Original Article

An *In silico* Method to Study Structure, Function, and Regulatory Role Alteration Mediated by Single-nucleotide Polymorphisms in Gallbladder Cancer

Arpit Kumar Pradhan[1,†], Rudrarup Bose[1], Shyamasree Ghosh[1,†], Amitava Datta[2]

[1]School of Biological Sciences, National Institute of Science Education and Research, HBNI, Bhubaneswar, Jatni, Khurda, Odisha, India, [2]School of Computer Science and Software Engineering, The University of Western Australia, Perth, WA 6009, Australia

[†]Both authors contributed equally to this work.

Abstract

Introduction: Gallbladder cancer (GBC) is a fatal malignancy of gallbladder and bile ductswhich shows delayed symptoms and sometimes can be asymptomatic, being fatal. Reported globally, for a very low survival rate, it suffers from the lack of early diagnostic and prognostic markers. Single nucleotide polymorphisms (SNPs) have been reported to be associated in different cancers. **Methods:** In this study using in silico methods, we report for the first time a combination of SNPs from the coding and noncoding region leading to alteration in GBC. Different pipelines were designed for the study of SNPs. Regulatory role alteration of Synonymous and non-coding SNPs were studied using RegulomeDB, DeepSEA analysis and funcPred. Structural alteration and energy parameters for non-synonymous SNPs were studied by Swiss-PDB, Chimera and Gromacs. Protein stability analysis was done using MutPred, mCSM and I-mutant. **Results:** As a result, three potential variants from the coding region rs1042838, rs11887534, and rs700519 associated with progesterone receptor, ATP binding cassette subfamily G member 8 (ABCG8), and cytochrome P450 19A1, respectively, were predicted to be potentially damaging SNPs in GBC leading to structure and function alteration. Three noncoding SNPs (rs2978974, rs4633 and rs2830) and 1 missense SNP(rs523349) were shown to be associated with damaging effect in GBC, and one of these SNPs (rs2978974) showed significant chromatin feature alteration. **Conclusion:** Our study strongly shows that SNPs both in the coding and noncoding region may be exploited as a combination of potential biomarkers in early diagnosis of GBC due to structure function alteration by nonsynonymous SNPs and regulatory role alteration by noncoding SNPs.

Keywords: ABCG8, cancer, gallbladder cancer, single-nucleotide polymorphism

Introduction

Gallbladder carcinoma (GBC) is a fatal malignant adenocarcinoma arising from the epithelial lining of gallbladder and bile ducts, involving the chronic biliary tract. It is a disorder with a high mortality rate and is reported as one of the most aggressive biliary cancers, having the shortest median survival duration. With the location of the gallbladder behind the liver and symptoms common to other disorders such as nausea, jaundice, vomiting, stomach pain, abdominal lumps, or being completely asymptomatic at the initial stages, the diagnosis often becomes possible only at an advanced or late stage. This disease suffers from the limitation of suitable markers for early diagnosis and very low rates of survival. Females are with an increased risk. The global occurrence of GBC is reported higher from Korea, Japan, Slovakia, Poland, and the Czech Republic. It is also enlisted in the category of most common cause for cancer-related mortality from Northern and North-Eastern parts of India, South Karachi, Pakistan, and Quito, Ecuador. Globally, Chile has been reported with highest mortality rates with higher mortality rate in men (7.8/100,000) as compared to that of

Access this article online

Quick Response Code:

Website: www.jnsbm.org

DOI: 10.4103/jnsbm.JNSBM_233_17

Address for correspondence: Arpit Kumar Pradhan, School of Biological Sciences, National Institute of Science Education and Research, HBNI, Bhubaneswar, Bhimpur-Padanpur, Jatni, Khurda - 752 050, Odisha, India.
E-mail: arpit.pradhan@niser.ac.in

How to cite this article: Pradhan AK, Bose R, Ghosh S, Datta A. An *In silico* method to study structure, function, and regulatory role alteration mediated by single-nucleotide polymorphisms in gallbladder cancer. J Nat Sc Biol Med 2018;9:137-49.

women (16.6/100,000).[1,2] Thus, the current day research in gallbladder cancer is focused in the search of biomarkers for early diagnosis and prognosis.

Association of single-nucleotide polymorphisms (SNPs) with the risk in gallbladder cancer[3-9] and gallbladder stones[10-12] has been reported globally and also from isolated populations. What we do not know at this point is whether common susceptibility or risk factors are associated with the origin of both gallbladder stones and gallbladder cancer or whether one leads to the other.

Intraspecies variation is largely attributed to SNPs.[13] Broadly, SNPs are classified into synonymous (silent) SNP and nonsynonymous (missense) SNP (nsSNPs). Synonymous SNPs, arising due to wobble transfer RNA (tRNA) base pairing and redundancy in the genetic code, in coding regions, do not lead to a change in amino acid or primary polypeptide sequence. Yet, synonymous SNPs find importance as they alter the messenger RNA (mRNA) secondary structures and interfere with various processes of mRNA splicing, mRNA stability, protein translation, and co-translational protein folding, thus leading to changes in *cis* and *trans* factors that affect the mRNA stability which, in turn, may affect gene expression, both events being very closely linked.[13] nsSNPs, on the other hand, lead to changes in the amino acid sequence. Therefore, their study finds importance as they directly influence the translated primary polypeptide. Such changes in amino acid sequence are not only related to their alteration in the primary sequence but also may reduce protein solubility or destabilize the protein structure. We tried to understand through *in silico* approaches using computational tools for search of biomarkers and the structural-functional relation of SNP with respect to its coded protein in gallbladder cancer.

Materials and Methods

Identifying single-nucleotide polymorphisms involved in gallbladder carcinoma

The SNPs for GBC were obtained from PolySearch 2.0[14] and were subjected to various *in silico* analysis, and a pipeline was designed for selection of SNPs having structural and functional importance. SNPs involved in the GBC were enlisted through PolySearch 2.0,[15] with the query keyword "Gallbladder Cancer." A list of 41 SNPs was found to be related with GBC and had a greater global minor allele frequency (MAF) value [Table 1]. Location of these SNPs, either in the coding region or noncoding region, was found using Ensembl genome browser.[16,17] SNPs were sorted based on their location in coding or noncoding region. The dbSNP database, being the most extensive database, was availed for our study in spite of its limitations of containing both validated and nonvalidated polymorphisms.[18] The rsIDs of SNPs under study were entered in the *Ensembl* and their precise location was obtained. We selected missense nsSNPs, synonymous SNPs, and noncoding SNPs for our investigation. Separate methods were designed to study the functionality of nsSNPs [Figure 1].

Functional analysis of nonsynonymous single-nucleotide polymorphisms

nsSNPs located in the coding region result in amino acid variations. PolyPhen-2 web server[19] and SIFT were used to predict any damaging effect of nsSNPs on structure and

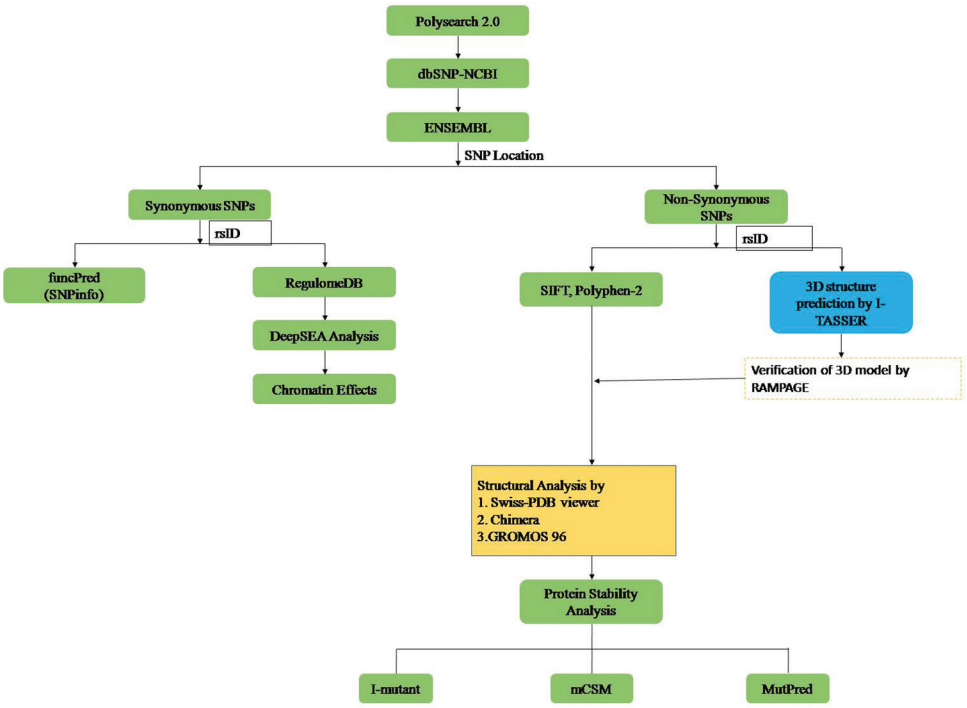

Figure 1: Schematic representation of computational tools for *in silico* analysis of single-nucleotide polymorphisms in gallbladder cancer

Table 1: List of single-nucleotide polymorphisms in coding and noncoding region predicted by FuncPred

dbSNP ID	Chromosome	Position	TFBS	miRNA (miRanda)
SNPs affecting Regulatory role alteration				
rs1065778	15	49307498	-	-
rs1065779	15	49292103	-	-
rs11267919	4	3487899	-	-
rs11614913	12	52671866	Y	-
rs1361530	1	119767087	-	Y
rs1569686	20	30830740	-	-
rs1800775	16	55552737	Y	-
rs1801132	6	152307215	-	-
rs1819698	1	119767042	-	Y
rs2234693	6	152205028	-	-
rs2304463	15	49295412	-	-
rs2606345	15	72804229	Y	-
rs2695121	19	55572553	-	-
rs2830	17	37958089	Y	-
rs2910164	5	159844996	Y	-
rs2976392	8	143759934	Y	-
rs2978974	8	143748866	Y	-
rs35463555	19	55569492	Y	-
rs3746444	20	33041912	Y	-
rs3808607	8	59575478	Y	-
rs3824260	8	59575744	Y	-
rs4633	22	18330235	-	-
rs4646	15	49290136	-	Y
rs4818	22	18331207	-	-
rs523349	2	31659210	-	-
rs700518	15	49316404	-	-
rs708272	16	55553789	Y	-
rs743572	10	104587142	Y	-
rs7922612	10	95801429	-	-
rs9340799	6	152205074	-	-

SNPs in the coding region (nsSNP)				
dbSNP ID	**Chromosome**	**Position**	**Protein associated**	**Predicted damaging by SIFT or PolyPhen**
rs10012	2	38155894	Cytochrome P450	-
rs1042838	11	100438622	Progesterone receptor	✓
rs1048943	15	72800038	Cytochrome P450 1A1	-
rs1056836	2	38151707	Cytochrome P450 1B1	-
rs11887534	2	43919751	ABCG8	✓
rs2066479	9	98037631s	Testosterone 17-beta-dehydrogenase 3	-
rs2274223	10	96056331	Phospholipase C epsilon 1	-
rs2294008	8	143758933	Prostate stem cell antigen	-
rs4148217	2	43952937	ABCG8	-
rs6259	17	7477252	Sex hormone-binding globulin	-
rs700519	15	49295260	Cytochrome P450 19A1	✓

Y: SNPs that affect function, ✓: SNPs predicted to be damaging, -: SNPs that don't affect function, SNPs: Single-nucleotide polymorphisms, TFBS: Transcription factor binding site, nsSNP: Nonsynonymous SNP, dbSNP: SNP database, ABCG8: ATP-binding cassette subfamily G member 8, SIFT: Sorting intolerant from tolerant

Table 2: Parameters of protein in RAMPAGE

Protein	Favored (%)	Allowed (%)	Outliers (%)
Progesterone receptor	65.2	22.1	12.7
Aromatase	88.8	9	2.2
ABCG8	76.2	14.6	9.2

ABCG8: ATP-binding cassette subfamily G member 8

function of the protein by analysis of multiple sequence alignment and protein three-dimensional (3D) structure,[20] the protein sequence, database identifiers/accession number, the position at which substitution takes place, the amino acid being substituted, and the amino acid present in the variant type. For our study, each SNP with their respective rsIDs was

uploaded, and the study was done for every nsSNPs. Prediction outcomes could be classified as probably damaging, possibly damaging, or benign according to the PolyPhen-2 score ranging from 0 to 1,[18] The score refers to the amino acid substitution in the variant type being damaging. These scores are represented as HumDiv scores, compiled from all damaging alleles with known effects on molecular function, and HumVar scores which consist of human disease-causing mutations. The closer the HumDiv score is to 1, it is indicative of greater damaging nature of the SNP.

SIFT is a module that takes in a query sequence and uses information of multiple alignments for the prediction of tolerated and deleterious substitutions for every position of the query sequence. It obtains the multiple alignments of chosen sequences and calculates the normalized probabilities for all possible substitutions at each position from the alignment. Substitutions at each position are predicted to be either damaging or tolerated based on whether the normalized probabilities are less than or greater than a tolerant index of 0.05. The nsSNPs were analyzed to obtain the damaging SNPs. The SNPs found to be damaging were shortlisted for further analysis.

Protein structure prediction

FASTA mRNA sequence was obtained for the proteins wherein the SNP was found to be damaging. The site of SNP was identified in the protein sequence, and the SNP was subjected to further analysis.

I-TASSER web server[21] enabled generation of the 3D structure of the protein and also predicted the biological functions of protein molecules from amino acid sequence. It provides five models based on the amino acid sequence, and each model is assigned an individual confidence scores (C-score) calculated based on the significance of threading template alignments and the convergence parameters of the structure assembly simulations.[18] A higher C-score indicates greater confidence and vice versa. It also provides a template modeling (TM) score which measures the structural similarity between two structures. A TM score >0.5 indicates a model of correct topology whereas a TM score <0.17 indicates a random similarity. It also predicts solvent accessibility and ligand-binding sites. Protein sequence was submitted to the webserver, and the protein structure was obtained.

Verification of three-dimensional model of protein

RAMPAGE Ramachandran plot analysis was used for verification of 3D structures. It provides the number of residues in the favored, allowed, and outlier region.[22] If a good proportion of residues lie in the favored and allowed region, then the model is predicted to be good.

Modeling nonsynonymous nucleotide polymorphisms on protein structure

Generation of the mutated models of the selected protein structure for the corresponding amino acid substitution was achieved using Swiss-PdbViewer (v4.10).[23] The native amino acid is replaced with the variant and the.pdb files for the model were saved. GROMACS is used as a default force field used by the server for energy minimization calculations. Several parameters such as total energy (KJ/mol), total electrostatic constraint, total bond energy (KJ/mol), torsion (KJ/mol), and nonbonded energy (KJ/mol) were calculated after energy minimization for native as well as mutant structure.

The.pdb files were further analyzed by Chimera, an effective molecular modeling system used for interactive visualization and analysis of molecular structures, sequence alignment, docking analysis, supramolecular assemblies, trajectories, and conformational analysis.[24]

Protein stability prediction

I-Mutant 2.0, mCSM, and MutPred were used to predict the protein stability upon the mutation. I-Mutant 2.0[25] was used to predict the stability upon single-site mutation.[26] The user can either provide protein structure or sequence as the input. Along with the protein structure or sequence, temperature and pH need to be specified. The protein sequence consisting the damaging SNP was uploaded to I-Mutant and output obtained is in the form of protein stability change upon mutation and Gibbs-free energy change. MutPred[27] predicts the probabilities of gain or loss of a function due to a particular polymorphism, thus providing an insight to molecular mechanism responsible for the disease.[28] The output contains top 5 property scores (p) where P is the *P* value that certain structural and functional properties are impacted.

mCSM machine learning method is used to predict the effects of missense mutations based on structural signatures. mCSM extracts geometric and physicochemical patterns (represented in terms of pharmacophores) using a graph representation. These are then used to represent the 3D chemical environment during supervised learning. Application of these signatures is done in a range of tasks including the protein structural classification and function prediction, as well as prediction of large-scale receptor based on protein-ligand prediction.[29]

Noncoding single-nucleotide polymorphisms functional analysis

Tools predicting potential functional effects of SNPs in noncoding binding sites such as intron/exon border consensus sequences (splice sites), transcription factor binding sites, exonic splicing enhancers (ESEs), and microRNA (miRNA) binding were used. SNPinfo (FuncPred).[30] and RegulomeDB[31] were used to screen SNPs based on their functionality for further genetic mapping services. The functionality of the SNPs was determined by SNPinfo (FuncPred) web server which helps in studying the SNPs for genetic association studies and consists of three pipelines for SNP selection. Among the several tools, i.e., TagSNP, FuncPred, and SNPseq in SNPinfo web server, to study the functionality of the SNPs, we chose FuncPred prediction software, which is a composite of PolyPhen, SNP3D, MATCH, TRANSFAC 12.1, RESCUE-ESE, ESEfinder, FAS–ESS, miRanda, and miRBase. Queries were

submitted for all the SNPs in the genes or chromosomal region or with their respective rsID. The SNPs related to the GBC were entered in a batch submission with their respective rsID. The output was a list of SNPs with possible functional effect.

RegulomeDB software was also used to complement the SNP prioritization. Utilizing the online composite database integrating a large collection of regulatory information and prediction tools to annotate and prioritize potential regulatory variants derived from genomic sequencing, RegulomeDB uses databases that take their datasets from chromatin immunoprecipitation sequencing (ChIP-seq), histone ChIP-seq, chromatin state information, and expression quantitative trait loci (eQTL) information. Thus, it helps in in silico predictions through DNase footprinting to identify protein binding sites, transcription factors binding domain and regulatory binding motif variations of nucleotide variants.[31] It divides the variants into six categories. Category 1 variants are likely to affect binding and linked to expression of a gene target. Category 2 variants are likely to affect binding. Category 3 variants are less likely to affect binding. Category 4, 5, and 6 variants have minimal binding evidence.[18] A batch of noncoding SNPs and synonymous SNPs was uploaded to the software with their respective rsIDs and each SNP was assigned a score from 1 to 6. The SNPs predicted to affect binding were shortlisted for further analysis by DeepSEA.

The chromatin effects of single nucleotide alteration in sequences were predicted using DeepSEA.[32] Through the "*in silico* saturated mutagenesis" approach, prediction of chromatin feature informative sequence elements for any sequence can be identified. DeepSEA accurately predicts the effect of individual SNPs on TF binding with the DeepSEA TF-binding classifiers, which are demonstrated for several SNPs that have experimentally validated well-known effects on TF binding. DeepSEA prioritizes functional SNPs on the basis of the predicted signals of chromatin effect. DeepSEA supports three types of input: vcf, FASTA, and bed. We used vcf format for predicting effects of noncoding variants. In vcf format, each line has at least 5 tab-separated fields. The first column consists of chromosome name. The second column contains position in a chromosome. The variant name is specified in the third column. The fourth and the fifth column contains reference allele (wild type) and alternative allele (mutant). Log fold changes along with probability differences and E-values are essential for evaluating variant's impact.

Network analysis of potential proteins with other proteins involved in gallbladder carcinoma

GeneMANIA was availed for network analysis of potential biomarker proteins. Protein-protein, protein-DNA, and genetic interactions, pathways, reactions, gene and protein expression data, protein domains, and phenotypic screening profiles are included in GeneMANIA searches.[33]

Results

Using FuncPred we found, among the 41 SNPs, 30 SNPs [Table 1] were found to be either synonymous or noncoding and 11 SNPs were found to be non-synonymous. Separate pipelines were designed for the study of SNPs in the coding region and in the noncoding region [Figure 1].

From the PolyPhen-2 and SIFT analysis, the SNPs in the coding region predicted to have deleterious effect were selected for further analysis.

All the 11 SNPs in the coding region [Table 1] were submitted to the PolyPhen-2 server. One nsSNP rs11887534 [Table 1] associated with ATP-binding cassette (ABC) subfamily G member 8 was found to be possibly damaging with a very high HumDiv score of 0.769 and HumVar score of 0.525. Analysis by SIFT predicted two other variants (rs1042838 and rs700519) to be highly damaging. rs1042838 associated with progesterone receptor was found to be damaging and had a median conservation score of 3.56 [Table 1]. rs700519 associated with cytochrome P450 19A1 (aromatase) was predicted to be damaging and had a median conservation score of 3.05. We moved forward with these three highly damaging SNPs (rs11887534, rs1042838, and rs700519) to find their structure-function alteration [Figures 2-4].

Structural analysis of native and mutant protein

rs11887534 is associated with ABCG8 gene. FASTA sequence of native ABCG8 was obtained. The location of SNP was obtained from NCBI, and the position of mutation was found out. Native protein sequence was submitted to I-TASSER in their respective FASTA format. Five models of native ABCG8 were obtained as an output; of these, the first model with highest C score of − 0.90, TM score of 0.60 ± 0.14, and estimated RMSD 10.1 ± 4.6 Å was considered for the analysis. Similarly, out of the five models obtained for native aromatase protein, the first model with highest C score of −0.69, TM score of 0.81 ± 0.09, and estimated RMSD 5.8 ± 3.6 Å, and

Table 3: Native versus protein mutant out of single-nucleotide polymorphism

	ABCG8 native	ABCG8 mutant	Progesterone native	Progesterone mutant	Aromatase native	Aromatase mutant
Total Energy (KJ/mol)	-27918.332	-29442.268	-21743.91	-22558.859	-25110.941	-25472.408
Total electrostatic constraint	-21661.68	-21797.66	-20554.77	-20262.72	-14933.51	-14746.16
Total bond energy (KJ/mol)	675.189	639.677	990.564	926.881	446.003	406.257
Torsion (KJ/mol)	7057.365	6738.392	10358.655	9962.072	3511.209	3376.156
Non Bonded energy (KJ/mol)	-20744.88	-21399.17	-21323.85	-21740.44	-17653.67	-17933.77

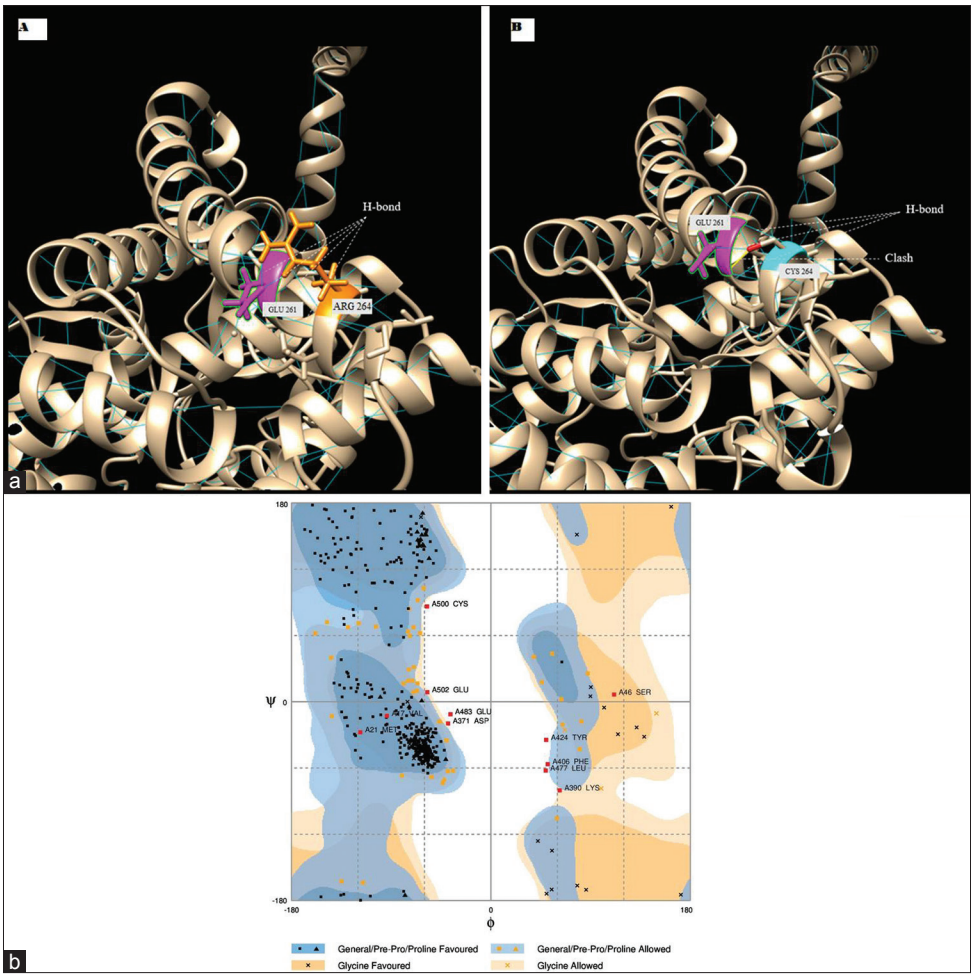

Figure 2: (a) Three-dimensional analysis of wild and valiant residues of aromatase at position 263. A: Arg (orange color at 264 position). B: Network clashes (yellow lines, indicated with white dotted lines) appeared between variant residue Cys 264 (light blue color) with Glu (magenta color). (b) Ramachandran of predicted aromatase secondary structure

Table 4: Prediction by MutPred for protein structural stability

Mutation	Protein	Top feature alteration
D19H	ABCG8	Gain of glycosylation at T16
		Gain of catalytic residue at D19
		Loss of phosphorylation at S21
		Loss of helix
		Gain of loop
R264C	Aromatase	Loss of methylation at K262
		Loss of MoRF binding loss of catalytic residue at R264
		Gain of ubiquitination at K262
V660L	Progesterone receptor	Loss of catalytic residues at G661
		Gain of glycosylation at P663
		Loss of MoRF binding

MoRF: Molecular recognition feature, ABCG8: ATP-binding cassette subfamily G member 8

for progesterone receptor, the first model with highest C score of −0.29, TM score of 0.75 ± 0.10, and estimated RMSD 8.1 ± 4.4 Å were considered for the analysis.

RAMPAGE was used to support the quality of predicted protein models. All the three protein models showed good proportion of residues in the favored and allowed regions in RAMPAGE [Table 2 and Figures 2-4].

To get the variant modelled structure, Swiss-PdbViewer was used. The D19H polymorphism in the ABCG8 [Tables 3 and 4] showed deviation from the native model in various parameters including total energy after minimization, total electrostatic constraint, total bond energy, torsion, and nonbonded energy. The molecular analysis by Chimera showed change in hydrogen bonding due to the D19H point mutation [Figure 4]. The R264C mutation in the cytochrome P450 19A1 (aromatase) showed deviation from the native model [Table 3]. The Arg → Cys variation at position 264 showed clashes with its neighboring glutamate residue at position 261 [Figure 2]. This variation also caused a change in hydrogen bonding of Cys at 264 with its neighboring residues. The Val → Leu variation at position 660 in the progesterone receptor [Figure 3] showed clashes with its neighboring Phe residue at 794 position and also caused a change in hydrogen bonding.

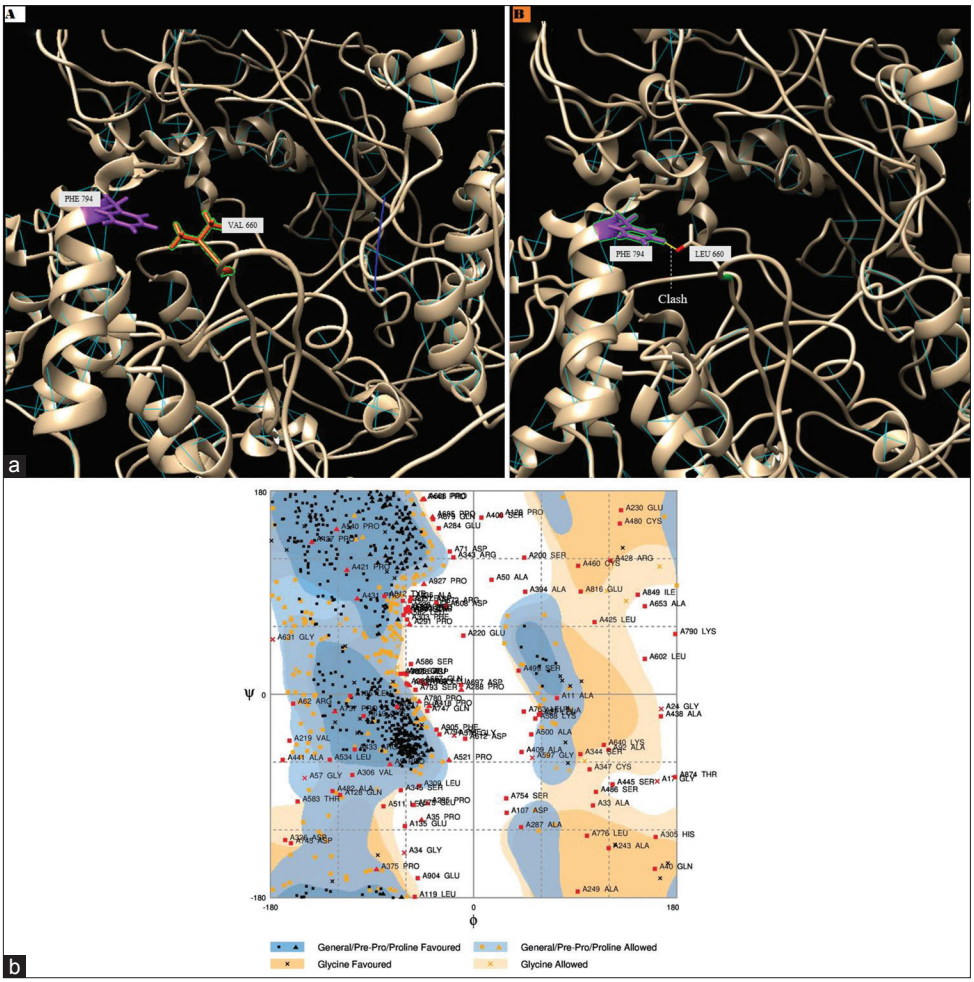

Figure 3: (a) Three-dimensional analysis of wild and valiant residues of progesterone at position 660. A: Val (red color) at 660 position. B: Network of clashes (yellow lines, indicated with white dotted lines) appeared between variant residue Leu 660 with Phe 794 (violet color). (b) Ramachandran of predicted progesterone receptor secondary structure

Table 5: Protein structural stability analysis

SNP	Protein associated	ΔΔG (predicted by mCSM)	I-mutant results
D19H	ABCG8	−1.107 kcal/mol	Destabilizing
V660L	Progesterone receptor	−0.349 kcal/mol	Destabilizing
R264C	Aromatase	−0.689 kcal/mol	Destabilizing

SNP: Single-nucleotide polymorphism, ABCG8: ATP-binding cassette subfamily G member 8

Prediction of protein structural stability and other effects

Analysis by I-Mutant indicates that conversion of negatively charged aspartic acid to positively charged histidine amino acid at position 19 decreases the stability of ABCG8 gene. I-Mutant also predicted the other two conversions, Arg → Cys variation at position 264 and Val → Leu variation at position 660 to be destabilizing. All the three SNPs were predicted to be destabilizing by mCSM software due to their negative ΔΔG values [Table 5].

MutPred predicted structural alterations due to the SNPs. The D19H mutation leads to gain of glycosylation at T16, gain of catalytic residue at D19, loss of phosphorylation at S21, and loss of a helix [Table 4]. The R264C mutation causes loss of methylation at K262, loss of molecular recognition feature (MoRF) binding, loss of catalytic residue at R264, and gain of ubiquitination at K262. The V660 L mutation causes loss of catalytic residues at G661, gain of glycosylation at P663, and loss of MoRF binding. The results obtained from MutPred have been summarized in Table 4.

Functional analysis of noncoding and synonymous single nucleotide polymorphisms

A list of 30 SNPs was submitted to the FuncPred program and the results obtained are summarized in Table 1. Out of these 30 SNPs [Table 1], a total of 16 SNPs were predicted to have effect on function; out of which, 3 SNPs were found to affect miRNA-binding site and 13 SNPs were found to affect transcription-binding site. RegulomeDB was used to complement SNP prioritization. RegulomeDB divided 30 SNPs into six categories (Category 1 to Category 6), where 23 SNPs had annotation scores [Table 6] and the rest 7 SNPs had no annotation data (not shown in table). Out of 23 SNPs, four

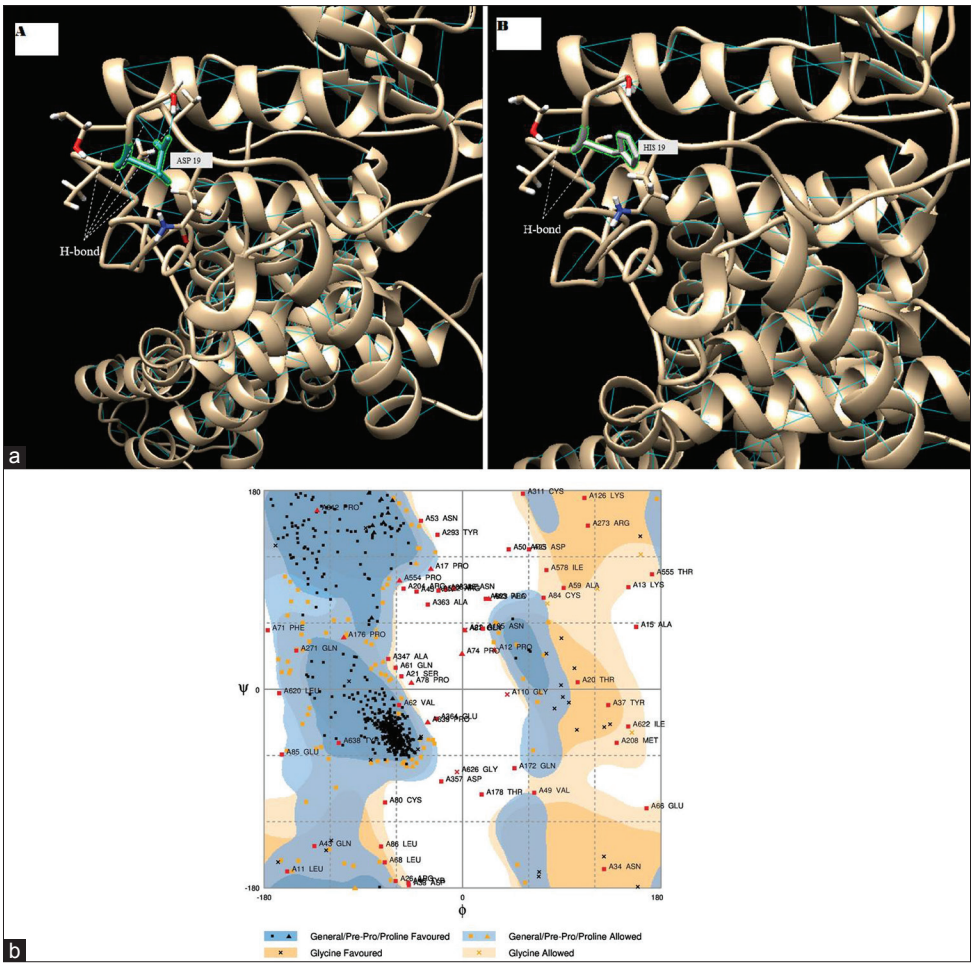

Figure 4: (a) Three-dimensional analysis of wild and valiant residues of ABCG8 at position. A: Asp (green color) at 19 position. B: Variant residue His19 (grey color). (b) Ramachandran plot of predicted ABCG8 secondary structure

SNPs were found to likely affect binding with a RegulomeDB score of 2b (Category 2), one SNP was found to less likely affect binding with a RegulomeDB score of 3a (Category 3), and the remaining 18 SNPs had minimum functional evidence (Category 4, 5, and 6). The four SNPs, i.e. rs2978974, rs4633, rs2830, and rs523349 which were predicted to likely affect binding had annotations for transcription factor binding, DNase peak, and motif hit and thus likely to have regulatory roles [Table 6].

All the 4 SNPs predicted to affect binding were subjected to DeepSEA analysis to predict the chromatin effects of sequence alterations with single-nucleotide sensitivity. The.vcf files for all four SNPs were submitted to DeepSEA. However, DeepSEA could analyze only two variants [Figure 5a-b]. It provided chromatin feature probability log fold changes for both the variants. The variant (rs2978974) associated with the PSCA gene showed five chromatin feature alterations to be significant with an $E \leq 0.02$ [Figure 5a and Table 7]. The gene interactions of the four essential biomarker proteins (ABCG8, cytochrome P450 19A1, progesterone receptor, and PSCA) with other proteins involved in GBC are shown in Figure 6.

Discussion

Gallbladder carcinoma, a malignant adenocarcinoma of gallbladder epithelial tissue, has high mortality rate and lacks suitable markers for early diagnosis and prognosis. It is essential to detect the potential biomarker out of the several genetic markers and protein markers available. Since it is difficult to study all the SNPs associated with GBC, interpretation of clinically essential novel markers is always challenging. In the current study using *in silico* analysis, we could screen important variants which could be potential therapeutic targets.

In this study the first of its kind, by *in silico* analysis of SNPs involved in GBC performed by several bioinformatics tools, we report (i) seven novel SNPs in coding as well as noncoding region which could serve as essential biomarker and could be potential therapeutic targets, (ii) Both synonymous and nsSNPs as potential biomarker, (iii) understanding of the structure-function alteration in nsSNPs, and (iv) regulatory role alteration by SNPs in noncoding region.

Separate pipelines designed for the study of SNPs in coding and noncoding region are unique in this study to understand

Table 6: List of single-nucleotide polymorphisms predicted by RegulomeDB score

dbSNP ID	RDB score	Category	Description
rs3824260	6	Minimal binding evidence	Motif hit
rs3808607	4	Minimal binding evidence	TF binding + DNase peak
rs2976392	5	Minimal binding evidence	DNase peak + motif hit
rs2978974	2b	Likely to affect binding	TF binding + DNase peak + motif hit
rs708272	5	Minimal binding evidence	DNase peak + motif hit
rs1800775	3a	Less likely to affect binding	TF binding + DNase peak + motif hit
rs11267919	Nil	Nil	Nil
rs2606345	4	Minimal binding evidence	TF binding + DNase peak
rs700518	5	Minimal binding evidence	DNase peak
rs1065778	4	Minimal binding evidence	TF binding + DNase peak
rs2304463	Nil	Nil	Nil
rs1065779	6	Minimal binding evidence	Motif hit
rs4646	Nil	Nil	Nil
rs4633	2b	Likely to affect binding	TF binding + DNase peak + motif hit
rs4818	4	Minimal binding evidence	TF binding + DNase peak
rs2830	2b	Likely to affect binding	TF binding + DNase peak + motif hit
rs1819698	Nil	Nil	Nil
rs1361530	6	Minimal binding evidence	Motif hit
rs523349	2b	Likely to affect binding	TF binding + DNase peak + motif hit
rs1569686	5	Minimal binding evidence	DNase peak
rs2910164	Nil	Nil	Nil
rs11614913	5	Minimal binding evidence	DNase peak
rs3746444	5	Minimal binding evidence	DNase peak + motif hit
rs743572	4	Minimal binding evidence	TF binding + DNase peak
rs7922612	5	Minimal binding evidence	DNase peak
rs35463555	4	Minimal binding evidence	TF binding + DNase peak
rs2695121	4	Minimal binding evidence	TF binding + DNase peak
rs1801132	4	Minimal binding evidence	TF binding + DNase peak
rs9340799	Nil	Nil	Nil
rs2234693	Nil	Nil	Nil

RDB: RegulomeDB, TF: Transcription factor, dbSNP: SNP database, SNP: Single-nucleotide polymorphism

the functional role of each variant. Three potential variants present in the coding region were predicted by SIFT and PolyPhen. rs1042838, rs11887534, and rs700519 associated with progesterone receptor, ABC subfamily G member 8, and cytochrome P450 19A1, respectively, were predicted to be potentially damaging SNPs in GBC.

Human ATP transporters, 48 in number, comprising of 7 subfamilies, sharing similar structure and function by transportation of molecules across membranes by utilizing the energy derived from hydrolysis of ATP. Of the seven distinct subfamilies of ABC genes including (ABC1, MDR/TAP, MRP, ALD, OABP, GCN20, and White), ABCG8 belongs to the subfamily White. Very recently, the crystal structure of the human sterol transporter ABCG5/ABCG8 has been deciphered.[34,35] ABCG8 or ATP-binding cassette subfamily G member 8 is an apical membrane sterol export pump, expressed in specific tissues of gallbladder, liver, and intestine. It belongs to the superfamily of ABC transporters that boost the active efflux of cholesterol from hepatocytes to bile.[35,36] Excessive amount of cholesterol content in bile is associated with development of gallstones. The D19H polymorphism in ABCG8 gene is proposed to increase the expression of ABCG8 or enhance its function, resulting in more efficient transfer of cholesterol into bile and the accumulation of cholesterol in the gallbladder forming the key step for gallstone formation.[36] Although D19H polymorphism in ABCG8 has been earlier reported to be associated with the risk in gallbladder stones,[12] no study has been done to identify the functional significance of SNP in gallbladder cancer. Since full-length protein structure of native ABCG8 is not yet available in the protein data bank, it was essential to generate a full-length model of native ABCG8 and find the structure-function alteration of D19H polymorphism. Only a partial structure of ABCG8 with PDB ID 5DO7 exists. The model of ABCG8 generated was compared in terms of several parameters such as total energy after minimization, total electrostatic constraint, total bond energy, torsion, and nonbonded energy for both native and mutant structure, and a clear deviation was observed for native and mutant structure. A variation in terms of hydrogen bonding difference was observed. The D19H mutation was also predicted to cause a gain of glycosylation at T16, gain of catalytic residue at D19, loss of phosphorylation at S21, and loss of a helix. Our

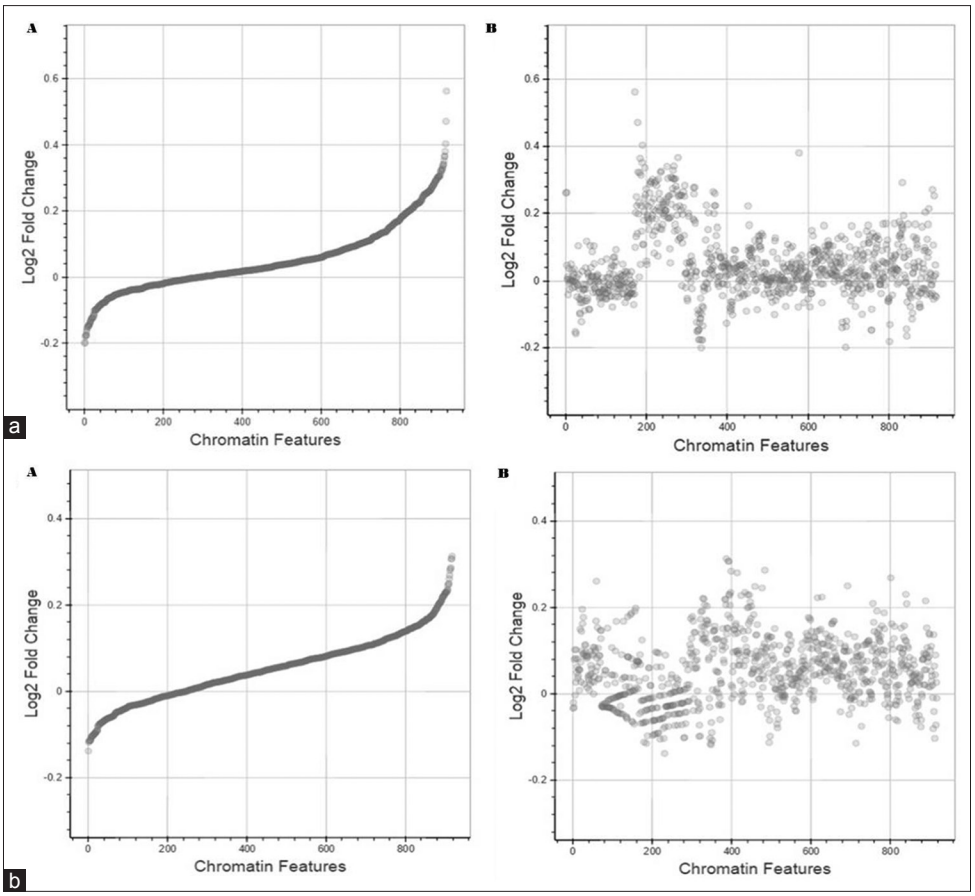

Figure 5: (a) A: Chromatin feature alteration sorted by log2 fold change in PSCA gene. B: Chromatin feature alteration sorted by chromatin feature in PSCA gene. (b) A: Chromatin feature sorted by log2 fold change in SRD5A2 gene. B: Chromatin feature alteration sorted by chromatin in SRD5A2

Table 7: Chromatin feature alteration for PSCA gene

Chromatin features	Cell type	Effect (log 2-fold change)	*E*	Normalized (*P*)	
				Reference	Alternative
EZH2	H1-hESC	−0.17634	0.011047	0.11462	0.102839
SUZ12	NT2-D1	−0.14663	0.014651	0.074535	0.067712
EZH2	HUVEC	−0.17546	0.015497	0.050659	0.044996
EZH2	NHDF-Ad	−0.14576	0.015604	0.060729	0.054996
EZH2	NHLF	−0.11262	0.018788	0.104873	0.097851

findings are indicative of the fact that there could be a possible single-point factor D19H polymorphism in ABCG8 gene that is associated with the genesis of both gallbladder cancer and gallbladder stones and finds importance as a biomarker in both disorders as leading to significant alteration in the structure and function of the protein.

The R264C polymorphism in aromatase and V660 L polymorphism in progesterone receptors show a clear deviation of energy parameters for native and mutant structure. A possible reason for R264C polymorphism to decrease the stability could be due to the smaller size cysteine residue in the mutant structure which shows clash with the neighboring glutamate residue at position 261. V660 L polymorphism causes a clash between leucine and phenylalanine residue at 794 position. This clash could possibly be the reason due to the decrease in stability of the mutant structure.

SNPs either noncoding or synonymous could impair the regulatory roles and thus find an important place to be potential biomarker for GBC. RegulomeDB provided three SNPs (rs2978974, rs4633 and rs2830) [Table 6] either synonymous or non-coding, which had annotations for transcription factor binding, DNase peak, and motif hit and thus have regulatory roles. Out of these three SNPs, one SNP rs2978974 associated with PSCA showed significant chromatin feature alteration of EZH2 and SUZ12 with $E \leq 0.02$ [Table 7]. Thus, this variant associated with

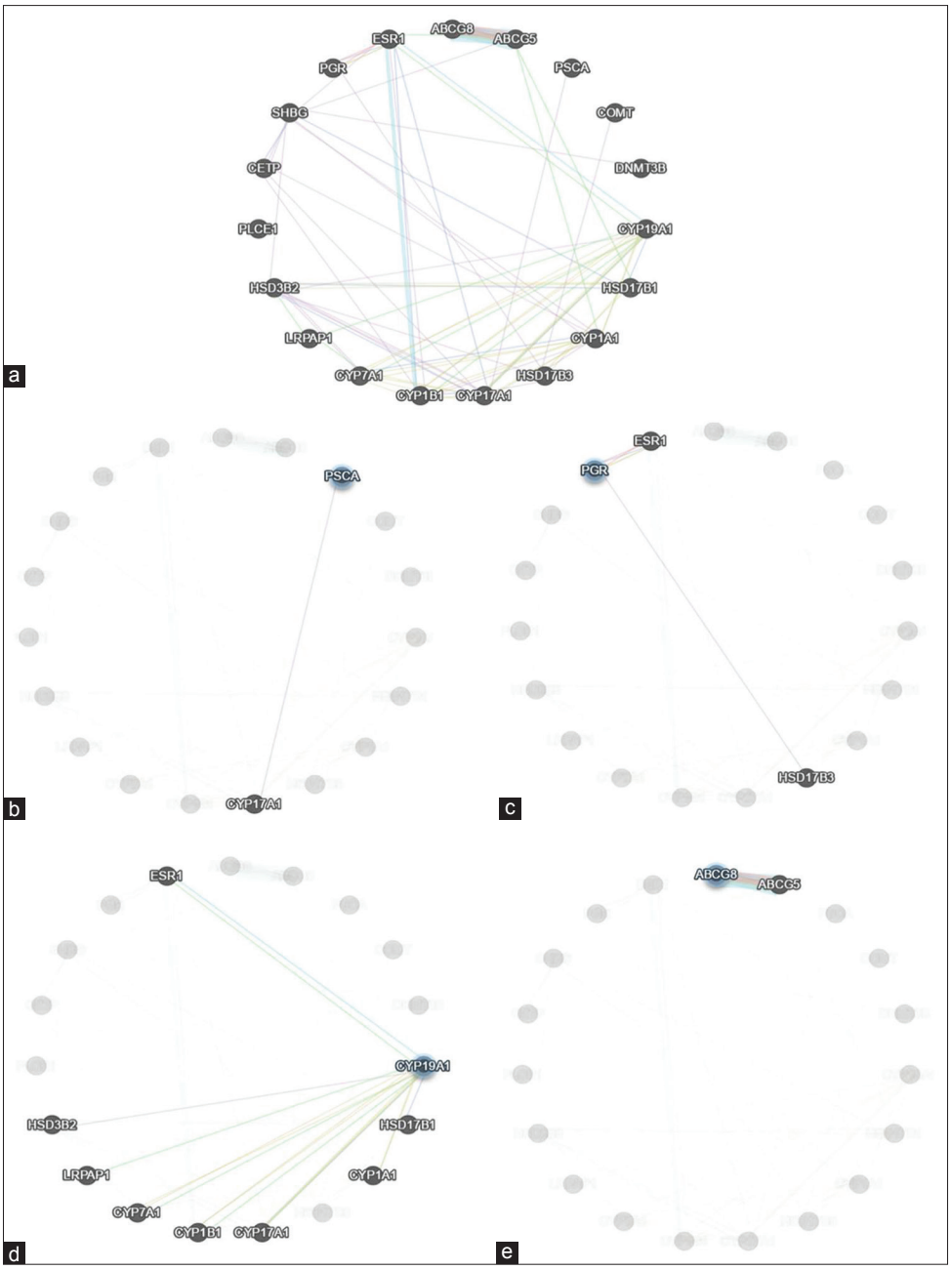

Figure 6: Gene interaction network. (a) Overall interaction of the genes involved in gallbladder cancer. (b) PSCA interacts with CYP17A1. (c) PGR interacts with ESR1 and HSD17B3. (d) CYP19A1 interacts with HSD17B1, CYP1A1, ESR1, HSD3B2, LRPAP1, CYP7A1, CYP7B1, and CYP17A1. (e) ABCG8 interacts with ABCG5

PSCA can be potential biomarker. PSCA has been shown to be downregulated due to methylation in nonneoplastic gallbladder lesions with growth-suppressive effects in adenocarcinoma[5] suggesting its function as a tumor suppressor. Although the role of PSCA has been reported in several physiological functions including cell adhesion, signal transduction, inhibition of cell proliferation, and/or induction of cell death, its biological functions in carcinogenesis are not yet fully understood.

In our study through *in silico* approaches, we strongly report for the first time that (i) D19H polymorphism in ABCG8, R264C polymorphism in cytochrome P450 19A1, and V660 L polymorphism in progesterone receptor are the most deleterious nsSNPs in gallbladder cancer. Due to the observed prominent alteration of the protein 3D structure, these markers find significance in drug targeting and therapeutic importance. (ii) We also report three SNPs (rs2978974, rs4633 and rs2830) [Table 6] either noncoding or synonymous and a missense SNP(rs523349), which had annotations for transcription factor binding, DNase peak, and motif hit and thus reveal a deleterious effect in GBC and finds significance as biomarkers for early diagnosis of GBC. (iii) The variant

rs2978974 in PSCA [Figure 5a] showed significant chromatin feature alteration and thus could be potential biomarker for early diagnosis.

The gene network analysis showed the interactions of these potential biomarker proteins with other proteins involved in gallbladder carcinoma. It is pretty much clear that the SNPs in the potential biomarker proteins could alter their interactions with other protein downstream, thus resulting in a major malfunction in the cascade of reaction. CYP19A1 interacts with HSD17B1, CYP1A1, ESR 1, HSD3B2, LRPAP1, CYP7A1, CYP7B1, and CYP17A1. PSCA interacts with CYP17A1. PGR interacts with ESR1 and HSD17B3. ABCG8 interacts with ABCG5.

Our current study has highlighted SNPs, both in the coding and noncoding region, that play important role in GBC and may be exploited as markers to the disease. Due to the common SNP of D19H in ABCG8 region being associated with both gallbladder cancer and gallbladder stones, the questions as to whether one disease leads to the other or are there other factors involved in the genesis of gallbladder cancer, which remains the future scope of research. The other two SNPs, R264C polymorphism in aromatase and V660 L polymorphism in progesterone receptors in the coding region along with the rs2978974 in noncoding region of PSCA, could also be important biomarkers and thus potential drug targets for gallbladder carcinoma.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Singh TD, Barbhuiya MA, Poojary S, Shrivastav BR, Tiwari PK. The liver function test enzymes and glucose level are positively correlated in gallbladder cancer: A cancer registry data analysis from North central India. Indian J Cancer 2012;49:125-36.
2. Pradhan A, Saha S, Ghosh S. Study of gallbladder cancer in the light of proteomics. PeerJ Preprints 2015;15;2167-9843.
3. Sharma KL, Rai R, Srivastava A, Sharma A, Misra S, Kumar A, *et al.* A multigenic approach to evaluate genetic variants of PLCE1, LXRs, MMPs, TIMP, and CYP genes in gallbladder cancer predisposition. Tumour Biol 2014;35:8597-606.
4. Rai R, Sharma KL, Sharma S, Misra S, Kumar A, Mittal B, *et al.* Death receptor (DR4) haplotypes are associated with increased susceptibility of gallbladder carcinoma in North Indian population. PLoS One 2014;9:e90264.
5. Rai R, Sharma KL, Misra S, Kumar A, Mittal B. PSCA gene variants (rs2294008 and rs2978974) confer increased susceptibility of gallbladder carcinoma in females. Gene 2013;530:172-7.
6. Sharma KL, Misra S, Kumar A, Mittal B. Association of liver X receptors (LXRs) genetic variants to gallbladder cancer susceptibility. Tumour Biol 2013;34:3959-66.
7. Wannhoff A, Hov JR, Folseraas T, Rupp C, Friedrich K, Anmarkrud JA, *et al.* FUT2 and FUT3 genotype determines CA19-9 cut-off values for detection of cholangiocarcinoma in patients with primary sclerosing cholangitis. J Hepatol 2013;59:1278-84.
8. Jiao X, Ren J, Chen H, Ma J, Rao S, Huang K, *et al.* Ala499Val (C and T) and lys939Gln (A and C) polymorphisms of the XPC gene: Their correlation with the risk of primary gallbladder adenocarcinoma – A case-control study in China. Carcinogenesis 2011;32:496-501.
9. Srivastava A, Srivastava A, Srivastava K, Choudhuri G, Mittal B. Role of ABCG8 D19H (rs11887534) variant in gallstone susceptibility in Northern India. J Gastroenterol Hepatol 2010;25:1758-62.
10. Wang Y, Jiang ZY, Fei J, Xin L, Cai Q, Jiang ZH, *et al.* ATP binding cassette G8 T400K polymorphism may affect the risk of gallstone disease among Chinese males. Clin Chim Acta 2007;384:80-5.
11. Katsika D, Magnusson P, Krawczyk M, Grünhage F, Lichtenstein P, Einarsson C, *et al.* Gallstone disease in Swedish twins: Risk is associated with ABCG8 D19H genotype. J Intern Med 2010;268:279-85.
12. Buch S, Schafmayer C, Völzke H, Becker C, Franke A, von Eller-Eberstein H, *et al.* A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. Nat Genet 2007;39:995-9.
13. Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C. Silent (synonymous) SNPs: Should we care about them? Methods Mol Biol 2009;578:23-39.
14. Liu Y, Liang Y, Wishart D. PolySearch2: A significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res 2015;43:W535-42.
15. Cheng D, Knox C, Young N, Stothard P, Damaraju S, Wishart DS, *et al.* PolySearch: A web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Res 2008;36:W399-405.
16. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, *et al.* Ensembl 2018. Nucleic Acids Res 2018;46:D754-61.
17. Aken BL, Ayling S, Barrell D, Clarke L, Curwen V, Fairley S, *et al.* The ensembl gene annotation system. Database (Oxford) 2016;2016. pii: baw093.
18. Kalia N, Sharma A, Kaur M, Kamboj SS, Singh J. A comprehensive *in silico* analysis of non-synonymous and regulatory SNPs of human MBL2 gene. Springerplus 2016;5:811.
19. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, *et al.* A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248-9.
20. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using polyPhen-2. Curr Protoc Hum Genet 2013;Unit7.20.
21. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols. 2010;5:725.
22. Lovell SC, Davis IW, Arendall WB 3rd, de Bakker PI, Word JM, Prisant MG, *et al.* Structure validation by calpha geometry: Phi, Psi and Cbeta deviation. Proteins 2003;50:437-50.
23. Guex N, Peitsch MC. SWISS-MODEL and the swiss-Pdbviewer: An environment for comparative protein modeling. Electrophoresis 1997;18:2714-23.
24. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, *et al.* UCSF chimera – A visualization system for exploratory research and analysis. J Comput Chem 2004;25:1605-12.
25. Capriotti E, Fariselli P, Casadio R. I-mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306-10.
26. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. ProTherm, version 4.0: Thermodynamic database for proteins and mutants. Nucleic Acids Res 2004;32:D120-1.
27. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, *et al.* MutPred2: Inferring the molecular and phenotypic impact of amino acid variants. Biorxiv 2017;134981.
28. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 2009;25:2744-50.
29. Pires DE, Ascher DB, Blundell TL. MCSM: Predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30:335-42.
30. Perron U, Provero P, Molineris I. *In silico* prediction of lncRNA function using tissue specific and evolutionary conserved expression. BMC Bioinformatics 2017;18:144.

31. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, *et al.* Annotation of functional variation in personal genomes using regulomeDB. Genome Res 2012;22:1790-7.
32. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods 2015;12:931-4.
33. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function. Genome Biol 2008;9 Suppl 1:S4.
34. Lee JY, Kinch LN, Borek DM, Wang J, Wang J, Urbatsch IL, *et al.* Crystal structure of the human sterol transporter ABCG5/ABCG8. Nature 2016;533:561-4.
35. Yoon JH, Kuver R, Choi HS. ABCG8 D19H polymorphism: A basis for the genetic prediction of cholesterol gallstone disease. J Gastroenterol Hepatol 2010;25:1713-4.
36. Srivastava K, Srivastava A, Mittal B. Caspase-8 polymorphisms and risk of gallbladder cancer in a Northern Indian population. Mol Carcinog 2010;49:684-92.