystal structure of BRCT domain of BRCA1 at 3.50 Å resolution. (B) DNA binding domain of p53 at 1.92 Å resolution. (C) Modelled 3D structure of the catalytic domain of ATM. We used 5NP0 as template to model the domain.

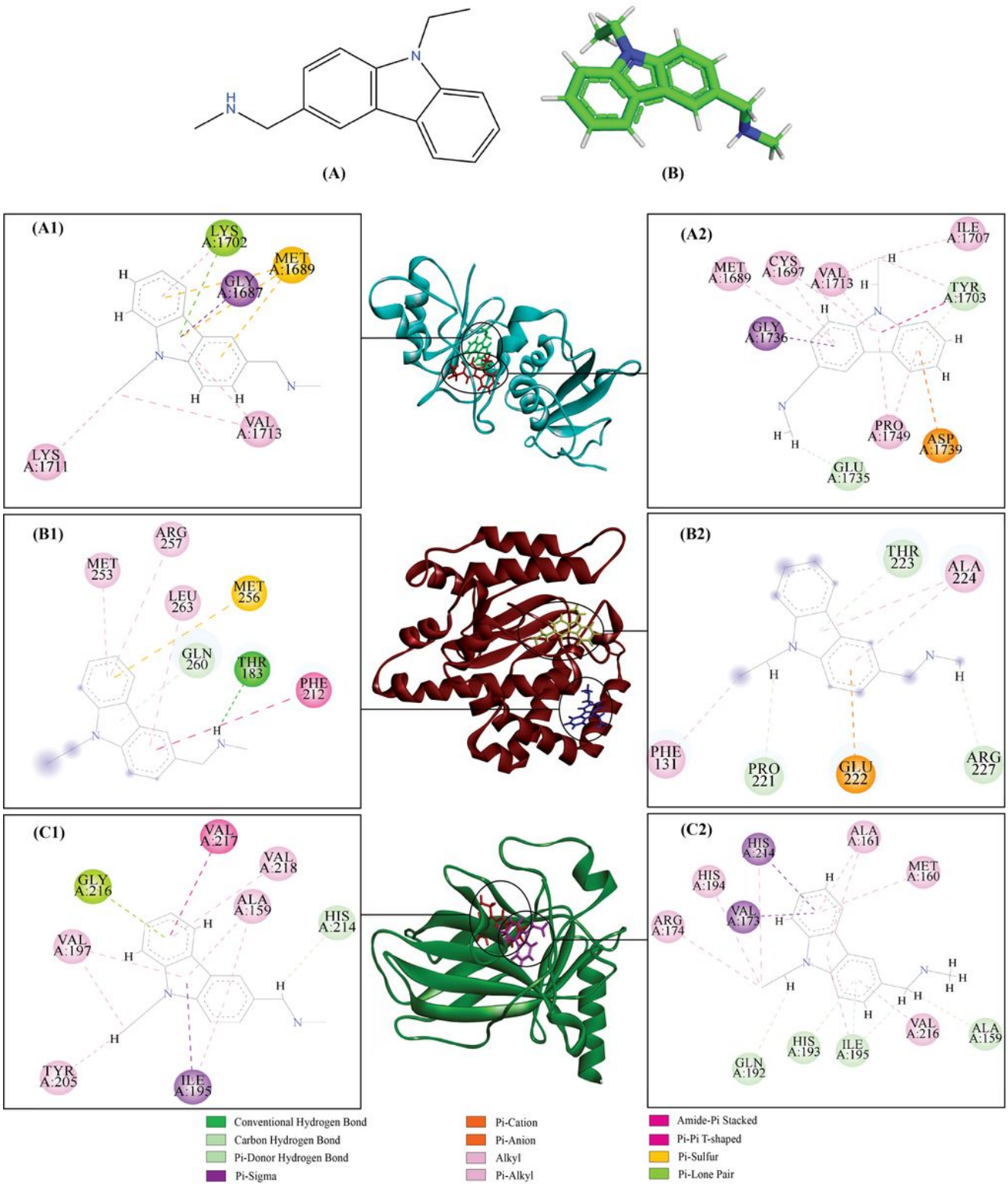

Figure 6

Structure of PhiKan083 and interactions with mutants. 2D (A) and energy minimized 3D (B) structure of PK083. Interaction of PK083 with (A1) V1687G and (A2) V1736G mutant BRCT domains, (B1) I183T (I2865T) and (B2) V224A (V2906A) mutant Catalytic domains, (C1) V216G and (C2) L194H mutant DNA-binding domains.

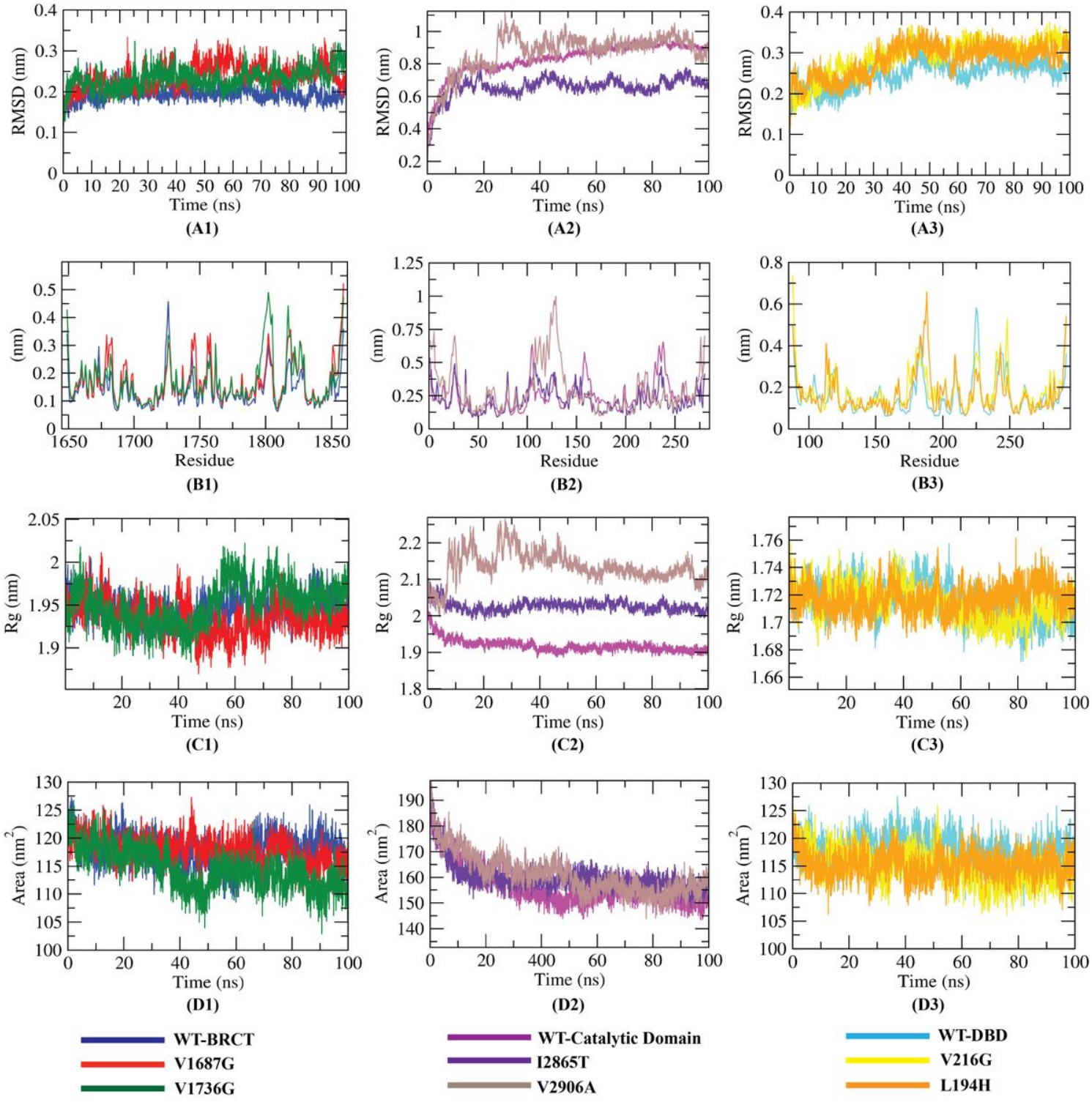

Figure 7

MD simulation results of WT and mutant variants. (A1-A3) RMSD analysis, (B1-B3) RMSF analysis, (C1-C3) radius of gyration (Rg) analysis, and (D1-D3) SASA analysis.

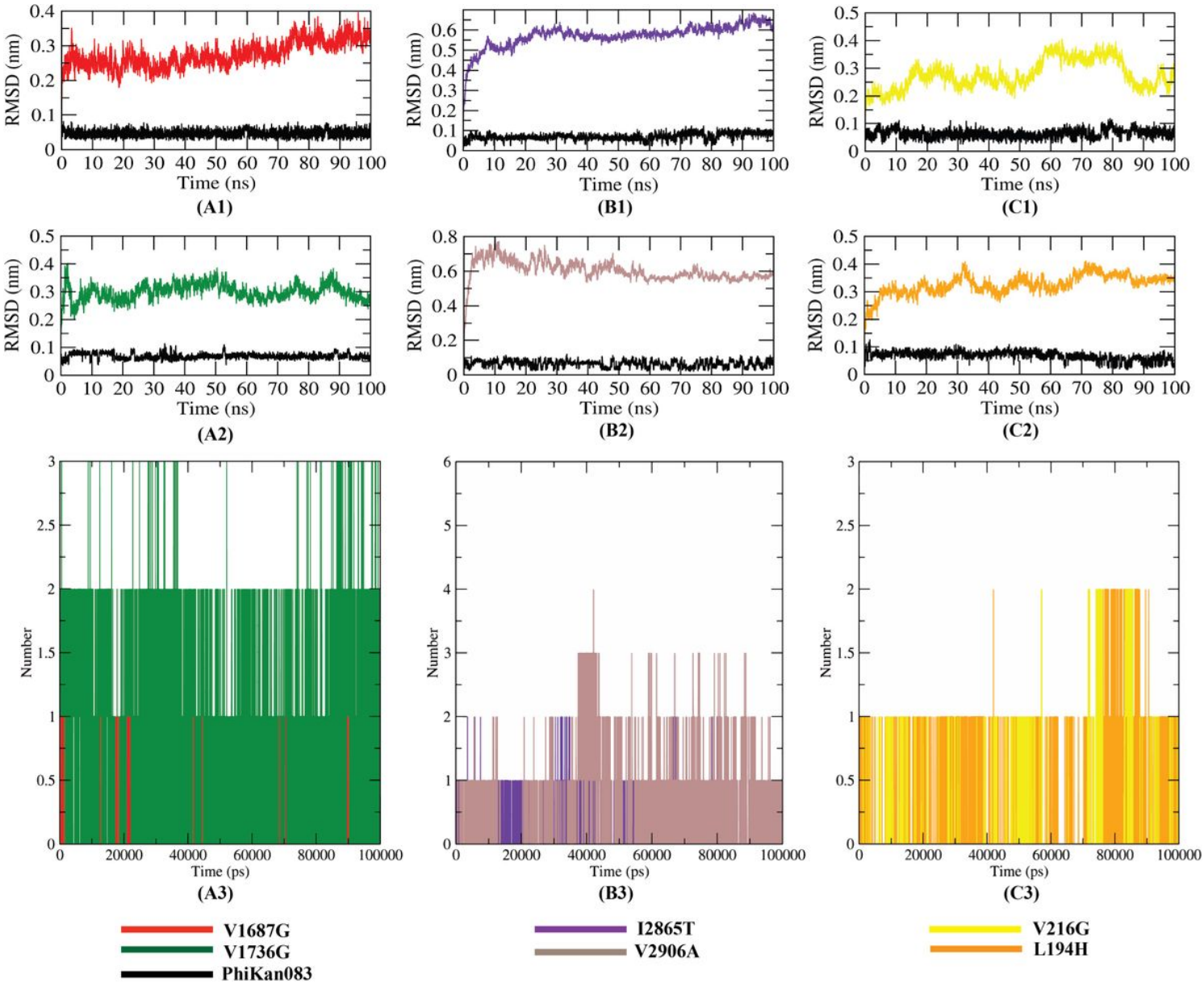

Figure 8

MD simulation results of mutant-PK083 complexes. (A1-C2) RMSD of PK083 and protein backbone obtained from 100 ns MD simulation. (A3-C3) Hydrogen bond analysis between PK083 and the mutants.

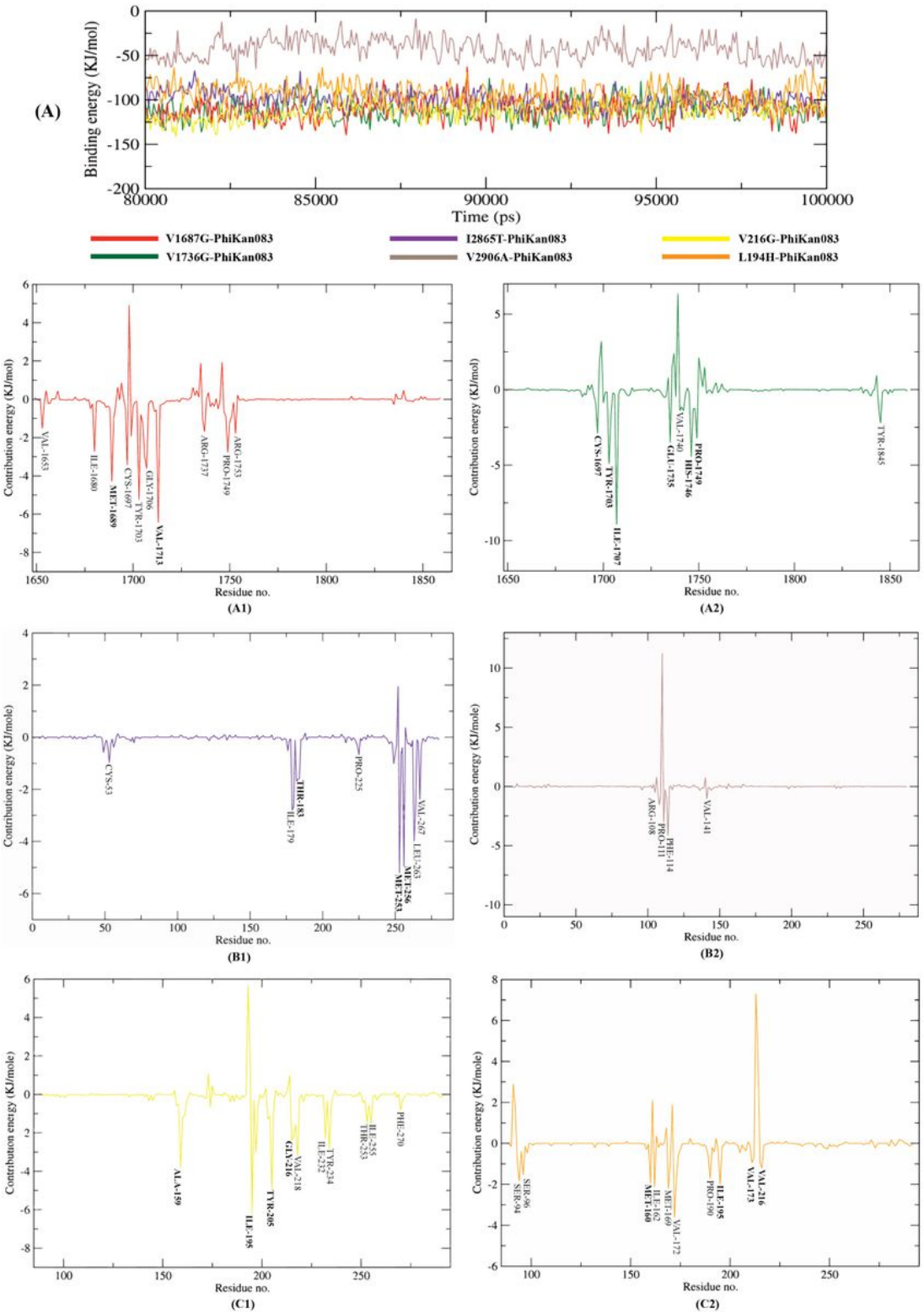

Figure 9

Graphical representation showing contribution energy plot of each residue of the six mutants, (A1) V1687G, (A2) V1736G, (B1) I183T (I2865T), (B2) V224A (V2906A), (C1) V216G and (C2) L194H. The residues that were involved in molecular docking are shown in bold.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTableS1S8.docx
- SupplementaryfigureS5.tif
- SupplementaryfigureS6.tif
- Supplementaryfile1.xlsx
- Supplementaryfile10.pdf
- Supplementaryfile11.pdf
- Supplementaryfile12.pdf
- Supplementaryfile13.xlsx
- Supplementaryfile14.xlsx
- Supplementaryfile15.xlsx
- Supplementaryfile2.xlsx
- Supplementaryfile3.xlsx
- Supplementaryfile4.xlsx
- Supplementaryfile5.xlsx
- Supplementaryfile6.xlsx
- Supplementaryfile7.xlsx
- Supplementaryfile8.xlsx
- Supplementaryfile9.xlsx