In Silico Analysis of nsSNPs Affecting Stability and Dynamics of P-Glycoprotein - a Breast Cancer Associated Protein Identified from Gene-Environment Interaction Studies

Himani Gupta, Yasha Hasija*

Department of Biotechnology, Delhi Technological University
Main Bawana Road, Shahbad Daulatpur, New Delhi, 110042
*Corresponding Author: yashahasija@gmail.com

Abstract

The correct identification of functional SNPs of any gene is an important aspect in the study of genomics but loss of consistent genotype and phenotype data hampers any experiment to characterize the functional influence of all SNPs in humans. Therefore, *in silico* methods assist in providing useful information for characterizing functional aspect of SNPs. In this study, we have made an intense effort to identify potentially functional SNPs influencing protein function in environment susceptible genes discovered in Breast Cancer pathway. For this we used set of bioinformatic tools that utilize homology-based structure profile information, sequence-based conservation profile, and support vector algorithm in order to examine the nsSNPs found in the breast cancer patients. ABCB1 is one such environment susceptible gene coding for P-glycoprotein which is found to be overexpressed in tumour cells and is the root cause for drug efflux in breast cancer. Six different somatic missense mutations in the human ABCB1 gene in breast cancer patients have been reported in COSMIC database as of 2014.

In this study we have applied a set of tools like PolyPhen 2.0, PhD-SNP, and MutPred to display with accurate prediction the disease-associated mutations in ABCB1 gene and their structural impact. Further, we have carried out molecular dynamic simulations (MDS) to study the molecular as well as structural role of predicted disease associated nsSNPs. MDS was used to observe the atomic interaction and motion trajectory of native and mutant (R538S and M701R) P-glycoprotein. Out of these six nsSNPs, two mutations R538S which is present in the ATP binding domain at NMD interface and M701R present in the TMD domain of P-gp have been predicted to be deleterious by our analysis.

Introduction

Breast cancer is known to be most common type of cancer in women, other than for skin cancer. It has been found to be second most cause of cancer death in women, second to lung cancer. About 90% to 95% of breast cancers are considered sporadic, which means that the genes get damaged by chance after a person is born and no risk involves in passing on the gene to future generation. Only 5% to 10% of breast cancers are inherited which is quite less Scientists are gradually learning that many chemicals which are commonly found in daily products might as well contributing to the very high occurrence of breast cancer apart from conventionally recognized risk factors for breast cancer (genetic profile, obesity, age, reproductive history, alcohol intake, smoking, etc.).

In order to overcome this situation we have used a comprehensive bioinformatics procedure to understand the role of gene variants which are interacting with the environment and as well as identifying genes which highly interact with chemicals and have found to possess variations in breast cancer. Genes which may be influenced by both exogenous and endogenous environmental factors need to be identified, as well as discovering variant genes that have been potentially under reported or understudied relation to breast cancer. Integration of pharmacological and toxicological databases such as the Comparative Toxicological Database (CTD) and Environmental Genome Project (EGP) data on genetic variation has also been useful in developing understanding comprehensively for research of GEI related disease. One such protein that has been identified by this method is p-glycoprotein.

P-glycoprotein belongs to an ATP-Binding Cassette (ABC) transporter family and functions as a physiological barrier to toxins and xenobiotics by extruding them out of cells (Srivalli et al., 2012; Sharom et al., 2011).Efflux of various chemically distinct amphipathic compounds, which also include anticancer drugs is carried out by this transporter by deriving energy from hydrolysis of ATP (Ambudkar et al., 1999).

It has been found widely distributed in tissues (Liu, 2009) and was the first ABC transporter displaying MDR due to its overexpression in breast cancer cell lines (Riordan et al., 1985).It averts cellular intake of chemotherapeutic agents thus making chemotherapy unsuccessful almost in many cases. Thus, this protein acts as the major barrier in treatment of breast cancer (Martins et al., 2010; Bansal et al., 2009).A number of strategies are being implemented so as to overcome the problems related with P-gp in optimal drug delivery which include inhibition of P-gp, and other various methods to bypass it (Goren et al., 2000; Mazel et al., 2001).

From our analysis, ABCB1 gene (P-gp) has been found to be environmental susceptible gene and somatic mutations in this gene can be attributed to environmental exposures. In this study we aim to analyse the role of these somatic non synonymous SNPs which have been taken from the COSMIC database on the structure and function of p-glycoprotein

by carrying out mutational analysis using tools like PolyPhen, PHD-SNP, MutPred and then studying the dynamic behavior of these mutations through molecular dynamic simulations. Since p-glycoprotein inhibition is essential for the success of cancer therapy, it becomes important to study the effect of somatic mutations present in the breast cancer patients samples since this would help in better inhibitor designing, hence prove to be good strategy for prevention from cancer.

Review of Literature

Gene-environment alterations are seen in various cancers, like breast cancer, glioma, colorectal cancer, lung cancer etc. and are currently being used clinically as diagnostic markers. Such alterations occur in lieu of response to external or internal environmental cues, and have been noticed to occur because of long term carcinogen exposure. Environmental effects might be direct or indirect. For instance, external agents like a chemical toxin, can enter the cells of a tissue on prolonged exposure and may interfere directly with the genetic material. Otherwise, an environmental condition, such as chronic stress, might stimulate the body to produce its own intrinsic epigenetic factors and indirectly affect the cell function by disruption in several molecular pathways of the cell and causing self-sufficiency in growth signals, evasion of apoptosis, growth control signals insensitivity, increasing replicative potential, invasion, metastasis and sustained angiogenesis (Hanahan and Weinberg, 2000; Balmain *et al.*, 2003).

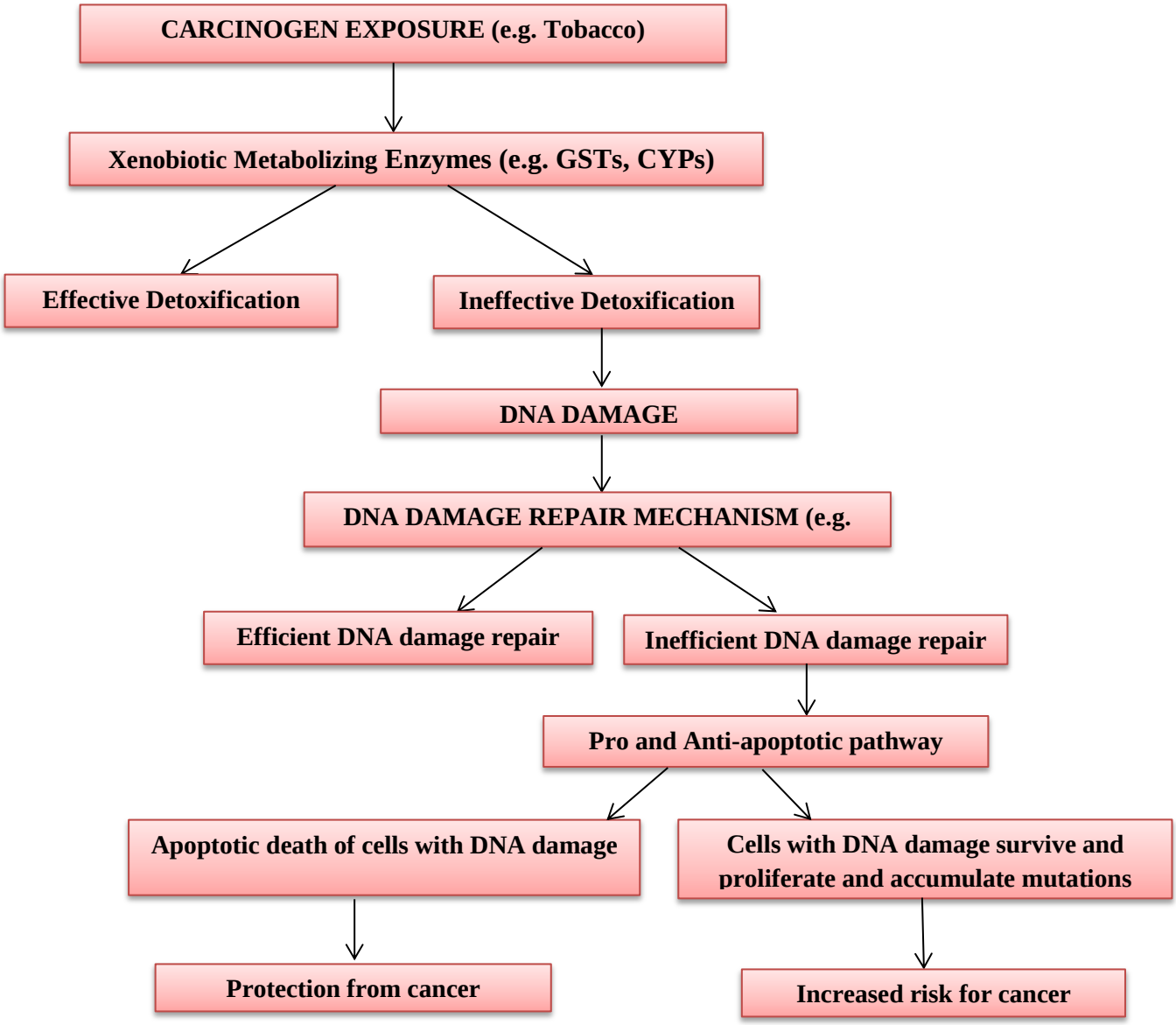


Figure 1: Multistep Carcinogenesis model

Therefore, it has been implicated that environmental exposure plays a key role in the etiology of cancer.

Scrotal cancers were found to be caused by soot which is carcinogenic and present in chimney sweepers (Doll, 1975). Since then it was established that play a dominant role is played by environmental factors in many sporadic cancers (Bostwick *et al.,* 2004).However, it is also known that all individuals which are exposed to the same type and amount of carcinogen do not develop cancer. So, cancer development is due to endogenous or exogenous carcinogens as well as their contact with genes that are involved in the carcinogen detoxification, DNA damage repair and cell signalling and cell cycle control. Development of sporadic cancers, due to carcinogen exposure may be aided by a cumulative effect of polymorphisms in these genes. Each allele, under this polygenic model, confers a little genotypic risk which combines additively to lead to a range of susceptibilities (Houlston and Peto, 2004).
Thus, only genetic predisposition is not responsible but a combination of exposures including environmental factors and susceptibility genes contribute to the development of sporadic cancers (Figure 1). Also, life styles such as alcohol consumption, smoking play a major role in increasing the risk of cancer. Genes can be altered by environmental chemicals in many ways, like physical interaction, altered methylation, and mutagenesis thereby affecting gene expression and protein function. Conversely, susceptibility to chemicals gets affected by genetic polymorphisms occurring naturally and thereby leading to augmented disease predisposition. The primary event in response to carcinogen exposure is damage of DNA which needs to be safeguarded. There are a number of genes which are responsible for this function and preventing cancer and thus are the cancer susceptibility genes.

Background/current status of activities in the area

INTERNATIONAL AND NATIONAL STATUS

A large number of studies have attempted to establish the role of polymorphisms in carcinogen metabolizing enzymes as well as other important genes involved in tumour susceptibility to environment interactions (Table 1).

Gene	Nucleotide/amino acid change	Associated cancer	Associated exposure	Mechanism of action	Reference
CASP8	rs17468277	Breast	Alcohol		Stefan Nickels *et al.*, 2013
CYP1A1	3¢non-coding region 6235 T > C	Breast, uterine	Oestrogen metabolites	Activating pro-carcinogens and catalyzing oxidative metabolites of oestrogen	Peto and Houlston, 2001
CYP1A1	Codon 462 Exon 7 Ile–Val	Lung	Tobacco habit	Activation of tobacco related PAH	London *et al.*, 2000
CYP1A2	5347 T > C	Lung, bladder, colorectal	Tobacco habit	Activation of nitrosamines and arylamines	Seow *et al.*, 2001
GSTM1 GSTT1	Deletion (null genotype)	Lung, bladder, breast, HNSCC, colon, uterine, stomach	Tobacco habit	Carcinogen detoxification of oxidative metabolites	Miller *et al.*, 2002;Jhavar *et al.*, 2004
NAT2	C282T and T341C	Bladder, colon, liver	Tobacco habit	Carcinogen detoxification of aromatic amines, hydrazines	Tiemersma *et al.*, 2004
CHEK2	1100 del. missense variant I157T	Breast, prostate		DNA damage and replication checkpoint	Varley and Haber 2003
p53	Codon 72 (Arg–Pro)	Lung	Tobacco habit	Apoptosis regulation	Fan *et al.*, 2000

XRCCI	Codon Arg399Gln Arg194Trp	Breast, oesophageal cancer, HNSCC	Tobacco habit	DNA repair	Shu *et al.*, 2003; Xing *et al.*, 2002
hOGGI	Ser326Cys	Lung	Exposure to tobacco smoke	Oxidatively damaged DNA repair e.g. 8-oxo-G DNA adducts	Park *et al.*, 2004
SULT1A1	Arg213His	Breast, bladder	Oestrogen, tobacco	Catalyzes the sulfation of phenolic and estrogenic compounds, metabolism of polycyclic aromatic hydrocarbons (PAHs) and aromatic amines	Han *et al*., 2004
Alcohol dehydrogenase 3 (ADH3)	Ile349Val	UADT, colorectal adenomas	Alcohol	Alcohol metabolism	Nishimoto *et al*., 2004

Table1: Role of polymorphisms in carcinogen metabolizing enzymes and other important genes.

Though few GEI studies in glioma exist, a computational modelling of glioma (de Andrade *et al*., 2001), and recent epidemiological studies, have begun to assess and support the potential role of GEI in the development of GBM and other brain tumours (Searles Nielson *et al*., 2005; 2010; De Roos *et al*., 2006; Rajaraman *et al*., 2006; Bhatti *et al*., 2009). While these studies are an important step, they have only assessed a small number of mutations potentially involved in GEI and brain tumour development. Given their generally positive results linking genetic mutations and environmental exposures to brain tumour development, identifying other potential GEI candidates for brain tumours is critical.

Various genetic susceptibility loci have been discovered for breast cancer; however, it is still unclear how they merge with environmental/ lifestyle risk factors to influence cancer risk. In one of the study they undertook an international collaborative study for assessing gene environment interaction for breast cancer risk (Stefan Nickels *et al* 2013).They pooled data from 24 studies of the Breast Cancer Association Consortium and used up to 34,793 invasive breast cancers and 41,099 controls, they examined whether the relative risks associated with 23 (SNP's) single nucleotide polymorphisms were modified by 10 established environmental risk factors (age at menarche, parity, breastfeeding, body mass index, height, oral contraceptive use, menopausal hormone therapy use, alcohol consumption, cigarette smoking, physical activity) in women of European ancestry. They used logistic regression models stratified by study and adjusted for age and performed

likelihood ratio tests to assess gene–environment interactions. They found interactions between several Breast Cancer associated genes which provides first strong evidence that the risk of Breast Cancer associated with some common genetic variants may vary with environmental risk factors which include diethylstilbestrol, a synthetic form of estrogen that was used to prevent miscarriages; steroidal estrogens used for menopausal therapy; X-ray and gamma radiation; alcoholic beverages; tobacco smoking; and the sterilizing agent, ethylene oxide.

Therefore, it is essential to study the potential genetic and environmental risk factors jointly in order to achieve a better understanding of the mechanisms underlying the disease .So; we need to follow up these genetic findings with more detailed analyses to better understand how the gene and environment interact in their influence on disease risk. Study involving an attempt to identify genes potentially important in environmentally related alterations in cancer by applying bioinformatics methods becomes crucial in this respect.

Recent developments in bioinformatics, such as tools for the assessment of pathway, text mining of published literature and gene relationships as well as integration of large amounts of diverse environmental and biological information allow for hypothesis driven investigation of gene–gene interaction and GEI. Methods that exploit these tools have been applied to modelling of GEI in depression and alcohol use (McEachin *et al*., 2008), and bipolar disorder and its interaction with both tobacco and lithium treatment. Integration of toxicological and pharmacological databases such as the Comparative Toxicological Database (CTD) and the Environmental Genome Project (EGP) (Rieder Q3 *et al*., 2008) with data on genetic variation has also proven useful in comprehensive understanding for research into GEI related diseases (Herbert *et al*., 2006; Bauer-Mehren *et al*., 2011).

40,000 women still die because of breast cancer each year in spite of advanced treatments and increased awareness among many women. Every two minutes a woman is being diagnosed with breast cancer. According to facts in the 1960s, lifetime risk for a woman's breast cancer was 1 in 20. Today it has worsened and reached 1 in 8. Breast tissue develops and matures during early childhood and adolescence, and according to recent studies during these critical stages of development, certain chemicals, environmental exposures, diet, and other social factors, may cause breast cancer risk later on in life. The National Toxicology Program, an interagency program headquartered at NIEHS, has listed six substances in its Report on Carcinogens (RoC) that cause or may cause breast cancer in humans. These include diethylstilboestrol, a synthetic form of estrogen that was used to prevent miscarriages; steroidal estrogens used for menopausal therapy; X-ray and gamma radiation; alcoholic beverages; tobacco smoking; and the sterilizing agent, ethylene oxide.
It is now clear that cancer development is not only due to endogenous or exogenous carcinogens but also their interactions with genes which are involved in carcinogen detoxification, DNA damage repair and cell signalling and cell cycle control. Because of

carcinogen exposure, sporadic cancers development may be aided by a cumulative effect of polymorphisms in these genes. Recent advances in high-throughput microarrays also have produced a treasure of information related to molecular biology of cancers. Therefore, use of microarrays to obtain epigenetic and genetic changes among cancerous tissue and non-tumor tissue. Because of relative rarity of cancer microarray data for these tumors which is often the result of small studies, so pooling this data becomes highly desirable.

Hence we carried out our GEI analysis in breast cancer. One such gene identified in breast cancer through our analysis which is environmentally susceptible is ABCB1 gene coding for p-glycoprotein. P-gp is found to be overexpressed in breast cancer cells and is major cause for drug efflux in cancers. Human P-gp is a part of a small gene family and has two isoforms. The class I isoform (MDR1/ABCB1) being a drug transporter while export of phosphatidylcholine into the bile is carried out by the class II isoform (MDR2/3/ABCB4) (Sharom, 2011; Ruetz et al., 1994). A single P-gp molecule alone can identify and transport various drugs having different chemical structures, with molecular weight ranging from 250 g/mol (cimetidine) up to 1202 g/mol (cyclosporin) (Lin et al., 2003).

Pgp is most widely studied mammalian ABC transporters but knowledge about its ability to recognize as well as transport a varied range of compounds, xenobiotics, cyclic peptides, amphipathic anticancer agents and lipids and is still not understood (Gutmann et al., 2010, Szakács et al.,2006) .Many studies are being done which aim at identification of drug-binding sites in P-glycoprotein.

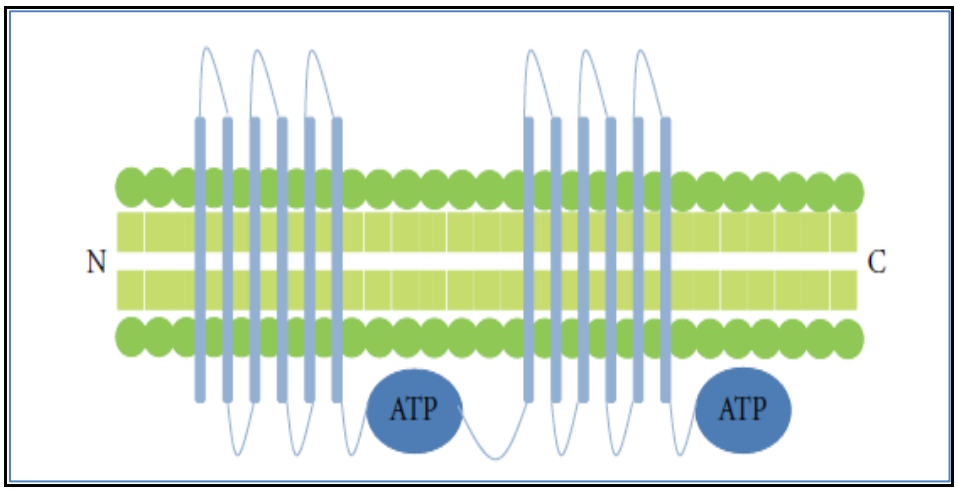


Figure 2: Structure of P-glycoprotein (PGP)—this ABC transporter consists of 12 transmembrane domains and 2 ATP binding sites.

Mouse PGP, having 87% sequence homology to human PGP while in drug-binding state, has been described recently (Aller et al., 2009). PGP structure has 12 transmembrane domains along with two ATP-binding sites (Chen et al., 1986).Human P-gp is a 170 kDa polypeptide consisting of 1280 amino acids (Bansal et al., 2009), organized in two homologous halves (Figure 2), each encompassing a transmembrane domain (TMD), which contain the drug binding sites and define the translocation pathway across the membrane, and one cytoplasmic nucleotide binding domain (NBD), which couple the energy associated with ATP binding and hydrolysis to drug transport (Chen et al. 1986, Goren et al., 2000).

Structurally diverse compounds can be extruded by P-gp out of the cells. Anticancer agents, immune suppressants, beta-adreno receptor blockers, calcium channel blockers, cardiac glycosides and several hundreds of substrates can interact with this protein in ATP dependent manner (Sharom et al., 2001; 2011).Weak substrates like less permeable drugs are extruded as well. Thus contribution in extrusion of several drugs from blood to intestinal lumen is huge. P-gp also enhances the removal of drugs out of renal tubes and hepatocytes into the surrounding luminal space. Thus, P-gp is responsible for reducing the oral bioavailability and absorption and lowering the retention time for several drugs. Also, it plays a crucial role in restricting cellular uptake of drugs while being present in BBB from blood circulation into the brain (Ma et al.,2010).P-gp is found to be overexpressed in tumour cells and is the root cause for drug efflux in cancer.

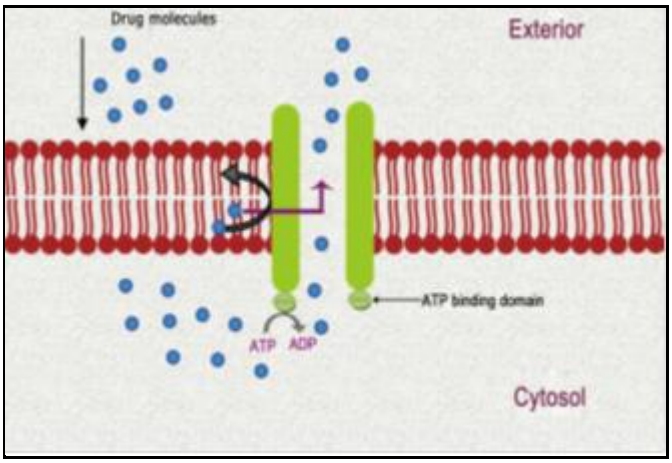


Figure 3: Drug efflux in tumour cells due to overexpression of P-glycoprotein

Therefore, the drugs administered remain futile and are unable to provide desired effect. To overcome P-gp mediated drug resistance several approaches have been taken (Mazel et al., 2001; Raub et al., 2005). Drugs localized in the plasma membrane alone are affected by P-gp. Parallel administration of inhibiting agents and cytotoxic drugs and, like cyclosporine or verapamil, can prevent P-gp facilitated extrusion and mediate the drug to reach the target site. Thus incorporation of both inhibiting agent and chemo-therapeutic agent into the carrier system can help in overcoming the resistance due to P-gp.

But presence of non-synonymous somatic mutations which have developed in response to environmental exposures may cause structural changes in the protein. Also, conformational deviations in the 3D structure of the protein are responsible for the variations in various biochemical pathway alterations as well as physiological affinities which are time dependent (Rajendran et al., 2012; 2013). In order to design an effective inhibitor we need to study the impact of such mutations by several available computational algorithms that can predict accurately uncharacterized mutations for their impact on functional and structural property of the concerned protein. Effectiveness of these algorithms in identifying the pathogenic mutations, have been stated in several research articles thus predicting the deleterious nsSNPs in correlation to their disease related property (Carvalho et al., 2007, Karchin, 2009). Underlying molecular mechanism and pathological consequences of genetic mutations can also be analysed efficiently by such computational studies.

Methodology

1. Dataset Retrieval

The copy number alterations and SNP mutations in cancer is to be collected from COSMIC database. Genes of three NIH-sponsored environmental databases were used for the cross-referencing of environmentally important genes with genes with variants in cancer. These databases were chosen because of their focus on validation of the environmentally responsive genes included in them, either through laboratory work in the databases projects themselves, or through expert curation of literature.

a) The Environmental Genome Project (EGP) located at: http://www.niehs.nih.gov/research/supported/programs/egp/
b) The 194 Comparative Toxicogenomics Database (CTD) located at: http://www.mdibl.org/research/ctd.shtml
c) Seattle 182 SNPs located at: http://pga.gs.washington.edu

2. Gene lists from environmental databases containing possible environmentally important genes in cancer is to be taken. The variant genes from the cosmic database for breast cancer and environmental databases were inputted into the GeneVenn program (Pirooznia *et al*., 2007) to assess their overlap. Common genes between the three environmental gene databases and breast cancer variant genes list were determined. These will be called GEI genes.

3. **Text-mining Search**

Pubmatrix (http://pubmatrix.grc.nia.nih.-gov), a National Institute of Health (NIH) tool which allows cross referencing of gene lists with search terms, was used to assess whether GEI genes had been previously reported as significant to breast cancer development. The Pubmatrix search uses an algorithm to match user inputted lists to gene names/symbols, etc. from abstracts, keywords and titles of studies in Medline. Thus the number of mentions a gene receives in these locations could take as a proxy for their relative importance to date for breast cancer development and for research focus in its relation.

4. Simulation for functional change in a point mutant by structure homology-based method (PolyPhen)

PolyPhen (Polymorphism Phenotyping) is an automatic tool for prediction of possible impact of an amino acid substitution on the structure and function of a human protein available at http://coot.embl.de/PolyPhen/. This prediction is based on straightforward empirical rules which are applied to the sequence, phylogenetic and structural information characterizing the substitution (Adzhubei et al., 2010). Input options for the PolyPhen server are protein sequence, SWALL database ID or accession number,

together with the sequence position of two amino acid variants. The query is submitted in the form of a protein sequence with a mutational position and two amino acid variants. Sequence-based characterization of the substitution site, profile analysis of homologous sequences, and mapping of the substitution site to known protein 3D structures are the parameters taken into account by PolyPhen server to calculate the score. It calculates position-specific independent counts (PSIC) scores for each of the two variants and then computes the PSIC scores difference between them. The higher the PSIC score difference, the higher the functional impact a particular amino acid substitution would be likely to have.

5. Support Vector Machines based Predictor of human Deleterious Single Nucleotide Polymorphisms

PhD-SNP is based a SVM-based classifier. PhD-SNP is optimized to predict if a given single point protein mutation can be classified as disease-related or as neutral polymorphism (Capriotti et al., 2006). The required inputs are:

- **Protein Sequence**: the protein sequence can be provided in raw format or giving its Swiss-Prot or uploading a text file containing the protein sequence;
- **Position**: the position number in the sequence of the residue that undergoes mutation;
- **New Residue:** if you would ask for a specific mutation please insert the symbol of the mutated residue;
- **Prediction**: choose between Sequence-Based or Sequence and Profile-Based prediction.
- **Multi SVM**: choose if the prediction is performed using 20 different SVM model from cross validation procedure or a single SVM model (fast option).

The results can be sent to e-mail address, or obtained interactively.

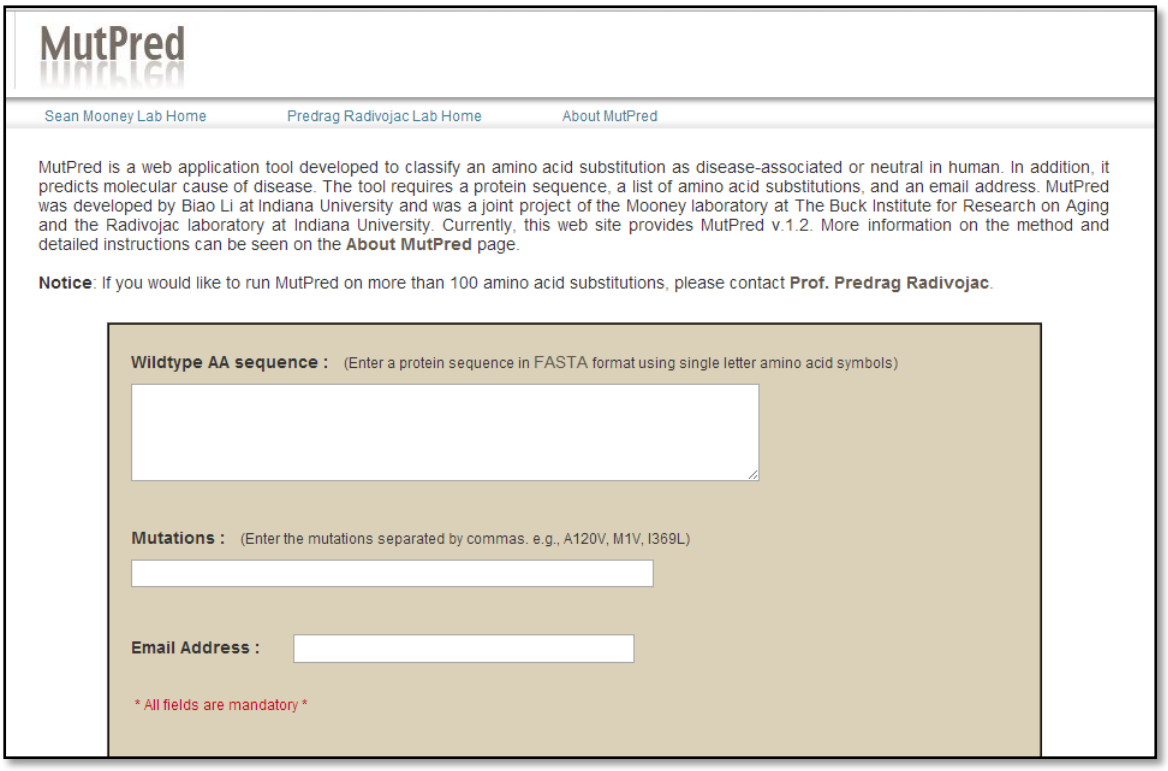
Outputs

The output consists of a table listing the number of the mutated position in the protein sequence, the wild-type residue, the new residue and if the related mutation is predicted as disease-related (**Disease**) or as neutral polymorphism (**Neutral**).

6. Molecule based prediction by MUTPRED

MutPred is a web application tool developed to classify an amino acid substitution (AAS) as disease-associated or neutral in human. In addition, it predicts molecular cause of disease/deleterious AAS (Li et al., 2009). MutPred is based upon SIFT and a gain/loss of 14 different structural and functional properties. For instance, gain of helical propensity or loss of a phosphorylation site. It was trained using the deleterious mutations from the Human Gene Mutation Database and neutral polymorphisms from Swiss-Prot. Current version of MutPred is 1.2. The update consists of replacing SIFT score by a more stable version of code that calculates evolutionary conservation. In addition, the I-mutant software was replaced by a more stable MUpro, by the Baldi group. The training data set

was updated to contain 39,218 disease-associated mutations from HGMD and 26,439 putatively neutral substitutions from Swiss-Prot.



The output of MutPred contains a general score (g), i.e., the probability that the amino acid substitution is deleterious/disease-associated, and top 5 property scores (p), where p is the P-value that certain structural and functional properties are impacted.

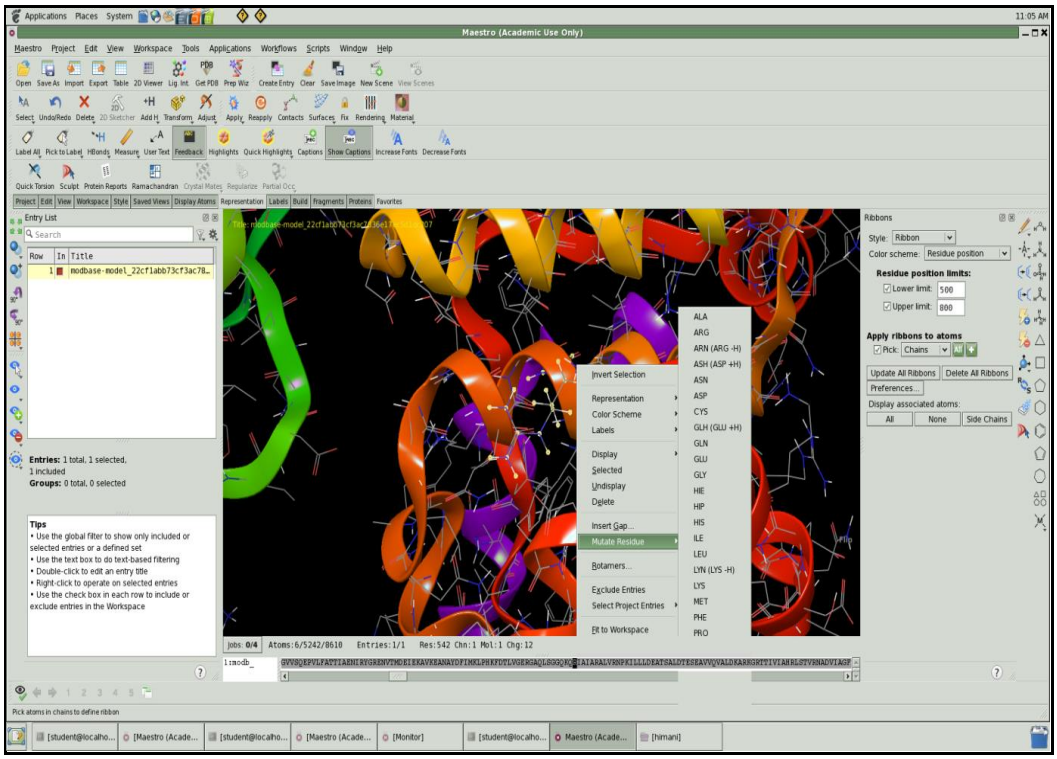
Certain combinations of high values of general scores and low values of property scores are referred to as hypotheses.

1. Scores with $g > 0.5$ and $p < 0.05$ are referred to as **actionable hypotheses**.
2. Scores with $g > 0.75$ and $p < 0.05$ are referred to as **confident hypotheses**.
3. Scores with $g > 0.75$ and $p < 0.01$ are referred to as **very confident hypotheses**.

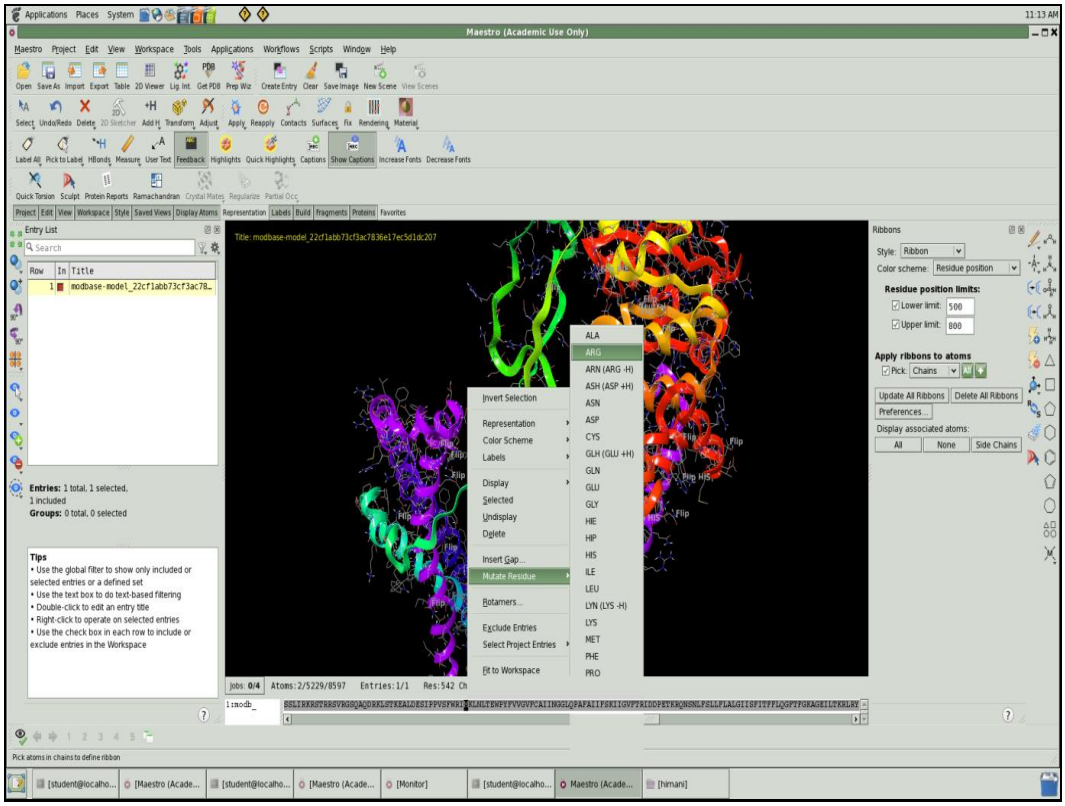
7. Modelling of the mutant protein structure

For understanding the significance of a single amino acid substitution on protein function, knowledge about 3D structure of protein is very important. We used the Uniprot to identify the protein coded by ABCB1. P08183 is the id of the human modelled protein based on the template 3G5UA for Mus musculus with sequence identity of 89%. Also mutation positions were confirmed and these positions and residues were in complete agreement with the results obtained with PolyPhen 2.0, PHD-SNP and MutPred. Mutations ware performed using Schrodinger software.

a) Mutation R538S



b) Mutation M701R



Mutations R538S and Mutation M701R were performed using mutate residue option in Scrodinger software

8. Protein preparation and Minimization.

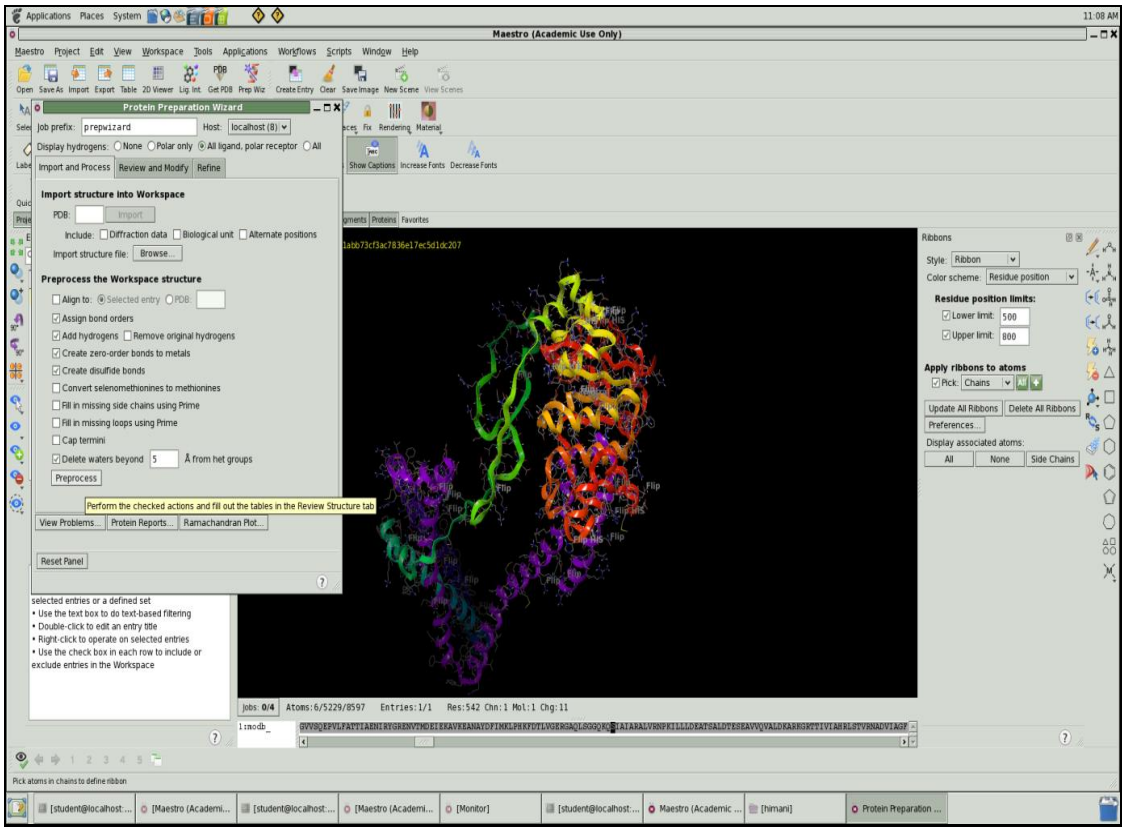
Structures imported are not usually suitable for molecular mechanics or dynamics calculations, because they have no hydrogen atoms, and include crystal water molecules. They might also have ill-defined bond orders, protonation states, formal charges, tautomerization states, disulfide bonds, and so on. All of these issues must be resolved before simulations can be performed.

Protein preparation and energy minimization for all three three-dimensional structures native and two mutants was performed using protein Preparation Wizard of Schrodinger software.

We can import structures with the Import panel, but the Protein Preparation Wizard panel provides a convenient facility for importing proteins.
1. From the Application menu, choose Protein Preparation Wizard. The Protein Preparation Wizard panel opens.

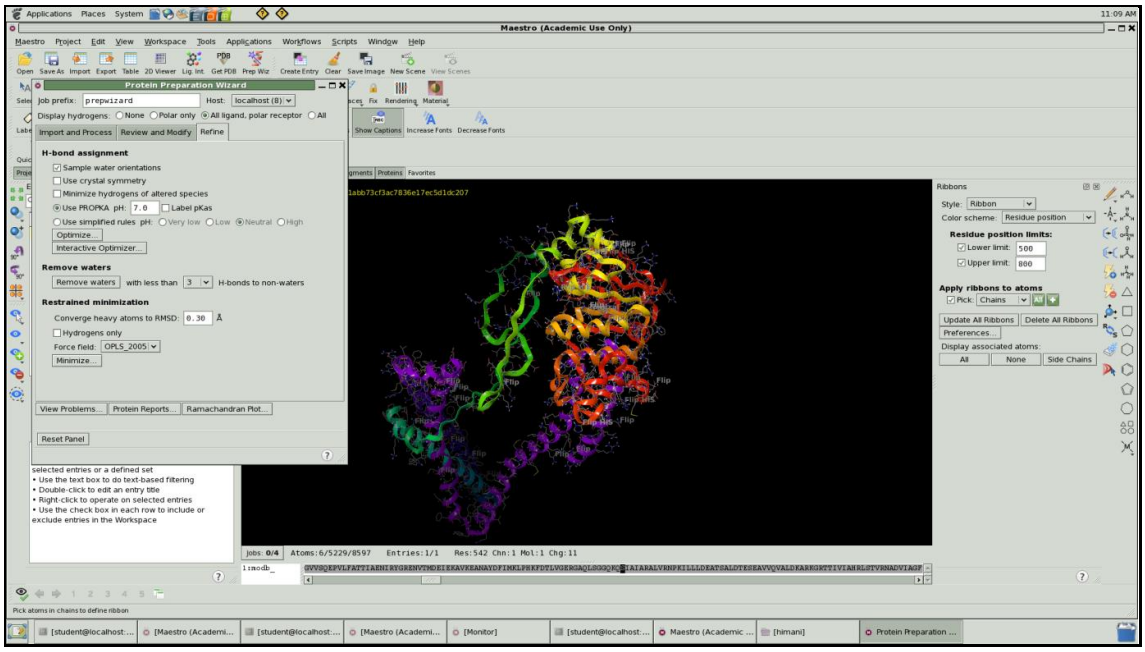
2. Select all the options in the Preprocess structure section except Selenomethionines, Fill loops, and Find side chains.

3. Click Preprocess.
The structure is preprocessed to correct the bonding information, add hydrogen atoms, cap the termini with NME and ACE, and delete water molecules. At the same time, the tables in the Chains, waters, and het groups section are filled in.



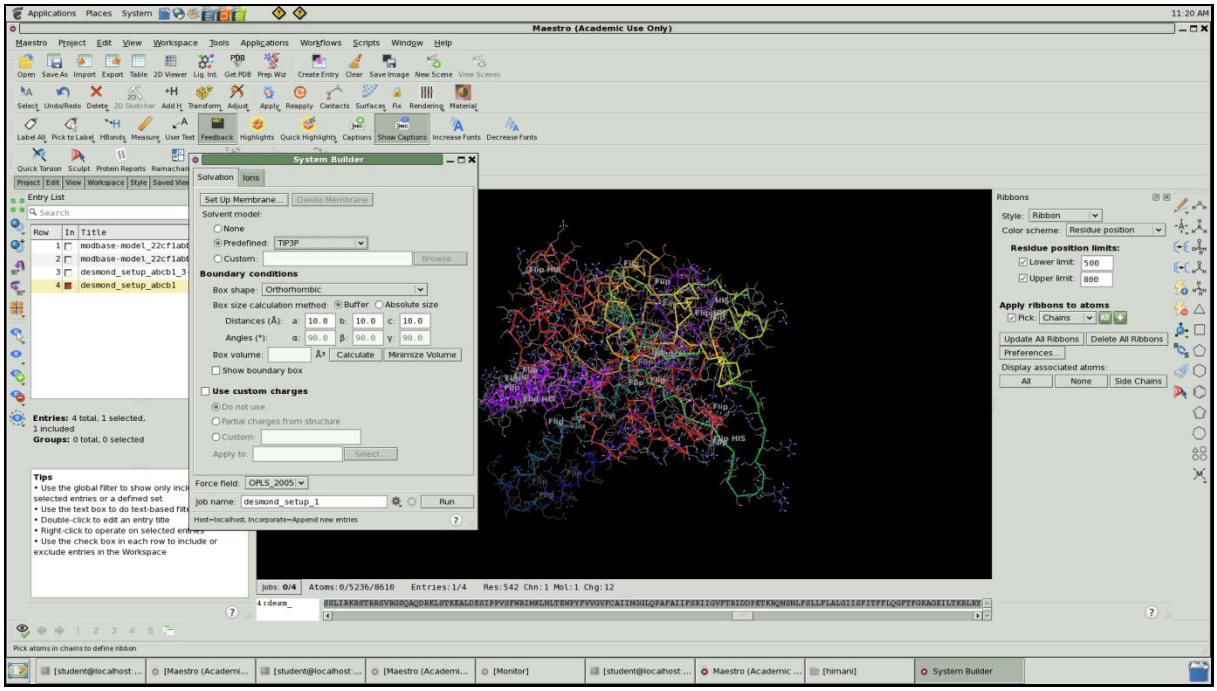
4. In the H-bond assignment section, click Optimize. The Start dialog box opens, because this step is run as a job, which is run on local host. This task optimizes the hydrogen bonding network in the protein, which includes orientation of hydroxyl and terminal amide groups in various residues.

5. In the restrained minimization tab click minimize to perform energy minimization of the structure.



9. Building model system for P-glycoprotein for MD simulation.

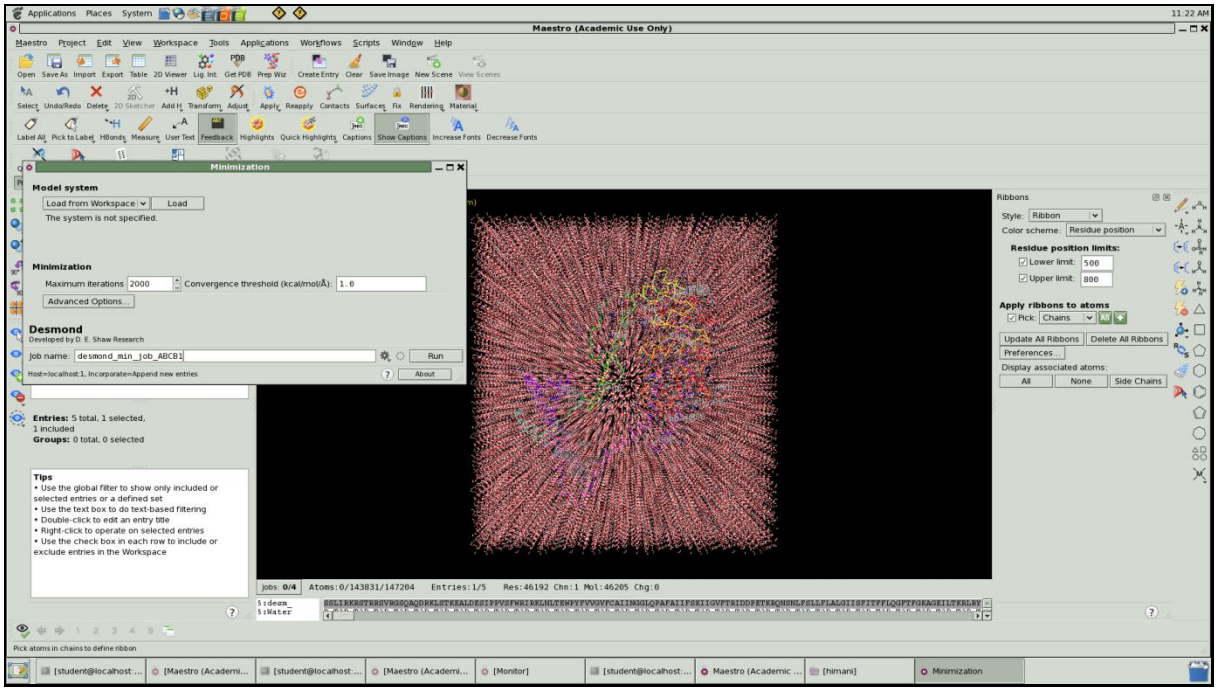
1. If the System Builder panel is not open, then from the Applications menu, choose Desmond, and then choose System Builder.
The System Builder panel opens with the Solvation tab displayed. If the System Builder panel is already open, click reset.
2. Ensure that Predefined is selected under Solvent model and we can choose the solvent model from drop down. We have chosen TIP3P in this case as it is most commonly used model. SPC is chosen in the option menu which is the default.
3. From the Box shape option menu, choose Orthorhombic. This is the shape that best fits the protein structure.
4. We have to ensure that Buffer is selected for the Box size calculation method, and that all three Distances text boxes contain 10.0.Default settings were used.
5. In the Ions tab, ensure that Neutralize is selected. The prepared structure is charged, so it needs to be neutralized with counter ions.
6. Select Add salt.



7. In the Salt concentration text box, enter 0.15.
Ions will be added to the simulation box that represents background salt at physiological conditions. By default, sodium chloride is added, but you can choose a variety of positive and negative ions for the salt.
8. Click Start. The Start dialog box is displayed.
9. Changed the job name to abcb1_setup.The job should not take more than a minute, so it can be run locally. No other changes are needed in this dialog box.
10. Click Start. The job is started, and the Monitor panel is displayed. When the job finishes, a new entry group is added to the Project Table, labelled **abcb1_setup-out**. It contains only a single entry, which includes the entire model system.

10. Running the simulation

1. In the main window choose Applications > Desmond > Minimization. The Minimization panel opens.
2. In the Model system section, ensure that Load from Workspace is chosen in the option menu, and click Load.
3. Change the job name to abcb1_min.
4. Click Start. The job starts and the Monitor panel opens. The job finishes in about a minute, and the output from the minimization is included in the Workspace. We can close or minimize the Monitor panel.
5. Close the Minimization panel.
6. In the main window choose Applications > Desmond > Molecular Dynamics. The Molecular Dynamics panel opens.
7. In the Model system section of the Molecular Dynamics panel, ensure that Load from Workspace is chosen in the option menu, and click Load.

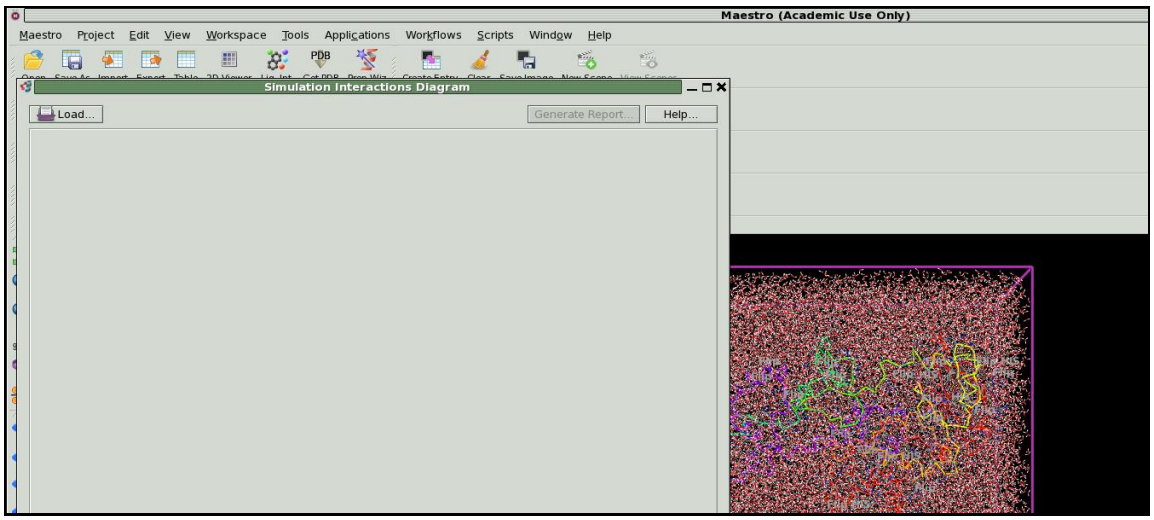


8. The controls at the top of the Simulation section allows us to specify the simulation time in ns and the recording interval in ps for the energy and for the trajectory.
9. The controls in the lower part of the Simulation section allows us to choose the ensemble class, from NVE, NVT, NPT, NPAT, and NPγT.
10. Change the job name to abcb1_md. Set the number of CPUs, and choose a host. Desmond MD simulations are CPU-intensive, and run very efficiently in parallel.
11. Click Start. The job starts and the Monitor panel opens. When the job finishes, the results are imported into the Project Table and the last structure in the simulation is displayed in the Workspace.
12. Close the Molecular Dynamics panel.

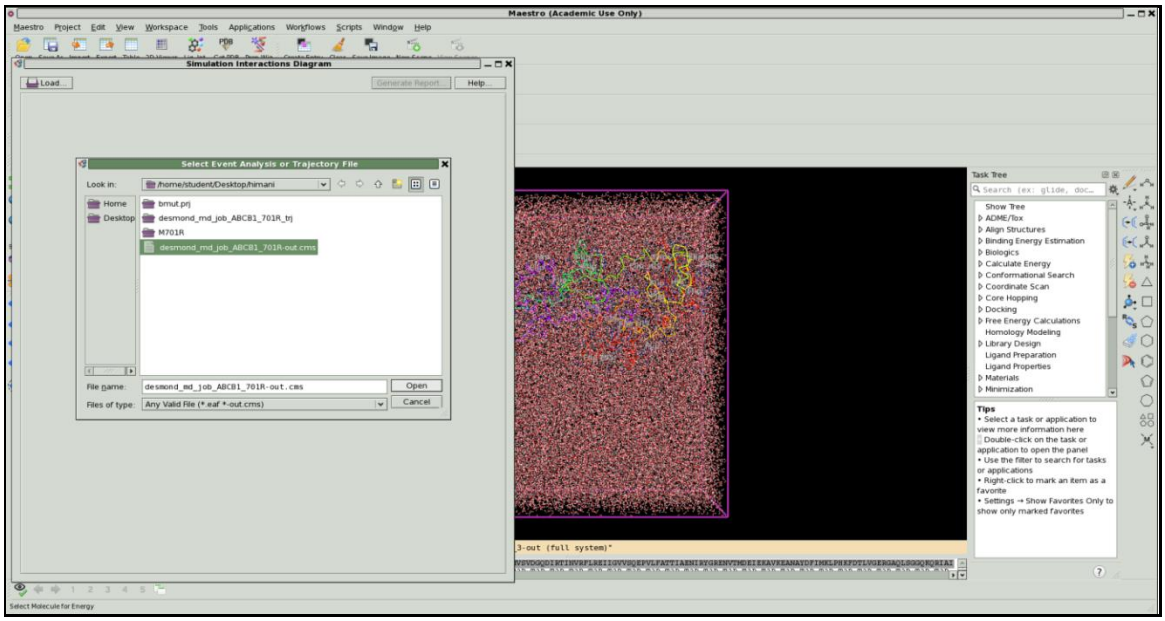
11. Simulation Analysis

The Simulation Interactions Diagram panel creates graphical displays of a variety of information about the behaviour and interactions of proteins and ligands during the course of a simulation.

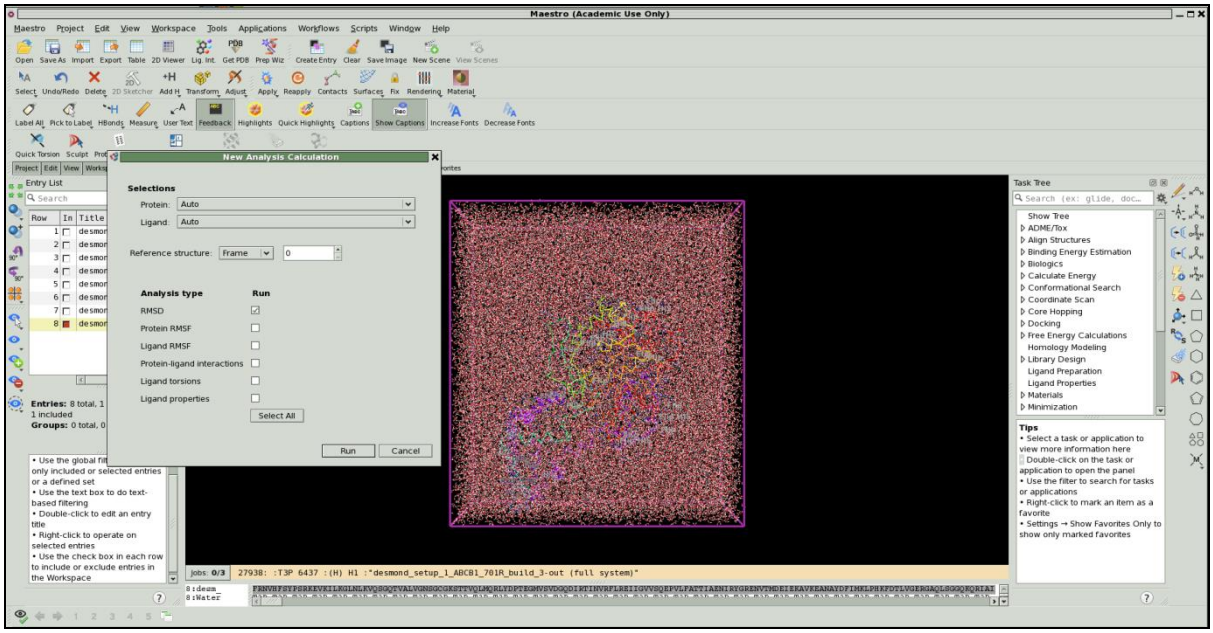
To open the Simulation Interactions Diagram panel :
• Choose Applications → Desmond → Simulation Interactions Diagram.



To generate the analysis data, click the Load button and select an output - out.cms file that has an associated trajectory in the file selector that opens.



Here we can select RMSD, RMSF in the check box and run the program.



Click the Load button and load the event analysis file (.eaf) and examine the various graphical representations of the RMSD, RMSF.

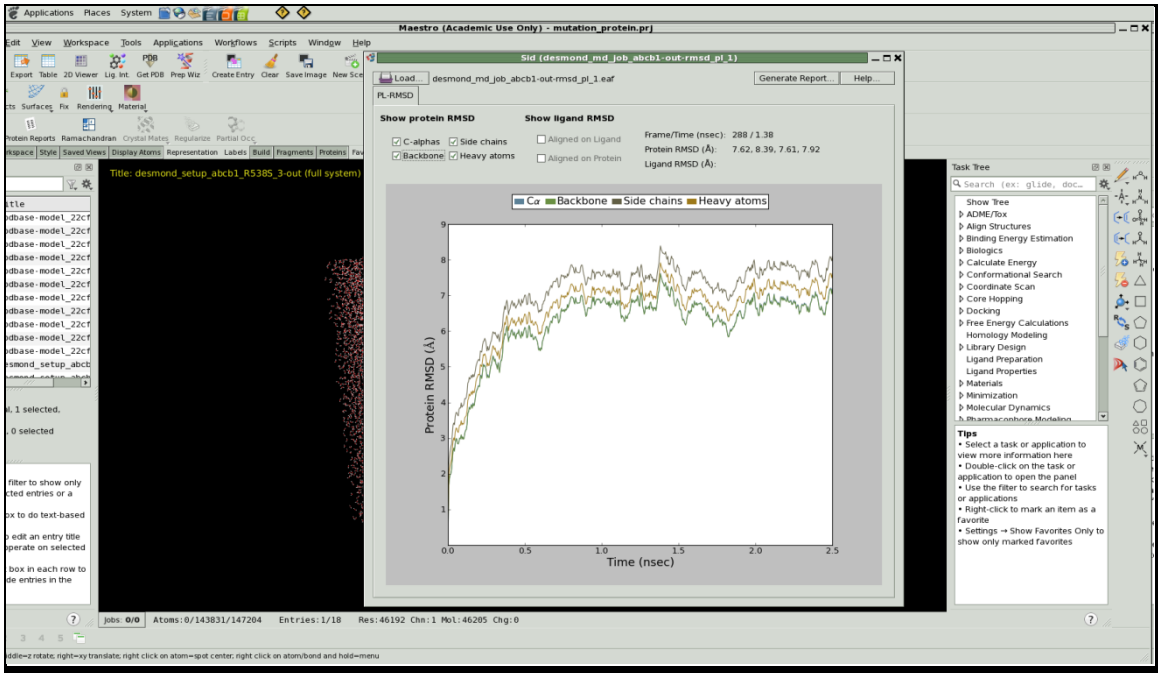
PDF file that contains all the charts in the panel, along with explanatory text can be created. We can export just the charts as images. We can also export the data to a plain text file if some further analysis is desired.

12. Protein RMSD

RMSD tab displays plots of the RMSD of selected protein with respect to a reference frame, as a function of simulation time. The RMSD is calculated after superimposing the frame for a given time step on the reference frame. The superposition depends on the choice of atom set to display, as explained below.

Atom set for which to display plots of protein RMSD values can be selected in the Show protein RMSD section. There are four choices, C-alphas, Backbone, Side chains, and Heavy atoms. The superposition is done for the chosen atom set, except that the backbone is used for superposition for side-chain measurements. Each RMSD is plotted in a different color, shown in the legend.

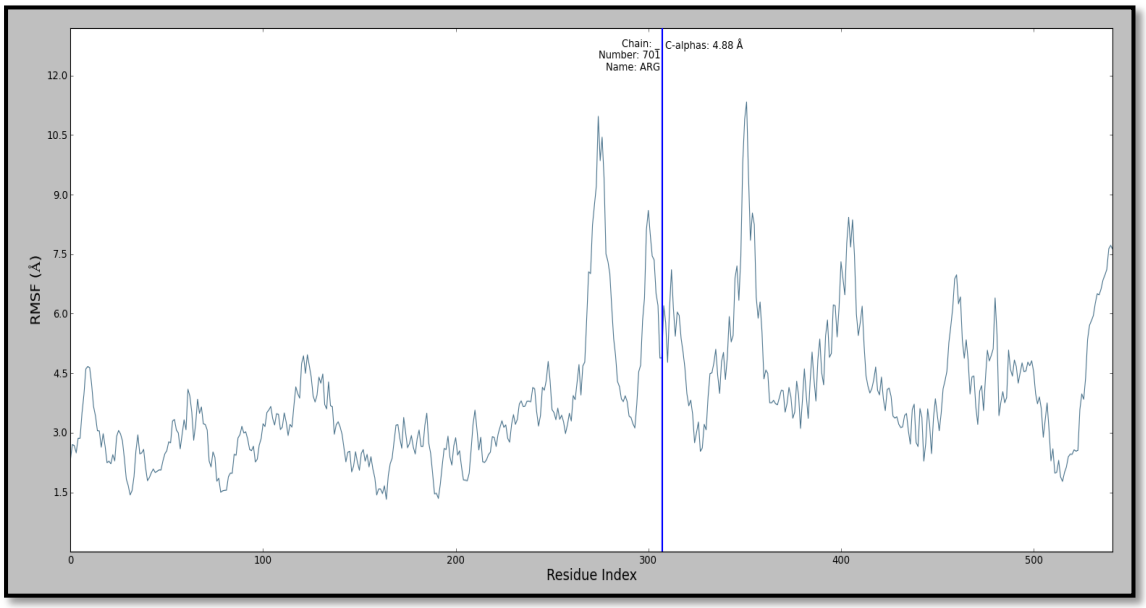
If the simulation has equilibrated, the protein RMSD should be fluctuating around some thermal average, by around 1-3 Å. If the RMSD is still increasing or decreasing, the simulation has not equilibrated, and the simulation may not be long enough for rigorous analysis. Large changes in the protein RMSD may indicate conformational changes.



13. Protein RMSF

The P-RMSF tab displays root-mean-square fluctuations (RMSF) for each residue in the protein chain. The RMSF for the residues is the time-averaged fluctuation of the square deviation of a designated set of residue atoms over the entire simulation time, after superposition on the reference frame. Peaks indicate areas of the protein that fluctuate most. These usually include the termini. Helices or strands usually fluctuate less.

To show the RMSF for different components of the protein, choose an option in the Show protein RMSF section. There are four choices, C-alphas, Backbone, Side chains, and Heavy atoms. Each RMSF is plotted in a different colour, shown in the legend. The superposition is done on the specified atom set, except for the side chains, where the protein backbone is used for superposition.



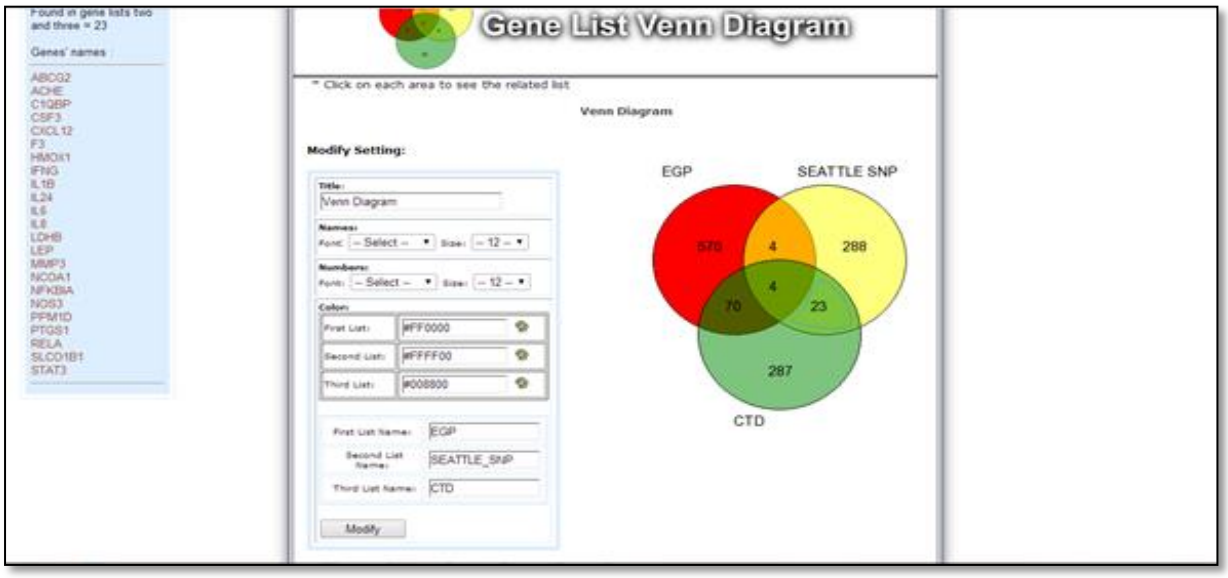
Results

1. Dataset Retrieval

The search of the environmental databases returned 648 EGP (environmentally responsive genes), 320 SSNP genes (inflammatory genes), and 397 CTD genes (Toxicogenomics genes)
In the COSMIC database, 18,736 genes were mutated from which 1066 genes were selected on basis of no. of mutated samples greater than 10.

2. Gene Overlap Analysis using Gene Venn

Gene Venn tool was used to study the overlap between all three environment databases. Little overlap was seen between them showing the three databases are mutually exclusive and all three together should be used for further analysis.



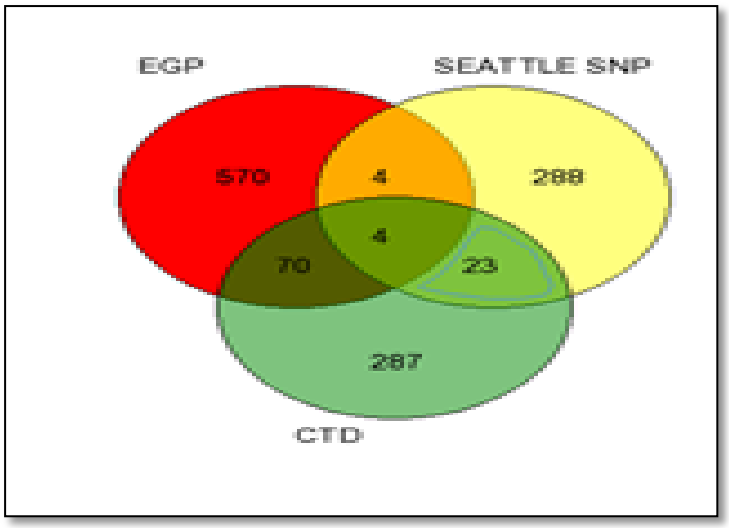
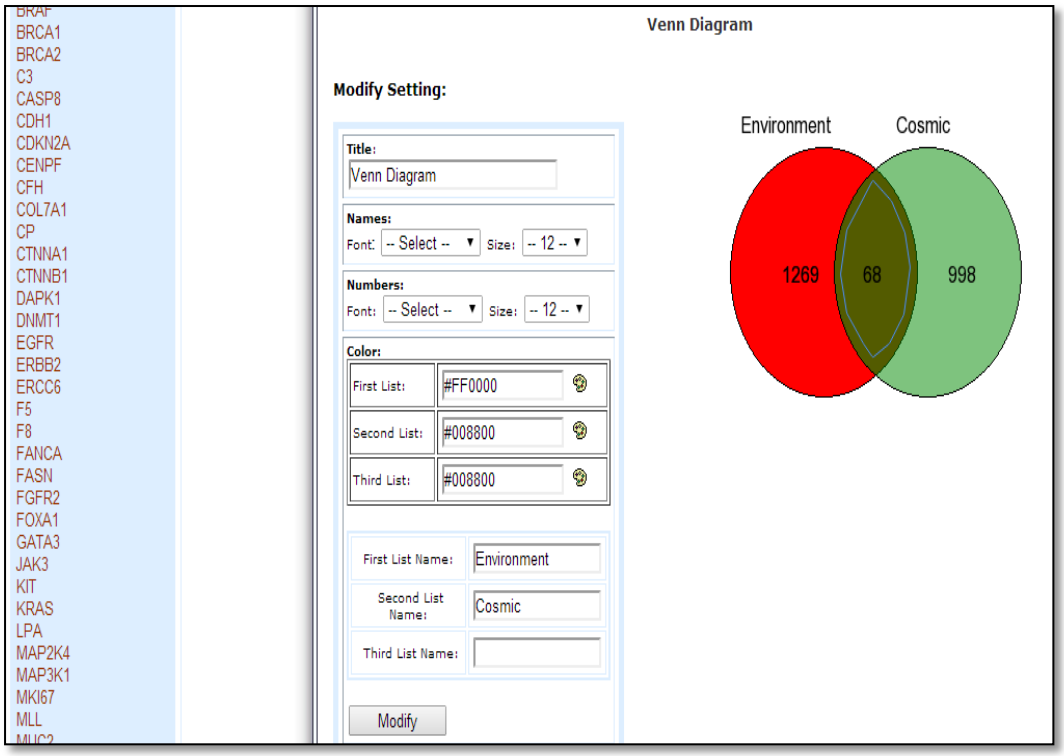


Figure 4: Little overlap of the three environmental gene datasets from Gene Venn tool indicate that all three datasets should be taken together for analysis.

3. Gene Overlap Analysis using Gene Venn

68 genes were identified as common between the environment susceptible genes and Cosmic database.



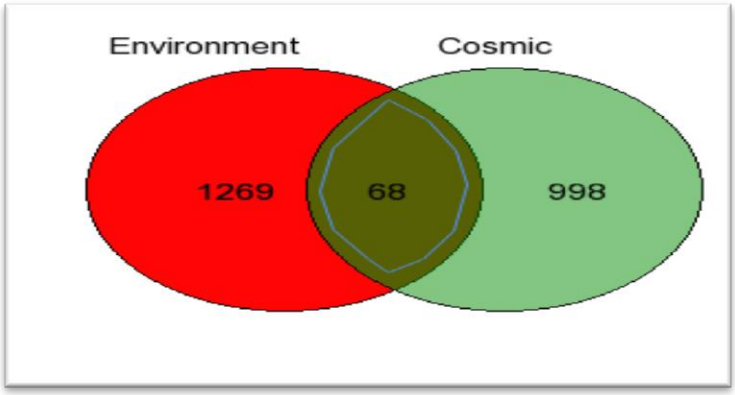


Figure 5. Overlap of the environment susceptible and breast cancer genes from cosmic. 68 genes were found to be common.

4. Pubmatrix Search

Control x Search for HIMANI GUPTA	
PubMatrix	Breast Cancer
ABCA3	4
ABCB1	1883
ABL2	1
AKT1	445
ALK	85
AR	1473
ATM	573
ATRX	1
BRAF	109
BRCA1	7613
BRCA2	5002
C3	306
CASP8	492
CDH1	252
CDKN2A	470
CENPF	3
CFH	5
COL7A1	0
CP	604
CTNNA1	50
CTNNB1	929
DAPK1	42
DNMT1	99
EGFR	2944
ERBB2	12049
ERCC6	6
F5	23
F8	12
FANCA	29
FASN	59
FGFR2	228
FOXA1	124
GATA3	122
Gene	47583
JAK3	23
KIT	766
KRAS	332
LPA	79
MAP2K4	19
MAP3K1	76
MKI67	1712
MLL	60
MUC2	64
MUC5AC	30
NCOA3	143
NCOR1	48
NELL1	0
NF1	56
NOTCH1	129
NOTCH2	30
NOTCH4	37
NRAS	21
PIK3CA	274
PLCG1	4
PMS1	9
POLE	188
PRKDC	37
PTEN	937
PTPRD	13
RB1	131
SMAD4	115
STAT4	5
SYNE1	1
TAF1	2
TBX3	26
TGFBI	4
TJP1	26
TP53	1040
VWF	41

Table 2. Pubmatrix search showing understudied and novel susceptible genes in relation to breast cancer as can be seen from number of studies carried out for them.

Pubmatrix search gave number of studies conducted per gene with respect to breast cancer which listed BRCA1, BRCA2, EGFR, MK1712, ERBB2, and ABCB1 as most studied genes with respect to breast cancer.

Among these genes ABCB1 gene has not yet been studied with respect to mutational simulation analysis and since it is major cause for failure of chemotherapies, it is crucial to study the impact of nsSNPs in P-glycoprotein structure and function so as to design better treatment regimens for breast cancer.

5. Analysis of deleterious mutation for ABCB1 gene (p-Glycoprotein)

The functional impact of nsSNPs can be assessed by evaluating the importance of the amino acids they affect. A total of 6 nsSNPs retrieved from COSMIC for our analysis. Protein sequence with mutational position and amino acid residue variants were submitted as input in PolyPhen. PolyPhen 2.0 reports a score ranging from 0 (neutral) to 1 (damaging), which represents the confidence of its internal classifier. Out of 6 nsSNPs, 5 nsSNPs were predicted to be damaging with probability score ranging from 0.989 to 1, and, the remaining 1 nsSNPs was categorized as benign.

Position	Mutation (CDS)	Mutation (Amino Acid)	Mutation ID (COSM)	Mutation Type	Polyphen	Polyphen score	PHD-SNP
521	c.1561G>C	p.D521H	COSM453510	Substitution – Missense	Probably Damaging	1	Disease,2
538	c.1614G>T	p.R538S	COSM158793	Substitution – Missense	Probably Damaging	1	Disease,4
701	c.2102T>G	p.M701R	COSM213711	Substitution – Missense	Benign	0.242	Disease,7
774	c.2320G>A	p.G774S	COSM1488745	Substitution – Missense	Probably Damaging	0.989	Neutral,1
939	c.2816G>C	p.G939A	COSM453507	Substitution – Missense	Probably Damaging	1	Neutral,6
989	c.2966G>A	p.G989E	COSM453506	Substitution – Missense	Probably Damaging	1	Neutral,0

Table 3: nsSNPs in P-gp and prediction of their functional impact and disease association by Polyphen 2.0 and PHD-SNP respectively.

We applied PhD-SNP which is based on support vector machine tool to further classify the predicted deleterious nsSNPs as disease associated. Total 6 nsSNPs were further used in PhD-SNP server, 3 of them were predicted to be disease associated. Results has been shown above in Table 3.

6. Prediction by MutPred

These 6 mutations were further analysed by MutPred tool to predict the SNP disease-association probability and probable change in the molecular mechanism in the mutant. We found R538S to be highly deleterious with general probability (g) scores of 0.958 and was predicted to have a Gain of ubiquitination at K536 (P = 0.0301) and loss of MoRF binding (P = 0.0464).Also M701R has been found to be deleterious with general probability (g) scores of .761 and was predicted with Loss of helix (P = 0.028) and Gain of methylation at K702 (P = 0.0379) and Loss of catalytic residue at M701 (P = 0.0413) (Table 4).

Mutational Analysis by MutPred

Mutation	Probability of deleterious mutation	MOLECULAR MECHANISM DISRUPTED			Top 5 features
		Actionable Hypotheses	Confident Hypotheses	Very Confident Hypotheses	
D521H	0.685				Gain of catalytic residue at L523 (P = 0.0752) Gain of disorder (P = 0.1354) Gain of MoRF binding (P = 0.1496) Gain of helix (P = 0.1736)
R538S	0.958		Gain of ubiquitination at K536 (P = 0.0301) Loss of MoRF binding (P = 0.0464)		Gain of ubiquitination at K536 (P = 0.0301) Loss of MoRF binding (P = 0.0464) Gain of disorder (P = 0.079) Loss of methylation at R543 (P = 0.0874) Loss of catalytic residue at R538 (P = 0.1086)
M701R	0.761	Loss of helix (P = 0.028)			Loss of helix (P = 0.028)

		Gain of methylation at K702 (P = 0.0379) Loss of catalytic residue at M701 (P = 0.0413)			Gain of methylation at K702 (P = 0.0379) Loss of catalytic residue at M701 (P = 0.0413) Loss of stability (P = 0.0789) Loss of ubiquitination at K702 (P = 0.1091)
G774S	0.666				Loss of helix (P = 0.2271) Gain of MoRF binding (P = 0.2495) Gain of ubiquitination at K779 (P = 0.2763) Loss of catalytic residue at F770 (P = 0.326) Loss of glycosylation at K779 (P = 0.4471)
G939A	0.884				Gain of MoRF binding (P = 0.1365) Gain of loop (P = 0.2045) Loss of stability (P = 0.3003) Loss of catalytic residue at I937 (P = 0.3067) Loss of ubiquitination at K934 (P = 0.3325)
G989E	0.864				Gain of catalytic residue at G989 (P = 0.119) Gain of relative solvent accessibility (P = 0.1259) Gain of solvent accessibility (P = 0.1903) Loss of helix (P = 0.2022) Gain of disorder (P = 0.2502)

Table 4. nsSNPs in P-gp and prediction of their structural impact by MUTPRED

From the above mutational analysis, R538S and M701R were selected for molecular dynamics simulation analysis.

7. Molecular Dynamics Simulation Analysis for R538S variant

To understand the structural and functional behaviour of the prioritized disease associated mutations, we performed molecular dynamics simulation for native and mutant p-glycoprotein.

To examine the extent to which mutation effects protein structure, RMSD values were determined for native and mutant protein structure. We calculated the RMSD for all the atoms from the initial structure, which were considered as a central criterion to measure the convergence of the protein system concerned.

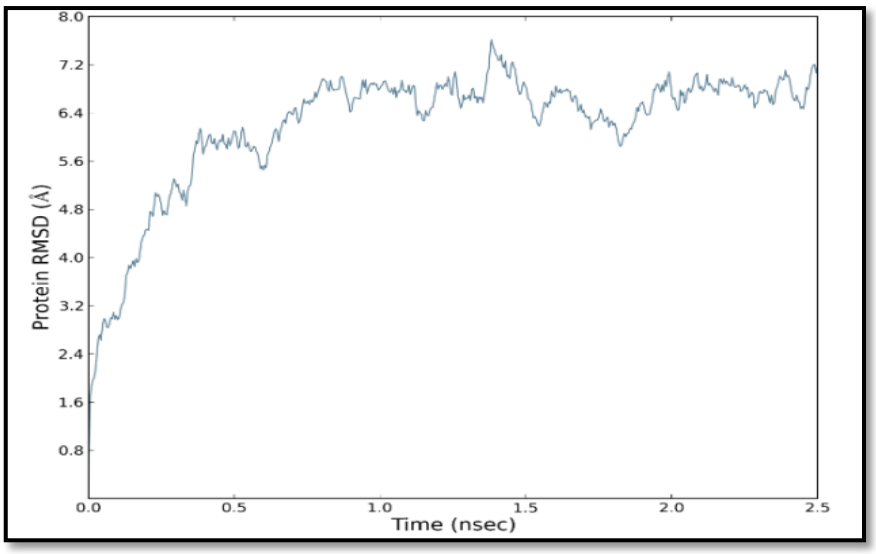


Figure 6 (a) Protein RMSD Native structure up to 2500 ps

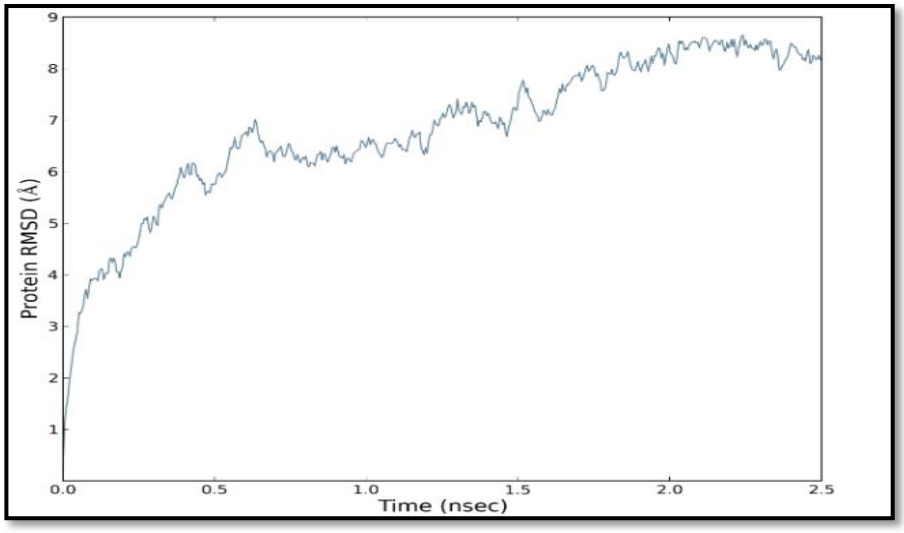


Figure 6 (b) Protein RMSD of mutant R538S structure up to 2500 ps

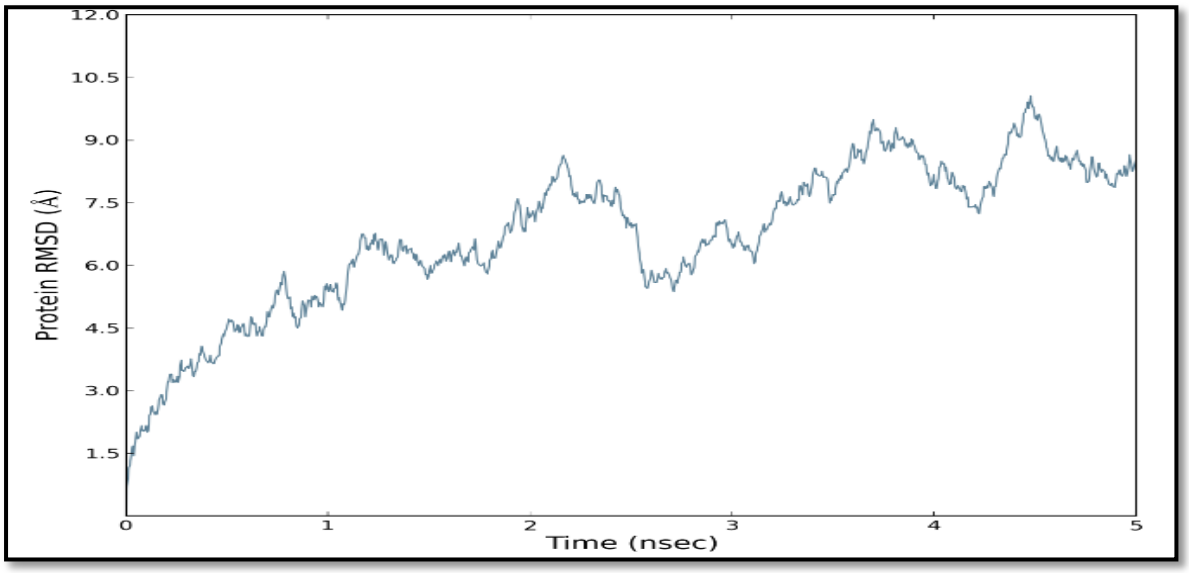


Figure 7(a) Protein RMSD native structure from 2500 to 7500 ps

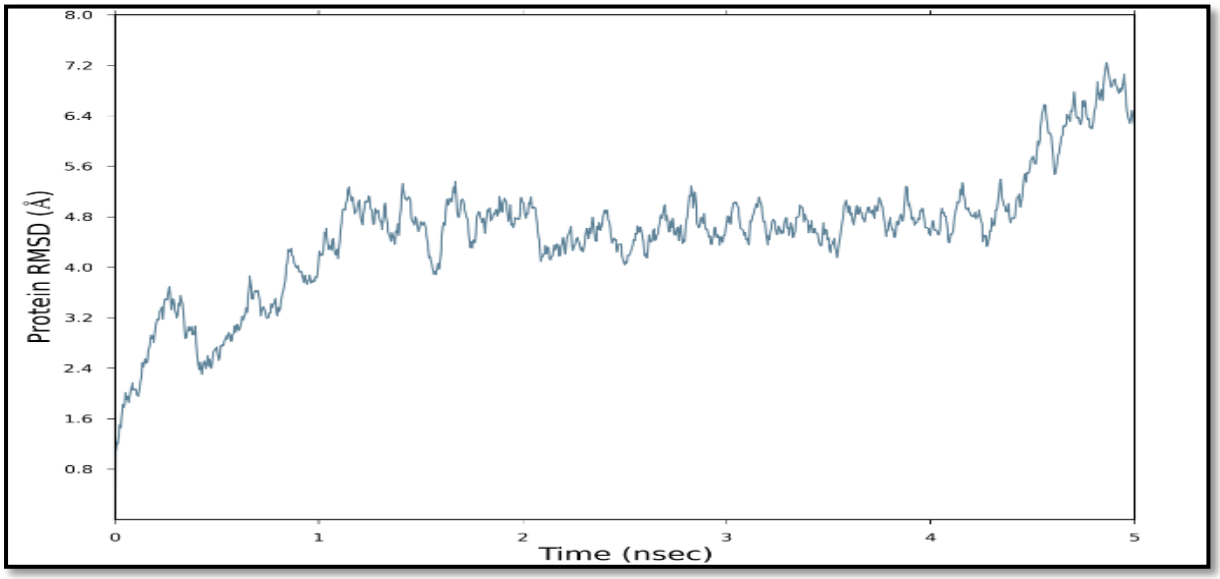


Figure 7 (b) Protein RMSD of mutant R538S structure from 2500 up to 7500 ps

In above figures 6(a) ,7(a) and figures 6(b),7(b) for native and mutant (R538S) structures respectively, show similar way of deviation till 3050 ps from their starting structure,

resulting in a backbone RMSD of ~0.14 to 0.72 nm during the simulations. After this, native structure retained the maximum deviation till the end of the simulation which is around 7500 ps resulting in the backbone RMSD of ~0.65 to ~0.96 nm while R538S mutant structure showed the minimum deviation till the end of the simulation, resulting in the backbone RMSD of ~0.38 to ~0.51nm.

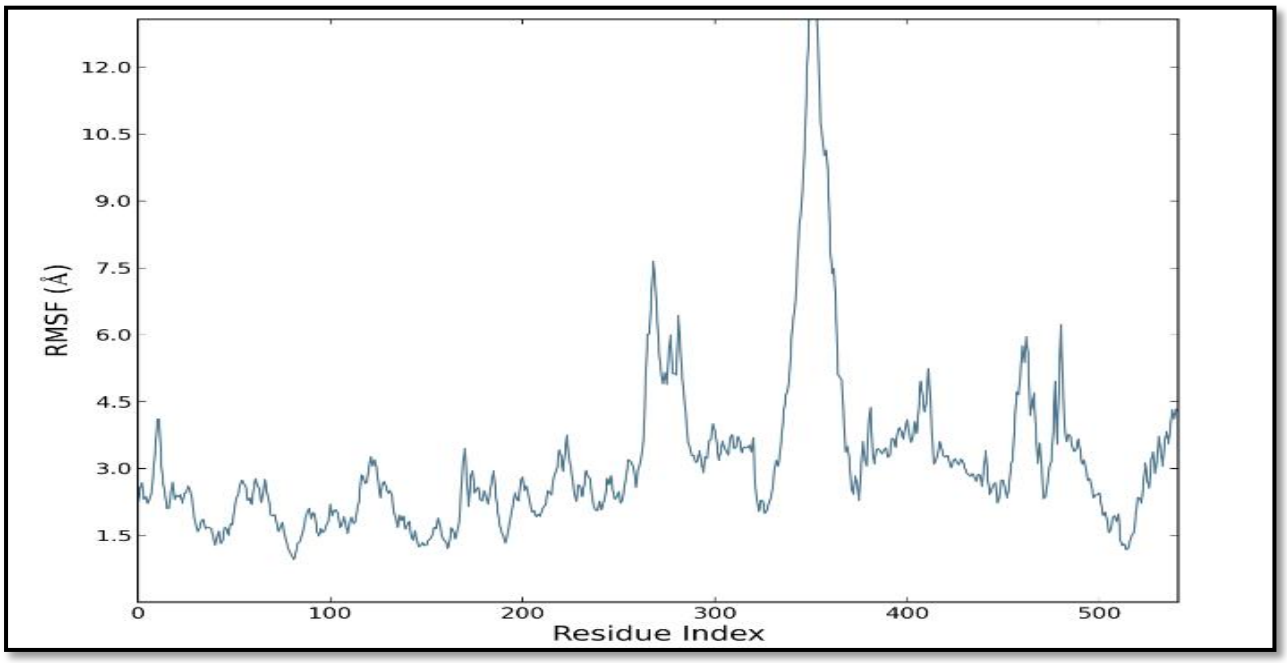


Figure 8(a) RMSF of native structure up up to 7500 ps

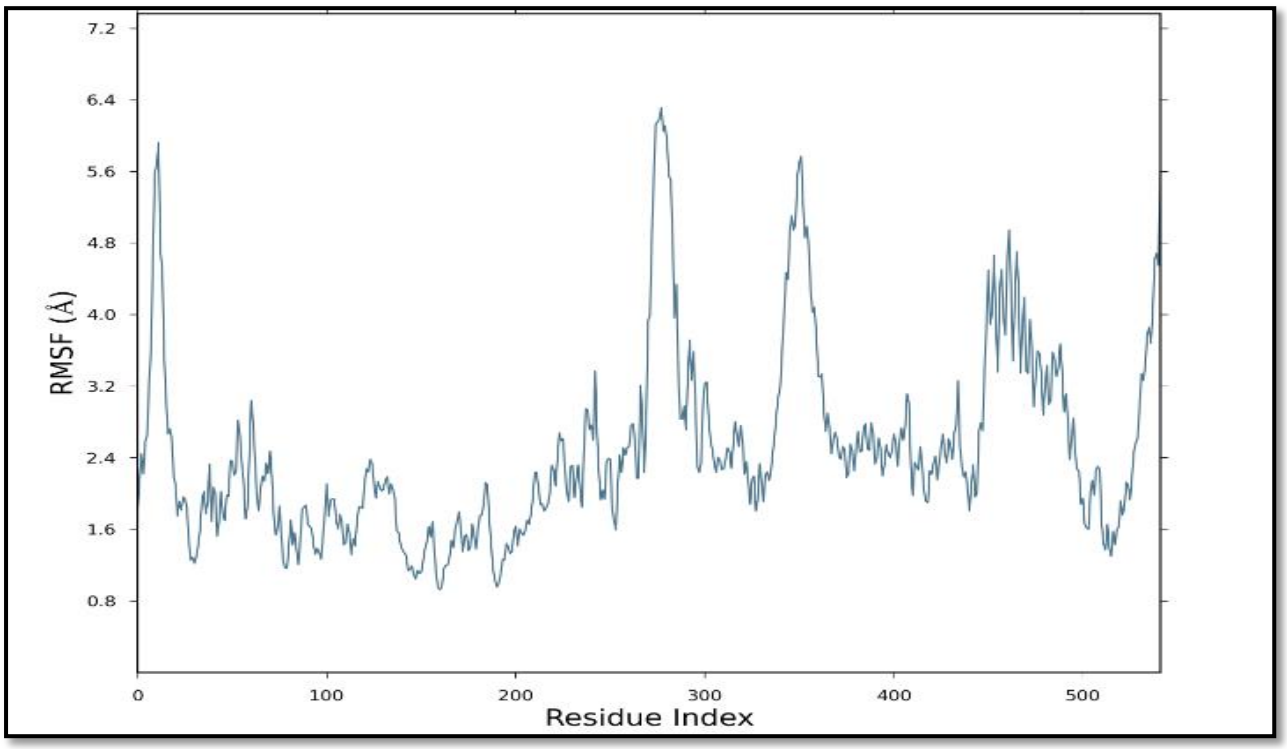


Figure 8(b) RMSF of mutant structure R538S up to 7500 ps

Through the aim of determining RMSF we predicted whether the mutation disturbs the dynamic behaviour of residues. The RMSF values of native and mutant (R358) structures were collected and shown in Figure 8(a) and 8(b) respectively. Analysis of fluctuation score depicted that the higher degree of flexibility was observed in native structure than mutant protein structure.

8 Molecular Dynamics Simulation analysis for M701R variant

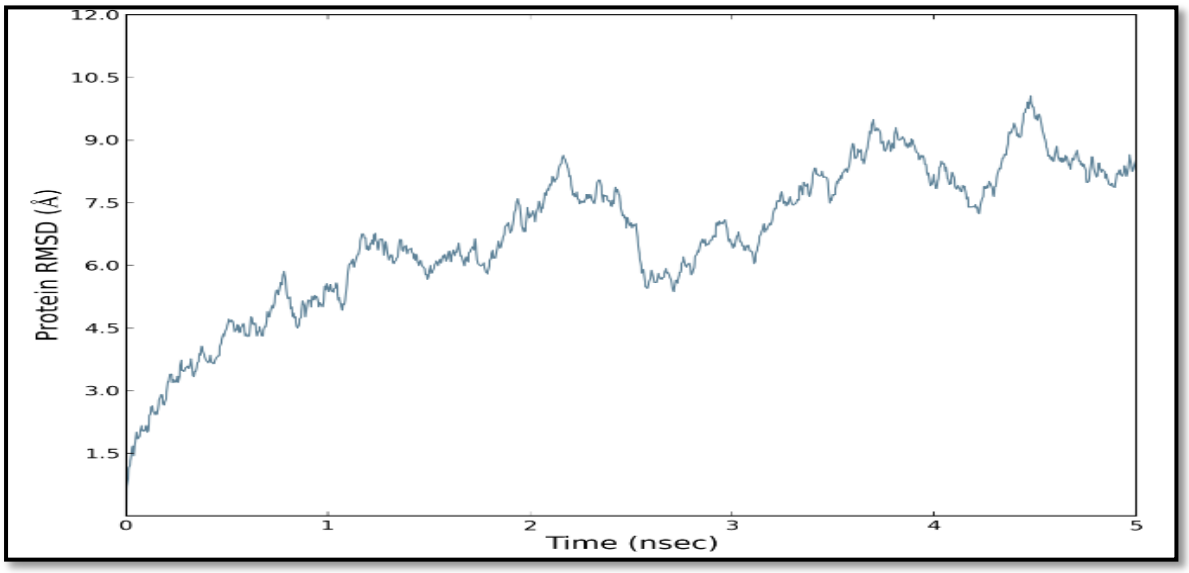


Figure 9(a) RMSD of native structure up to 5000ps

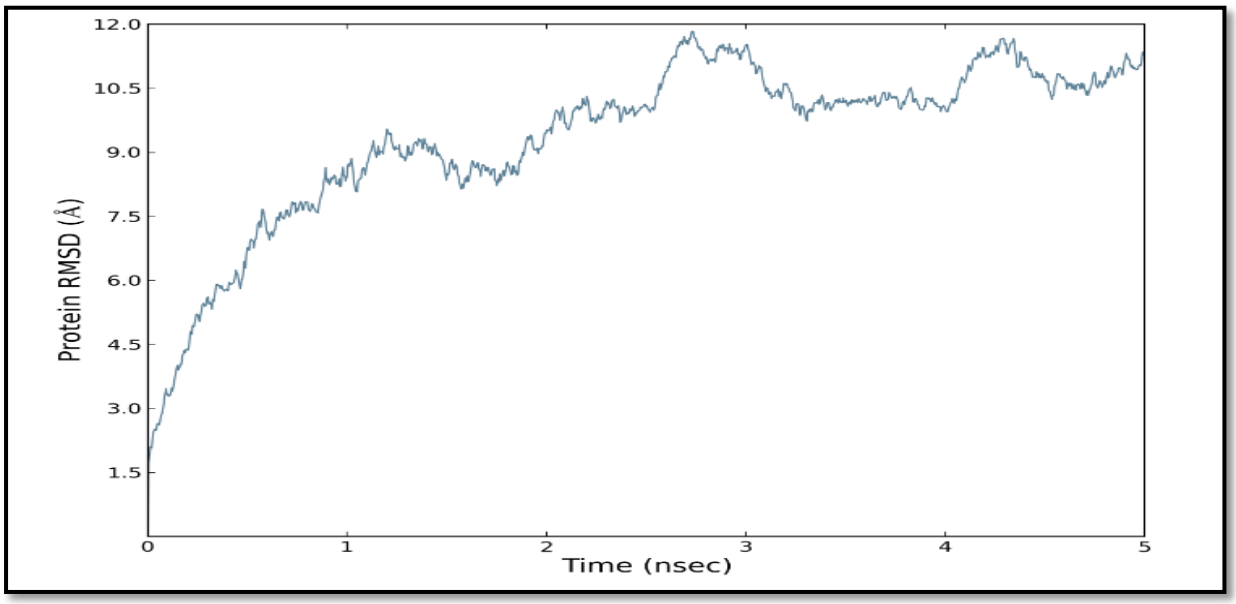


Figure 9(b) RMSD of variant structure M701R up to 5000ps

In above figure 9(a) and figure 9(b) for native and mutant (M701R) structures respectively, show similar way of deviation till 1050 ps from their starting structure, resulting in a backbone RMSD of ~0.14 to 0.85 nm during the simulations. After this, native structure retained the maximum deviation till the end of the simulation which is around 5000 ps resulting in the backbone RMSD of ~0.65 to ~0.96 nm while M701R mutant structure showed the minimum deviation till the end of the simulation, resulting in the backbone RMSD of ~0.75 to ~ 1.05 nm.

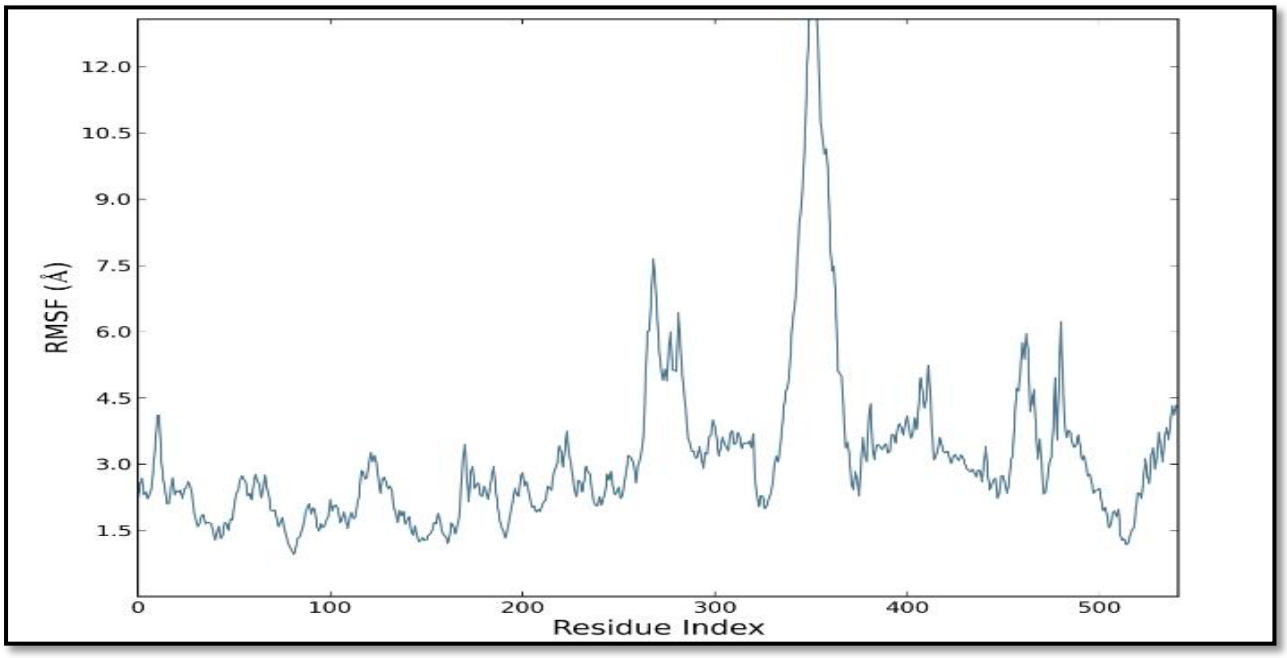


Figure 10(a) RMSF of native structure up up to 5000 ps

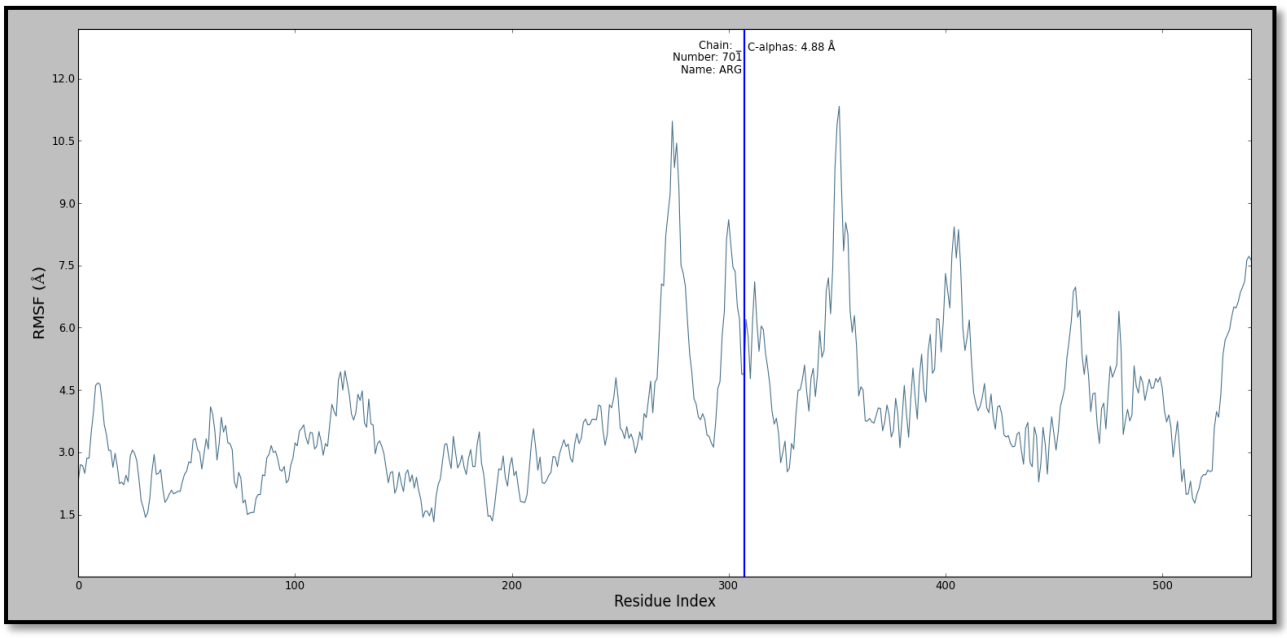


Figure 10(b) RMSF of variant structure M701R up to 5000 ps

Through the aim of determining RMSF we predicted whether the mutation disturbs the dynamic behaviour of residues. The RMSF values of native and mutant (R358) structures were collected and shown in Figure 10(a) and 10(b) respectively. Difference in fluctuation in RMSF is seen around 701th residue which is the mutated residue and can be very well seen from the highlighted figure 10(b) which suggests that mutation was responsible for the structural changes. Analysis of fluctuation score depicted that the higher degree of flexibility was observed in mutant structure than the native protein structure.

Conclusion

Of the 6 variants that were retrieved from COSMIC database, 5 variants were found to be probably damaging by PolyPhen 2.0, 3 variants were found to be disease associated by PHD-SNP and 2 variants were considered to have disease-association probability and probable change in the molecular mechanism by MutPred analysis. Two variants were selected as potentially detrimental point mutations because they were commonly found to have structural impact, disease association and damaging effect by the MutPred, PHD-SNP and PolyPhen 2.0 servers, respectively. The structures of these 2 variants were modelled and RMSD and RMSF were calculated with showed deviations and fluctuations as compared to the native structures.

Hence, it is evident that cancer results from a complex interaction of genetic and environmental risk factors. It is important to study the potential environmental and genetic risk factors combined in order to achieve a clearer understanding of the process underlying the disease and its treatment. Hence, our study showed that SNP analysis could be an ideal platform for identifying deleterious somatic and germline genetic variants that leads to various disease or treatment complications.

Discussion and Future Perspective

MDR is the major cause for the failure of chemotherapy in the cure of breast cancer. Chemotherapy is not effective, once MDR occurs, even when high doses of drugs are administered in order to overcome resistance, brings toxic effects and the drug resistance is further stimulated. Anticancer drugs resolve these problems by bypassing the resistance mechanism. Administration of substances that inhibit ABC transporters together with anticancer drugs is another method for overcoming resistance (Wu et al, 2008).

Binding sites for P-gp inhibitors are described as the efflux /drug binding site at the TMD, the allosteric residues which are intricate in communication pathways and the TP-binding site at the interface of NBD (McDevitt et al., 2007). In order to block substrate drugs from being transported away by competitive inhibition these drug-binding sites provide and efficient targets for inhibitor designing (Demmer et al., 1996, Tamai et al., 1990). Verapamil and Quinidine are some P-gp modulators that compete as substrates with the anticancer agent for transport by the pump (Thomas et al., 2003).This helps in diminishing the efflux of the anticancer drug, and this increases its intracellular concentration.

Pgp is a transport protein, consisting of two identical homologous halves, with total 1271 amino acids. It has a V shaped structure, having nucleotide binding domain (NMD) on both termini. The membrane spanning region is made up of 12 long alpha helices (H1: 44-87, H2:93-158, H3: 166-205, H4:210-249, H5: 266-317, H6: 327-365, H7: 708-736, H8: 743-794, H9: 807-847, H10: 852-904, H11:909-957 and H12: 968-1009). Each consecutive α-helix is connected through flexible loop regions which are capable of making the helices move apart while performing its functions. The remaining part of sequence forms two ATP binding domains which are α-β bundle. ATP binding domains are located from 378-626 and 1018-1271 amino acid residues (Neha et al., 2013).

Mutations in TMD or NMD domains can bring change in the structure due to which inhibitors designed to avert MDR may not bind to p-gp and hence fail to overcome MDR. So we need to study the impact of such polymorphisms prevailing in breast cancer patients in order to effectively design the inhibitors. Phenotypic effect of nsSNPs can be predicted using in silico methods which may provide a better understanding of genetic variations in disease susceptibility. Prioritizing candidate functional nsSNPs by combination of multiple algorithms served as powerful tool in our analysis. Six different somatic missense mutations in the human ABCB1 gene have been identified so far in breast cancer. Out of which two mutations namely R538S which is present in the ATP binding domain (378-626) and M701R present in the TMD domain in the vicinity of H7 alpha have been predicted to be deleterious by our analysis.

Divergence in mutant structure with native structure is due to mutation, deletions, and insertions and the deviation between the two structures is evaluated by their RMSD values which could affect stability and functional activity. To better understand how these mutations affect the structural behaviour of ABCB1, we incorporated molecular dynamic approach using Schrodinger DESMOND tool. The results that we have presented highlight the difficulty of unambiguously distinguishing native and mutant trajectories. The precise difference in the RMSD trajectories of R538S and M701 mutants, indicate the differences in the path of transition of structures from the starting conformation to their final states despite the initial structures being identical (except at the mutation sites). This information clearly speaks of the influence of amino acid substitutions on the dynamics of the protein.

Flexibility loss is observed in RMSF in case of R538S mutant. This may produce impact on the structural conformation of p-glycoprotein, which also affects its function. MutPred server predicted that there is a loss of motif binding site (P = 0.0464) which can also be observed from MD analysis where rigidity is seen which might have resulted in loss of MoRF binding. Since, NBD may also be an interesting target for the inactivation of P-gp due to the blocking of P-gp's catalytic cycle (non-competitive inhibition), as described for several flavonoid inhibitors (Conseil et al., 1998),this mutation present in the ATP binding site at the NBD interface is crucial in inhibitor designing. Also, distinct motifs contained within the NBD are of great value in *de novo* ligand design (McDevitt et al., 2007).

While in second mutant M701R, there is increased flexibility which can be seen from the RMSF. This may be because loss of helix due to point mutation which was also predicted by the MutPred server with (p= 0.028). In general, helices are mostly rigid, whereas spanning loop regions are mostly flexible (Verma et al., 2012; Ribeiro et al., 2013). Based on that, M701R mutant structure showed less helical content which might have resulted in more flexible conformation. Since this mutation lies in the cytoplasmic domain of TMD in the vicinity of alpha helix 7, which forms ligand binding cavity, it holds importance in inhibitor designing.

Therefore, it seems evident that both mutations (R538S and M701R) have damaging impact on protein structural orientation and its function. This study provides an essential insight into the underlying molecular mechanism of p-glycoprotein upon mutation and in future it may help to develop a personalized medicine for MDR in breast cancer. Further the predicted R538S and M701R mutations can be further studied by wet lab scientist to investigate the evidence of P-gp mutation in association to breast cancer and develop a potent drug target for breast cancer. Proper inhibitions of p-glycoprotein will not only increase in cellular uptake, transport, and retention of drugs, but will also help in precisely predicting their pharmacokinetics and fine tuning them for targeted drug delivery. These advancements will result bring about cost effective therapy by preventing the additional amount of drugs that used to get wasted previously by P-gp transport. Furthermore, treatment time will also get reduced because of optimal drug delivery.

References

Anderson, TW; Wright, C; Brooks, WS. (2010). The E3 Ubiquitin Ligase NARF Promotes Colony Formation in vitro and Exhibits Enhanced Expression Levels in Glioblastoma Multiforme in vivo. American Journal of Undergraduate Research. ***9***, 2-3.

A. A. Ribeiro and R. B. de Alencastro.(2013).Mixed Monte Carlo/molecular dynamics simulations of the prion protein. Journal of Molecular Graphics and Modelling.***42***, 1–6.

Ambudkar SV, Dey S, Hrycyna CA, Ramachandra M, Pastan I et al.(1999) Biochemical,cellular, and pharmacological aspects of the multidrug transporter. Annu Rev Pharmacol Toxicol. ***39***, 361-398.

Bansal T, Jaggi M, Khar RK, Talegaonkar S.(2009).Emerging significance of flavonoids as P-glycoprotein inhibitors in cancer chemotherapy. J Pharm Pharm Sci. **12(1),** 46–78.

Balmain A, Gray J and Ponder B. (2003). The Genetics and Genomics of Cancer. Nature Genet. ***33***, 238.

Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F, Furlong LI. (2011). Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. PLoS ONE. ***6***, e20284.

Bhatti P, Stewart PA, Hutchinson A, Rothman N, Linet MS, Inskip PD, et al. (2009). Lead exposure, polymorphisms in genes related to oxidative stress, and risk of adult brain tumors. Cancer Epidemiol Biomarkers Prev. ***18***, 1841–8.

B. Li, V. G. Krishnan, M. E. Mort et al. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics. ***25(21),*** 2744–2750.

Bostwick DG, Burke H B, Djakiew D, Euling S, Ho S M, Landolph J, Morrison H, Sonawane B, Shifflett T, Waters D J and Timms B. (2004). Human prostate cancer risk factors. Cancer. ***101*****(*****10***), 2371-490.

C.J. Chen, J. E. Chin, and K. Ueda. (1986). Internal duplication and homology with bacterial transport proteins in the mdr1 (P-glycoprotein) gene from multidrug-resistant human cells. Cell. ***47(3),*** 381–389.

Chung-Pu Wu, Anna Maria Calcagno, and Suresh V. Ambudkar . (2008). Reversal of ABC drug transporter-mediated multidrug resistance in cancer cells: Evaluation of current strategies. Curr Mol Pharmacol. ***1(2*****),** 93–105**.**

Conseil G, Baubichon-Cortay H, Dayan G, Jault JM, Barron D, Di Pietro A.(1998). Flavonoids: a class of modulators with bifunctional interactions at vicinal ATP- and steroid-binding sites on mouse Pglycoprotein. Proc Natl Acad Sci USA.***95***, 9831-6.

De Roos AJ, Rothman N, Brown M, Bell DA, Pittman GS, Shapiro WR, et al. (2006).Variation in genes relevant to aromatic hydrocarbon metabolism and the risk of adult brain tumors. Neuro Oncol. ***8***,145–55.

Demmer A, Dunn T, Hoof T, Kubesch P, Tummler B.(1996).Competitive inhibition of photoaffinity labelling of P-glycoprotein by anticancer drugs and modulators including S9788. Eur J Pharmacol. ***315***, 339-43.

D. Verma, D. J. Jacobs, and D. R. Livesay. (2012). Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged. PLoS Computational Biology. ***8(3),*** e1002409.

Doll R . (1975). Pott and the path to prevention. Arch. Geschwulstforsch. ***45***, 521–531.

E. Capriotti, R. Calabrese, and R. Casadio. (2006).Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. ***22 (22),*** 2729–2734.

Fan R, Wu M T, Miller D, Wain J C, Kelsey K T, Wiencke J K and Christiani D C. (2000). The p53 codon 72 polymorphism and lung cancer risk. Cancer Epidemiology. Biomarkers Prev. ***9***, 1037–1042.

F. S. Liu. (2009).Mechanisms of chemotherapeutic drug resistance in cancer therapy—a quick review," Taiwanese Journal of Obstetrics and Gynecology. ***48(3),*** 239–244.

Gutmann DAP, Ward A, Urbatsch IL, Chang G, van Veen HW. (2010). Understanding polyspecificity of multidrug ABC transporters: closing in on the gaps in ABCB1. Trends Biochem Sci. **35**, 36-42.

Goren D, Horowitz AT, Tzemach D, Tarshish M, Zalipsky S, Gabizon A. (2000).Nuclear delivery of doxorubicin via folate-targeted liposomes with bypass of multidrug-resistance efflux pump. Clin Cancer Res. 2000.***6(5),***1949–57.

Han D F, Zhou X, Hu M B, Wang C H, Xie W, Tan X D,Zheng F and Liu F. (2004). Sulfotransferase 1A1 (SULT1A1) polymorphism and breast cancer risk in Chinese women; Toxicol. Lett. ***150***, 167–177

Hanahan D and Weinberg R A. (2000). The hallmarks of cancer; Cell. ***100***, 57–70

Herbert MR, Russo JP, Yang S, Roohi J, Blaxill M, Kahler SG, et al. (2006). Autism and 580 environmental genomics. Neurotoxicology. ***27***:671–84.

Houlston R S and Peto J . (2004) .The search for low-penetrance cancer susceptibility alleles; Oncogene. ***23***, 6471–6476.

I. A. Adzhubei, S. Schmidt, L. Peshkin et al. (2010). A method and server for predicting damaging missense mutations. Nature Methods. **7 (4),** 248–249.

Jhavar S, Sarin R, Mulherkar R, Benner A, Agarwal J P and Dinshaw K .(2004). Glutathione S-transferase M1 or T1 null genotype as a risk factor for developing multiple primary neoplasms in the upper aero-digestive tract, in Indian males using tobacco. Oral Oncol. ***40*** ,84–91

J. R. Riordan, K. Deuchars, N. Kartner, N. Alon, J. Trent, and V. Ling.(1985).Amplification of P-glycoprotein genes in multidrugresistant mammalian cell lines. Nature. ***316(6031),*** 817–819, 1985.

Lin JH, Yamazaki M. (2003).Role of P-glycoprotein in parmacokinetics: clinical implications. Clin Pharmacokinet.**42(1),**59–98.

London S J, Yuan J-M, Coetzee G A, Gao Y T, Ross R K and Yu M C. (2000) .CYP1A1 I462V Genetic Polymorphism and Lung Cancer Risk in a Cohort of Men in Shanghai, China. Cancer Epidemiol. Biomarkers Prevention . ***9***, 987–991.

Martins A, Vasas A, Schelz Z, et al. (2010). Constituents of Carpobrotus edulis inhibit P-glycoprotein of MDR1-transfected mouse lymphoma cells. Anticancer Res. ***30(3),*** 829–35.

Mazel M, Clair P, Rousselle C, et al. (2001).Doxorubicin—peptide conjugates overcome multidrug resistance. Anticancer Drugs. ***12(2),***107–16.

M. A. Carvalho, S. M. Marsillac, R. Karchin et al. (2007). Determination of cancer risk associated with germ line BRCA1 missense variants by functional analysis. Cancer Research. ***67(4),***1494–1501.

Ma JD, Tsunoda SM, Bertino JS Jr, Trivedi M, Beale KK, Nafziger AN.(2010).Evaluation of in vivo P-glycoprotein phenotyping probes: a need for validation. Clin Pharmacokinet. ***49(4),*** 223–37.

McDevitt CA, Callaghan R. (2007) .How can we best use structural information on P-glycoprotein to design inhibitors? Pharmacol Ther. ***113***, 429-41.

McEachin RC, Keller BJ, Saunders EF, McInnis MG.(2008) . Modeling gene-by-environment interactio in comorbid depression with alcohol use disorders via an integrated bioinformatics approach. BioData Min. ***1(1),*** 2.

Mehdi Pirooznia, Vijayaraj Nagarajan, and Youping Deng. (2007). GeneVenn - A web application for comparing gene lists using Venn diagrams**.** Bioinformation. ***1(10),*** 420–422.

Miller D P, Liu G, De Vivo I, Lynch T J, Wain J C, Su L and Christiani D C .(2002). Combinations of the variant genotypes of GSTP1, GSTM1, and p53 are associated with an increased lung cancer risk; Cancer Res. ***62***, 2819–2823.

Neha Arora1, Shakti Sahi1 and Nagendra Singh. (2013). Structural Mapping of Inhibitors Binding Sites on P-glycoprotein: Mechanism of Inhibition of P-Glycoprotein by Herbal Isoflavones. International Journal of Biochemistry Research& Review. ***3(4),*** 421-435.

Nickels S. et al. (2013).Evidence of gene-environment interactions between common breast cancer susceptibility loci and established environmental risk factors. PLoS Genet. 9(3),e1003284.

Nishimoto I N, Pinheiro N A, Rogatto S R, Carvalho A L, de Moura R P, Caballero O L, Simpson A and Kowalski L P.(2004). Alcohol dehydrogenase 3 genotype as a risk factor for upperaerodigestive tract cancers; Arch. Otolaryngol Head Neck Surg. ***130*** ,78–82.

Park J, Chen L, Tockman M S, Elahi A and Lazarus P .(2004). The human 8-oxoguanine DNA N-glycosylase 1 (hOGG1) DNA repair enzyme and its association with lung cancer risk. Pharmacogenetics. ***14*** ,103–109.

Rajaraman P, Stewart PA, Samet JM, Schwartz BS, Linet MS, Zahm SH, et al. (2006). Lead,genetic susceptibility, and risk of adult brain tumors. Cancer Epidemiol Biomarkers Prev.***15***,2514–20.

Raub TJ. (2005).P-glycoprotein recognition of substrates and circumvention through rational drug design.Mol Pharm. ***3(1),*** 3–25.

Rieder MJ, Livingston RJ, Stanaway IB, Nickerson DA. (2008).The environmental genome project: reference polymorphisms for drug metabolism genes and genome-wide association studies. Drug Metab Rev. ***40,*** 241–61.

R. Karchin. (2009).Next generation tools for the annotation of human SNPs.Briefings in Bioinformatics. ***10(1),*** 35–52.

Ruetz S, Gros P. (1994). Phosphatidylcholine translocase: a physiological role for the mdr2 gene. Cell . ***77(7),*** 1071–81.

Searles Nielson S, Mueller BA, De Roos AJ, Viernes HM, Farin FM, Checkoway H. (2005). Risk of brain tumors in children and susceptibility to organophosphorus insecticides: The potential role of paraoxonase (PON1). Environ Health Perspect. ***113***, 909–13.

Searles Nielson S, McKean-Cowdin R, Farin M, Holly EA, Preston-Martin S, Mueller BA. (2010) .Childhood brain tumors, residential insecticide exposure, and pesticide metabolism genes. Environ Health Perspect. ***118***,144–9.

Seow A, Zhao B, Lee E J, Poh W T, Teh M, Eng P, Wang Y T,Tan W C and Lee H P .(2001). Cytochrome P4501A2 (CYP1A2) activity and lung cancer risk: a preliminary study among
Chinese women in Singapore. Carcinogenesis. ***22***, 673–677.

Sharom FJ, Liu R, Qu Q, Romsicki Y.(2001).Exploring the structure and function of the Pglycoprotein multidrug transporter using fluorescence spectroscopic tools. Seminars Cell Dev Biol. ***12(3),*** 257–65.

Sharom FJ. (2011). The P-glycoprotein multidrug transporter. Essays Biochem. ***50(1),*** 161–78.

S. G. Aller, J. Yu, A. Ward et al. (2009).Structure of P-glycoprotein reveals a molecular basis for poly-specific drug binding. Science. ***323(5922),*** 1718–1722.

Shu X O, Cai Q, Gao Y T, Wen W, Jin F and Zheng W. (2003). A population-based case-control study of the Arg399Gln polymorphism in DNA repair gene XRCC1 and risk of breast cancer; Cancer Epidemiol. Biomarkers Prev. ***12*** ,1462–1467

Srivalli KMR, Lakshmi PK. Overview of P-glycoprotein inhibitors: a rational outlook. (2012). Braz J Pharm Sci. ***48(3),*** 353–67.

Szakács G, Paterson JK, Ludwig JA, Booth-Genthe C, Gottesman MM. (2006).Targeting multidrug resistance in cancer. Nat Rev Drug Discov. ***5***, 219-234.

Tamai I, Safa AR. (1990). Competitive interaction of cyclosporins with the Vinca alkaloid-binding site of P-glycoprotein in multidrug-resistant cells. J Biol Chem. ***265***,16509-13.

Tiemersma E W, Bunschoten A, Kok F J, Glatt H, de Boer S Y and Kampman E. (2004) . Effect of SULT1A1 and NAT2 genetic polymorphism on the association between cigarette smoking and colorectal adenomas. Int. J. Cancer.**108**, 97–103.

Thomas H, Coley HM.(2003).Overcoming multidrug resistance in cancer: an update on the clinical strategy of inhibiting p-glycoprotein. Cancer Control. ***10***,159-65.

V. Rajendran and R. Sethumadhavan. (2013). Drug resistance mechanism of PncA in Mycobacterium tuberculosis. Journal of Biomolecular Structure and Dynamics. ***32(2),*** 209-21.

V. Rajendran, R. Purohit, and R. Sethumadhavan. (2012). In silico investigation of molecular mechanism of laminopathy caused by a point mutation (R482W) in lamin A/C protein. Amino Acids. ***43 (2),*** 603–615.

Varley J and Haber D A. (2003). Familial breast cancer and the hCHK21100delC mutation: assessing cancer risk; Breast Cancer Res. ***5***,123–125.

Xing D, Qi J, Miao X, Lu W, Tan W and Lin D. (2002). Polymorphisms of DNA repair genes XRCC1 and XPD and their associations with risk of esophageal squamous cell carcinoma in a Chinese population; Int. J. Cancer. ***100***, 600–605.

APPENDIX

Gene List Report – Overlapped gene list

Current Gene List: new_converted_list

Current Background: Homo sapiens

68 DAVID IDs

OFFICIAL_GENE_SYMBOL	Gene Name	Related Genes	Species
774727	collagen, type VII, alpha 1	RG	Homo sapiens
774789	antigen identified by monoclonal antibody Ki-67	RG	Homo sapiens
774934	alpha thalassemia/mental retardation syndrome X-linked (RAD54 homolog, S. cerevisiae)	RG	Homo sapiens
775913	GATA binding protein 3	RG	Homo sapiens

776365	spectrin repeat containing, nuclear envelope 1	RG	Homo sapiens
776910	fibroblast growth factor receptor 2	RG	Homo sapiens
777414	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	RG	Homo sapiens
777846	DNA (cytosine-5-)-methyltransferase 1	RG	Homo sapiens
778340	T-box 3	RG	Homo sapiens
778593	Notch homolog 1, translocation-associated (Drosophila)	RG	Homo sapiens
778806	tight junction protein 1 (zona occludens 1)	RG	Homo sapiens

779920	phosphatase and tensin homolog; phosphatase and tensin homolog pseudogene 1	RG	Homo sapiens
780460	Notch homolog 4 (Drosophila)	RG	Homo sapiens
781132	catenin (cadherin-associated protein), beta 1, 88kDa	RG	Homo sapiens
781362	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	RG	Homo sapiens
782841	ceruloplasmin (ferroxidase)	RG	Homo sapiens
783505	fatty acid synthase	RG	Homo sapiens
784487	polymerase (DNA directed), epsilon	RG	Homo sapiens

784618	protein tyrosine phosphatase, receptor type, D	RG	Homo sapiens
784930	TAF1 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 250kDa	RG	Homo sapiens
785344	mucin 5AC, oligomeric mucus/gel-forming; similar to hCG1778310	RG	Homo sapiens
790859	mitogen-activated protein kinase kinase 4	RG	Homo sapiens

791583	similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide	RG	Homo sapiens
791692	v-akt murine thymoma viral oncogene homolog 1	RG	Homo sapiens
792309	v-abl Abelson murine leukemia viral oncogene homolog 2 (arg, Abelson-related gene)	RG	Homo sapiens

793566	similar to Mast/stem cell growth factor receptor precursor (SCFR) (Proto-oncogene tyrosine-protein kinase Kit) (c-kit) (CD117 antigen); v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog	RG	Homo sapiens

Somatic mutations reported in COSMIC database

Position	Mutation (CDS)	Mutation (Amino Acid)	Mutation ID (COSM)	Count	Mutation Type
25	c.75A>G	p.K25K	COSM453513	1	Substitution - coding silent
34	c.102C>T	p.V34V	COSM453512	1	Substitution - coding silent
417	c.1251G>A	p.V417V	COSM453511	1	Substitution - coding silent
521	c.1561G>C	p.D521H	COSM453510	1	Substitution - Missense
538	c.1614G>T	p.R538S	COSM158793	1	Substitution - Missense
601	c.1803C>T	p.F601F	COSM453509	1	Substitution - coding silent
604	c.1812A>G	p.G604G	COSM453508	1	Substitution - coding silent
701	c.2102T>G	p.M701R	COSM213711	1	Substitution - Missense
774	c.2320G>A	p.G774S	COSM1488745	1	Substitution - Missense
939	c.2816G>C	p.G939A	COSM453507	1	Substitution - Missense
989	c.2966G>A	p.G989E	COSM453506	1	Substitution - Missense
1203	c.3609G>A	p.T1203T	COSM453505	1	Substitution - coding silent
?	c.703-1G>C	p.?	COSM158792	1	Unknown