

A DISSERTATION ON
In silico analysis of variants of cancer pharmacogenomic importance

SUBMITTED TO THE
DEPARTMENT OF BIOENGINEERING
FACULTY OF ENGINEERING
INTEGRAL UNIVERSITY, LUCKNOW



IN PARTIAL FULFILMENT
FOR THE
DEGREE OF MASTER OF TECHNOLOGY
IN BIOINFORMATICS

BY

Anam Tanveer

M. Tech Bioinformatics (IV Semester)

Roll No: 2001081001

UNDER THE SUPERVISION OF

Dr. Pramod Katara

Assistant Professor

Centre of Bioinformatics

University of Allahabad

Prayagraj, Uttar Pradesh



INTEGRAL UNIVERSITY, DASAULI, KURSI ROAD
LUCKNOW- 226026

DECLARATION FORM

I, **Anam Tanveer**, a student of **M Tech Bioinformatics** (II Year/IV Semester), Integral University have completed my six months dissertation work entitled “**In silico analysis of variants of cancer pharmacogenomic importance**” successfully from **Centre of Bioinformatics, University of Allahabad** under the able guidance of **Dr. Pramod Katara**.

I, hereby, affirm that the work has been done by me in all aspects. I have sincerely prepared this project report and the results reported in this study are genuine and authentic.

Anam Tanveer

Date:

Course Coordinator

Dr. Mohammad Kalim Ahmad Khan

UNIVERSITY OF ALLAHABAD

(A Central University, Government of India)

Allahabad-211002 UP-India

Dr. Pramod Katara
(Assistant Professor)
Centre of Bioinformatics
Institute of Interdisciplinary Studies



Email: pmkatara@gmail.com
Mobile: +7309465425

TO WHOM IT MAY CONCERN

This is to certify that the thesis entitled '*In silico* analysis of variants of cancer pharmacogenomic importance' has been submitted by Ms. Anam Tanveer (**Roll No. 2001081001**), in partial fulfillment of the requirements for the award of **DEGREE OF MASTER OF TECHNOLOGY IN BIOINFORMATICS** of **INTEGRAL UNIVERSITY, LUCKNOW**. She has successfully completed her six-month project, under my supervision from Feb-July, 2022 at the Centre of Bioinformatics, University of Allahabad. As per my knowledge, this work has not been previously submitted for the award of any Degree or Diploma of this or any other University or Institution and it is her original work.

A handwritten signature in blue ink, which appears to be 'P. Katara', is written over a diagonal line. Below the line, the date '18.7.2022' is written in blue ink.

Dr. Pramod Katara
Supervisor,
Centre of Bioinformatics,
University of Allahabad



INTEGRAL UNIVERSITY

Established Under the Integral University Act 2004 (U.P. Act No.9 of 2004)

Approved by University Grant Commission

Phone No.: +91(0522) 2890812, 2890730, 3296117, 6451039, Fax No.: 0522-2890809

Kursi Road, Lucknow-226026 Uttar Pradesh (INDIA)

CERTIFICATE BY INTERNAL ADVISOR

This is to certify that **Anam Tanveer** a student of **M. Tech. Bioinformatics** (II Year/IV Semester), Integral University has completed her six months dissertation work entitled “*In silico analysis of variants of cancer pharmacogenomic importance*” successfully. She has completed this work from Centre of Bioinformatics, University of Allahabad under the guidance of Dr. Pramod Katara, Assistant Professor, University of Allahabad. The dissertation was a compulsory part of her **M. Tech. Bioinformatics**

I wish her good luck and bright future.

Dr Salman Akhtar

Associate Professor

Department of Bioengineering

Faculty of Engineering



INTEGRAL UNIVERSITY

Established Under the Integral University Act 2004 (U.P. Act No.9 of 2004)

Approved by University Grant Commission

Phone No.: +91(0522) 2890812, 2890730, 3296117, 6451039, Fax No.: 0522-2890809

Kursi Road, Lucknow-226026 Uttar Pradesh (INDIA)

TO WHOM IT MAY CONCERN

This is to certify that Anam Tanveer, a student of **M. Tech. Bioinformatics** (II year/ IV Semester), Integral University has completed her six months dissertation work entitled “*In silico analysis of variants of cancer pharmacogenomic importance*” successfully. She has completed this work from Centre of Bioinformatics, University of Allahabad under the guidance of Dr. Pramod Katara. The dissertation was a compulsory part of her **M. Tech. Bioinformatics**

I wish her good luck and bright future.

Dr. (Er.) Alvina Farooqui
Head
Department of Bioengineering
Faculty of Engineering

ACKNOWLEDGEMENT

I bow in reverence to the Almighty for blessing me with strong will power, patience and confidence, which helped me in completing the present work.

The satisfaction and euphoria that accompany the accomplishment of any work would be incomplete without the mention of people who made it possible and whose consistent guidance and encouragement crown all the efforts. This project was not only a technical endeavor but also an interesting learning experience.

First of all, I would like to thank my mentor Dr. Pramod Katara for his continuous motivation, help and guidance in materializing the resources and providing accurate and adequate assistance at various stages of the work.

I would also like to thank my loving parents with whose blessings I was able to achieve my goal successfully. I would like to extend my thanks to my senior Ms. Anamika Yadav (Research scholar, University of Allahabad) for her sincere guidance and valuable suggestions without which it wouldn't have been materialized. I am thankful to my friends for having made everything possible by giving me strength and confidence to do this extraordinary work.

I gratefully acknowledge Chancellor, Prof. SW Akhtar, Pro chancellor Dr Syed Nadeem Akhtar, Vice Chancellor Prof Javed Mussarat, Registrar Dr. Haris Siddiqui and Controller of Examination, Dr Abdul Rahman Khan of Integral University Lucknow for their insight and constant encouragement. I would like to express my sincere thanks to Head Department of Bioengineering Dr Alvina Farooqui for instilling the confidence and for her constant encouragement. I am also thankful to Dr. Ashish, Assistant Professor Department of Bioengineering for constantly motivating me for better. A note of thanks and appreciation also goes to all the faculty members of the Department of Bioengineering for their support, help, and guidance.

I extend my heartfelt thanks to my Internal Advisor Salman Akhtar, Associate Professor, Department of Bioengineering for his sincere guidance and valuable suggestions at various steps of this project to bring out the best in me. I am thankful to my Course Coordinator Dr Mohammad Kalim Ahmad Khan for helping me to complete my task and answering my queries.

Thank you.

LIST OF ABBREVIATIONS

Name	Abbreviations
BReast CAncer gene:	BRCA
Single nucleotide polymorphisms:	SNP
Pharmacogenomic:	PGx
Pharmacogenomics Knowledgebase	PharmGKB
CYP2D6	Cytochrome P450 2D6
NCI	National Cancer Institute
ADMET	Absorption, distribution, metabolism, and excretion /toxicity
MDR	Multidrug resistance cancer
VDR	Variable Drug Response
ABC	ATP-binding cassette
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
COSMIC	Catalog of Somatic Mutations in Cancer
DGIdb	The Drug Gene Interaction Database
INDELs	Insertions and deletions
dbSNP	Single Nucleotide Polymorphism Database
LD	linkage disequilibrium

CONTENTS

S. No	PARTICULARS	Page No.
1	ACKNOWLEDGEMENT	i
2	ABBREVIATIONS	ii
3	LIST OF TABLES	iv
4	LIST OF FIGURES	v
5	INTRODUCTION	1-4
6	REVIEW OF LITERATURE	5-9
7	MATERIALS AND METHODS	10-16
8	RESULTS AND DISCUSSION	17-25
9	SUMMARY AND CONCLUSION	26
10	REFERENCES	27-30

LIST OF TABLES

Table No.	Particulars of the table	Page No.
1	Details of considered SNPs	11-14
2	SNPs Reported in Indian Population	18-19
3	Chromosome numbers showing significant correlations	19-20
4	Details of SNPs showing significant linkage disequilibrium (LD)	20-21
5	Top 20 SNPs with their Clinical Significance	22-23
6	Description of Clinical Significance of GSTP1 Protein	24
7	Results of energy difference calculation of wild type and mutant type protein (GSTP1)	27

LIST OF FIGURES

Figure No.	Particulars of the figure	Page No.
1	Mechanism of multi drug Resistance in cancer cells	7
2	Multiple factors contributing to variations in drug response	8
3	Concept of personalized medicine	9
4	Schematic representation of work flow	17
5	Chromosome showing correlation values greater than the reference value of 0.7	20
6	Alleles Frequencies of protein GSTP1 based on data from the 1000 genome project, phase III.	24
7	Ligand-free structure of GSTP1(PDB ID 6LLX)	25
8	Structure of GSTP1(6LLX) with complex MES and GSH	26
9	Snapshot of pocket rank one, two and three along with their score taken from prank web tool.	26
10	Wild type and mutant structure of GSTP1	27

1. INTRODUCTION

Pharmacogenomics studies how medicines interact with inherited genes. Genetic variation means that one drug can be effective for one person and toxic for the other. The term ‘pharmacogenetics’ was first coined by the German geneticist Friedrich Vogel in 1959. It is different from genetic testing as it looks for small variation within a gene instead of searching for BRCA1 and BRCA2 genes which are majorly responsible for cancer, this helps in choosing the safest and most effective drug and dose. Pharmacogenomic studies can reveal how genetic variation across individuals affects a drug’s pharmacokinetics and pharmacodynamics. It ultimately aims to improve the safety and effectiveness of drug treatment by using genetic information to inform prescription makers towards the efficacy of the particular drug for a particular type of individual. If the associations of genotypes with drug-induced phenotypes are reproducible and have large effect sizes, clinical use of such information can thus be implemented for patient’s benefit. Cancer is a leading cause of death in many developed and underdeveloped countries. Cancers are distributed worldwide due to many reasons and their occurrence vary globally across the population. Various factors contribute to differences in cancer incidence and mortality across countries including variations in age structure; prevalence of risk factors; and availability and use of preventive services, early detection tests (e.g., mammography), and high-quality treatment (mortality). Many of these factors are strongly influenced by level of development. While approximately 15% of all incident cancers worldwide are attributed to infections, the percentage is about three times higher in low to medium HDI countries than in very high-HDI countries. There is also a larger diversity in males than in females as far as cancer statistics is concerned. Hence population genetics and careful scrutiny of SNPs responsible for cancer can help researchers by giving a clear image of cancer and their direct relation with type of human races

This is therefore more beneficial in oncology because as mentioned cancer is one of the leading causes of morbidity and mortality in industrialized nations, and failed treatment is often life-threatening. Hereditary cancers represent between 5% to 10% of all the cancers and are characterized by a family history of the same, a earlier onset of the disease and a higher likelihood of primary cancers in multiple organs. They are often associated with germline alterations in oncogenes or tumor-suppressor gene (Rahner and Steinke, 2008). In 2017, 9.6 million people are estimated to have died from the various forms of cancer. Every sixth death in the world is due to cancer, making it the second leading cause of death after cardiovascular diseases. The ability to predict how a cancer patient will respond to a particular treatment

regimen is the ambitious goal of personalized oncology. The current treatment for most cancers comprises of cytotoxic chemotherapy, which is not precisely targeted to the somatic mutations that is responsible for driving the malignant transformation as such driver mutations are unknown for most patients.

One of the major achievements of the 1000 Genomes project has been the identification of numerous novel SNPs across different populations (Abecasis GR *et al.*, 2010). Every individual carry two copies of each gene and copies of a specific gene present within a population may not have identical nucleotide sequences. The distribution of SNPs is known to be non-random across the genome (Chuang, 2004). These single nucleotide changes are well scattered throughout the genome of all species and forms the basis for human diversity (Choudhury, *et al.*, 2014). SNP occur in humans every 300–2000 base pairs along the genome. In fact, they may occur at any nucleotide and those that are relatively common are of interest to the scientists and researchers. They occur almost once in every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNPs in a person's genome. These variations are found in at least 1 percent of the population (Brody, 2016). SNP is defined as a genomic locus where two or more alternative bases occur with appreciable frequency (>1%). Scientists have found more than 600 million SNPs in populations around the world. The vast majority of SNPs are functionally silent, occurring in non-coding or non-regulatory regions of the genome. However, some of the SNPs lead to altered protein structure or expression. These biologically functional SNPs are considered the essence and substrate of human diversity in both health and disease. Once identified this SNP-based ‘genetic profile’, can be viewed as a ‘fingerprint’, useful in defining the risk of an individual’s susceptibility to various illnesses and response to drugs. SNPs are currently the marker of choice due to their large numbers in virtually all human populations. From a clinical perspective, SNPs are supposedly potential ‘diagnostic and therapeutic biomarkers’ in many types of cancer. The location of these biomarker are of utmost importance in terms of prediction of functional significance, genetic mapping and population genetics (Shen *et al* , August 2009).The identification and study of SNPs in specific genes has provided useful confirmation of hypothesized models for gene and genome dynamics. Common population-specific SNPs are non-randomly distributed throughout the genome and are significantly associated with recombination hotspots (Alwi, 2005).

The pharmacogenetic study also tends to eliminate the high risk of pharmacokinetics as a prodrug must be metabolized, or bioactivated, to generate pharmacologic effects if it doesn't there will be a decreased drug action and individuals who are genetically extensive metabolizers may display the same pharmacologic outcome as poor metabolizers if an interacting drug is administered. In the absence of the metabolic pathway much higher concentrations of active parent drug can accumulate and cause serious toxic effects (Roden *et al.*, 2012).

The incidence of functionally-important CYP alleles can vary ancestrally, like the poor metabolizers with absence of CYP2D6 function are found in 5–10% of European and African populations, but are less common in Asian subjects. Contrastingly CYP2C19 poor metabolizers are commoner in Asian subjects compared to the other two major ancestry groups, and the frequency of the CYP3A5*3 variant is much higher in Caucasians (0.85) compared to African Americans (0.55), which correlates with higher hepatic CYP3A5 expression in African American subjects (Kuehl *et al.*, 2001).

The technological progress of following technologies like proteomics, genomics, epigenomics, pharmacogenomics, and metabolomics has allowed the concept of personalized medicine to become a clinical reality. Genomics has provided DNA sequences for a tremendous number of bacteria, viruses, and yeasts, as well as humans and a number of model higher organisms (Witzmann and Raymond) and Transcriptomics is a branch of Functional Genomics which is an approach that enables the analysis of gene expression through the detection and relative quantitation of individual messenger RNAs (mRNAs). Proteomics essentially is a subdiscipline of Functional Genomics which measures the qualitative and quantitative changes in protein content of a cell or tissue in response to treatment or disease and determines protein–protein and protein–ligand interactions whereas Metabolomics is an emerging field focused on comprehensive profiling of metabolites in a sample, whether intracellular or from circulating biofluids whose applications are also emerging in areas such as tumor staging and assessment of treatment efficacy. Therefore, precision medicine comprises of two different approaches one being stratified and second being the personalized medicine, which remains consistent in testing new drug therapies in groups of patients with specific molecular alterations and determining each patient's specific response to the treatment in order to get conclusions at population level.

In our study we have extracted the cancer related SNPs from the PharmGKB Database and performed a thorough analysis and annotation of the reported SNPs in the World as well as Indian Population using the In silico approaches with the help of various online tools and databases such as Ensemble, IGVB browser, SNP-nexus, prank web tool, UCF Chimera, etc., available to ultimately detect the SNP markers for the cancer specific pharmacogenomic importance at genetic level and by performing the site directed mutagenesis using the visualization tool study the effect of non-synonymous SNP on protein stability and activity.

OBJECTIVES:

1. To find SNP markers for Indian/World specific population.
2. Mining and annotation analysis of cancer related SNPs reported in World/Indian population and its role in pharmacogenomics.
3. Contribute to applications directly related to Personalized Medicine, drug therapeutics i.e., predictive biomarkers for patient's stratification and dose selection.

2. REVIEW OF LITERATURE

Cancer is a disease of the genome, characterized by a genomic instability in which numerous point mutations accumulate and structural alterations occur in the process of tumour progression (Zhang, 2020). Cancers are caused by mutations that may be inherited, induced by environmental factors, or result from DNA replication errors, ageing being the main risk for the same (Tomasetti *et al.*, 2017). One of the first studies involving the document interactions between tumours and their microenvironment was performed in 1863 by Rudolph Virchow who observed that leukocyte infiltration characterizes solid tumours (Schmidt *et al.*, 2006). Cancer research until the 1980s was dominated by a tumour-centric view suggesting that mutations in oncogenes and tumour suppressor genes were adequate to determine carcinogenesis and cancer progression (Vogelstein & Kinzler, 1993). It is now widely recognized that in response to signals derived from tumour cells, the Tumour microenvironment actively influences the progression of cancer (Maman & Witz, 2018). Effective immune responses could either suppress the malignant cells or diminish their phenotypes and functions. Also, the cancer cells have evolved multiple mechanisms, such as defects in antigen presentation machinery, the upregulation of negative regulatory pathways, and the recruitment of immunosuppressive cell populations to escape immune scrutiny leading to impeded effector function of immune cells and the annulment of antitumor immune responses (Matsushita, *et al.*, 2012). A cancerous site is actually a chaotic place where genetic mutations occur in multiple steps, producing strains of cells that vary in their capabilities. Some mutations are lethal for the cell, while some confer characteristics that enable further misbehaviour such as weight loss and decreased resistance to infections in humans.

Cancer is ranked as the second leading cause of death in 91 of 172 countries and is third to fourth in an additional 22 countries (Ferlay *et al.*, 2018). Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020, or nearly one in six deaths the most common among them being the breast with 2.26 million cases, lung with 2.21 million cases, colon and rectum along with 1.93 million cases and prostate cancers with approx. 1.41 million cases (Ferlay *et al.*, 2021). It is also the second and fourth leading cause of adult death in urban and rural India, respectively. By 2040, the number of new cancer cases per year is expected to rise to 29.5 million and the number of cancer-related deaths to 16.4 million (Source: International Agency for Research on Cancer). Statistics from NCI's Surveillance, Epidemiology, and End Results Program include the information specific to racial and ethnic populations as well as populations defined by age, gender, and geography. As a matter of fact

Blacks/ African Americans have higher mortality rates than all other racial/ethnic groups in United States including colorectal, lung, and cervical, breast and other cancers (NIH-National Cancer Institute).

2.1 *Multidrug resistance cancer*

Traditionally, the absorption, distribution, metabolism, excretion and/or toxicity (ADMET) of a drug were thought to be governed by the physicochemical properties of the molecule, protein binding and/or biotransformation (Lipinski *et al.*, 2001). However *Multidrug resistance cancer* comes onto play which is a term used to describe the phenomenon characterized by the ability of drug resistant tumours to exhibit simultaneous resistance to a number of structurally and functionally independent chemotherapeutic agents (Krishna *et al.*, 2000). Multi-drug resistance (MDR) in the cancer chemotherapy has been pointed out as the ability of cancer cells to survive against a wide range of anti-cancer drugs. MDR mechanism (figure 1) may be developed by increased release of the drug outside the cells so the drug absorption is reduced in these cells (Zahreddine *et al.*, 2013). A number of mechanisms have been described to explain the phenomena of MDR in mammalian cells such as non-cellular resistance mechanism typically associated with solid tumours arising as a consequence of *in vivo* tumour growth and cellular based resistance mechanisms characterized in terms of alterations in the biochemistry of the malignant cells which are further subdivided as non-classical MDR phenotypes and transport based classical MDR phenotype (Fan *et al.*, 1994). This is the major reason why anticancer drugs fail to kill cancer cells. Drugs are usually given systemically and are therefore subject to variations in absorption, metabolism and delivery to target tissues that can be specific to individual patients (Szakács *et al.*, 1984).

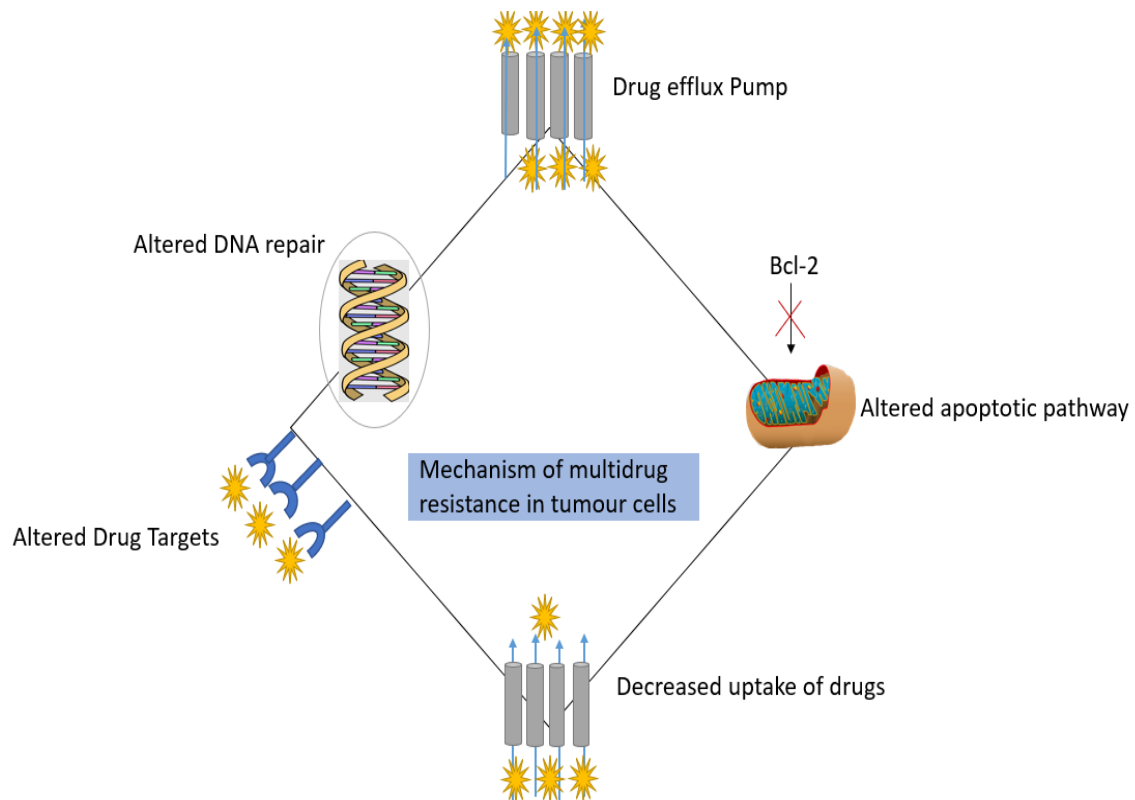


Figure 1: Mechanism of multi drug Resistance in cancer cells

2.2 Variable drug response against cancer medicine

Common cancer treatments include surgery, radiation therapy, chemotherapy, combination therapy and laser therapy and also selective therapies which are based on the better conception of the biology and molecular genetics in the tumor progression used for the promising treatments (Longley *et al.*, 2005). Currently it is estimated that around 90% of failures in the chemotherapy are during the invasion and metastasis of cancers related to drug resistance (Mansoori *et al.*, 2017). Some methods of drug resistance are disease-specific, while others, such as drug efflux, which is observed in microbes and human drug-resistant cancers, are evolutionarily conserved. There are different Intrinsic and extrinsic factors in drug resistance such as Intra-tumor heterogeneity which means Genomic instability such as mutation, gene amplifications, deletions, chromosomal rearrangements, transposition of the genetic elements, translocations and microRNA alteration etc. generates a great level of intercellular genetic heterogeneity in cancer. These factors change, increase, or diminish gene products which directly are involved in the generation of drug resistance and poor prognosis (Mansoori *et al.*, 2017). There is growing evidence that supports the important role of tumor microenvironment in drug resistance discussion as the main reason for the relapse and incurability of various cancers. Moreover, growth factor (GF), cytokines produced in the tumor microenvironment

provide additional signals for tumor cell growth and survival. One of the most studied mechanisms of cancer drug resistance involves reducing drug accumulation by enhancing efflux in which members of the ATP-binding cassette (ABC) transporter family proteins play a major role. Pharmacogenomics promises to explicate the effect of genetic inheritance on the individual variation of drug response and toxicity and has great potential to improve cancer treatment outcomes by either reducing toxicity or increasing efficacy. Various genetic factors such as genetic polymorphism in drug metabolism enzymes, drug targets, drug transporters and are collectively responsible for variable responses and tolerability of cancer chemotherapy.

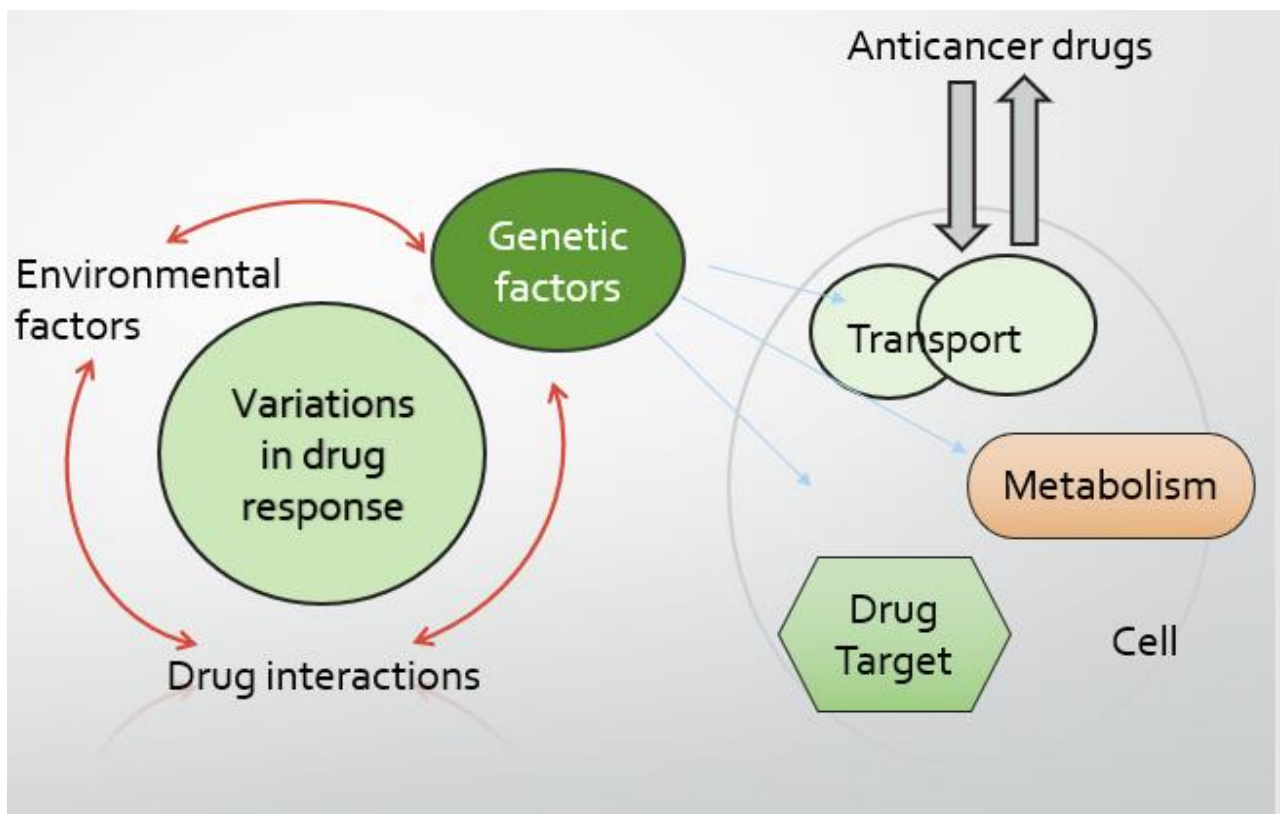


Figure 2: Multiple factors contributing to variations in drug response

2.3 Cancer pharmacogenomics

Each cancer has a unique combination of genetic mutations, such as an alteration in the nucleic acid sequence of the genome of an organism, virus, or extrachromosomal DNA, and even cells within the same tumor may have different genetic changes. It has also been commonly observed that the same types and doses of treatment can result in substantial differences in efficacy and toxicity across patients (Evans and Relling, 1999). This can effectively help in choosing drugs that target specific mutations within cancerous cells, identifying the patients at risk for severe toxicity to a drug, and also proposing those treatments which is beneficial for the patients.

Henceforth there is an increasing number of genomic variants being studied and identified as potential therapeutical targets and drug metabolism modifiers (Cascorbi *et al.*, 2013). This genomic information along with the tumor specific information is used to determine a personalized approach to cancer treatment. There are cancer driven alterations including the somatic DNA mutations and inherited DNA variants which impact the pharmacogenomic strategies affecting the pharmacokinetics and pharmacodynamics of metabolic pathways, making them potentially actionable drug-targets (Cascorbi *et al.*, 2013). Candidate's polymorphism can be searched for polymorphic DNA sequences within specific genes which can contribute to selecting effective therapeutic strategies for a patient (Crisafulli *et al.*, 2019). Hence helps to resolve pharmacokinetic or pharmacodynamic traits of a compound to a candidate polymorphism level and in turn contribute to selecting effective therapeutic strategies for a patient (Cockram *et al.*, 2010). One of the biggest challenges in using pharmacogenomics to study cancer is the difficulty in conducting studies in human. Drugs used for chemotherapy are too toxic to give to healthy individuals, which makes it difficult to perform genetic studies between related individuals.

Here personalized medicine also comes into play which is tailored and precise way of diagnosing and treating diseases like cancer. It results in early detection of mutations compared to previously existing methods. It also helps to lower the health care cost by avoiding unnecessary treatments and hence drugs with higher likelihood of success in subpopulation can be developed. Here (figure 3) describes the various events that occur in the process of personalized medicine.

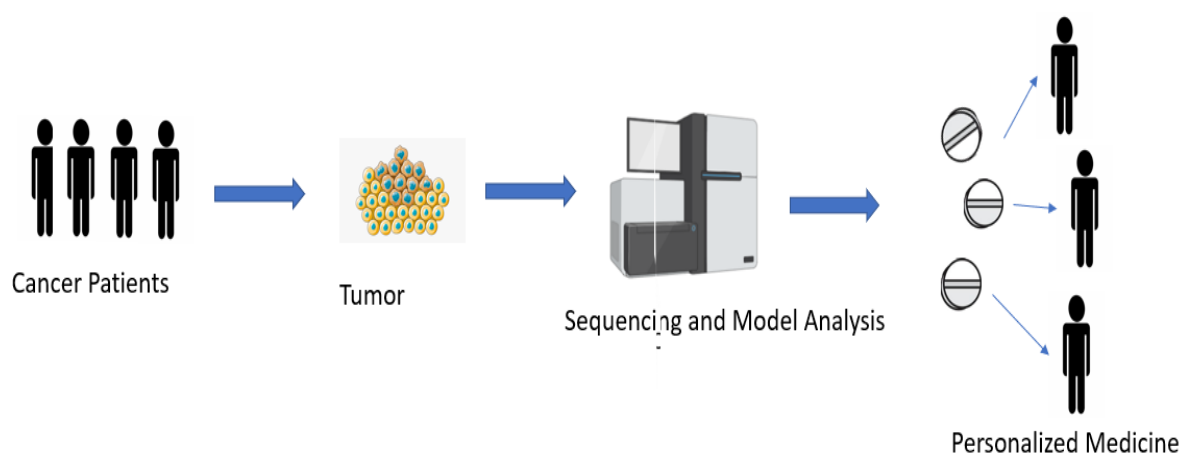


Figure 3: Concept of personalized medicine

2.4 Bioinformatics resources for cancer pharmacogenomics.

There are various available online tools for PGx Research such as the Pharmacogenomics Knowledgebase (PharmGKB) which is a comprehensive resource that curates knowledge about the impact of genetic variation on drug response including dosing guidelines, drug labels, gene-drug associations, and genotype-phenotype relationship. Similarly, The Drug Gene Interaction Database (DGIdb) is a database and web interface for identifying known and potential drug-gene relationships and The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) is a database of known and predicted protein interactions. There is a Catalog of Somatic Mutations in Cancer (COSMIC) which is a part of the Cancer Genome Project and stores and displays somatic mutation information and related details with information relating to human cancers.

3. MATERIALS AND METHODS

3.1 Mining of cancer related SNPs

Cancer related SNPs were searched to differentiate them from other mutations to focus on their major role at genetic level. Several tools and databases are freely available listing the clinically important SNPs. One such Database is PharmGKB. The PharmGKB contains genomic, phenotype and clinical information collected from PGx studies. PharmGKB contains SNP information and provides several tools for submitting, editing viewing and processing as well as accessing the information in the dbSNP database. The PharmGKB provides information about the curated pathways, Drug labeled annotation, Clinical guidelines annotation, and annotated drugs. In this study, we used PGx information about SNPs by manually searching for the neoplasm related SNPs and their PharmGKB variant annotations which reports the association between a variant and a drug phenotype from a publication hence total of 798 SNPs were retrieved.

Table 1: Details of considered SNPs

S. No	Genes	Variant
1	ABCA1	rs10024471
2	ABCB1	rs10144771, rs10405238, rs1042028, rs1042597, rs1042605,
3	ABCB11	rs10426377
4	ABCB4	rs10426628, rs1044457
5	ABCB5	rs1045642
6	ABCC1	rs1048977, rs10497346
7	ABCC10	rs10499563
8	ABCC11	rs10509681, rs1051640, rs1056892, rs1058930
9	ABCC2	rs10815019, rs10841661, rs10916825, rs10929302, rs10950831, rs11022922, rs11045585, rs11046217
10	ABCC3	rs1113129
11	ABCC9	rs11141915
12	ABCG2	rs11195419, rs1128503, rs1142345
13	ACYP2	rs1149222
14	ADGRG7	rs115349832
15	ADH7	rs11572080
16	ADRA2A	rs11572103, rs11598702
17	ALDH3A1	rs116134453
18	ANK3	rs11615
19	ARHGEF10	rs11646213
20	ARVCF; COMT	rs11671784, rs11692021
21	BDNF	rs117308378
22	CASP1	rs117412990
23	CASP5	rs117484357

24	CBR1	rs117876855
25	CBR3	rs11861118
26	CCDC70	rs11882256
27	CCDC77	rs1214763
28	CD96	rs12201199
29	CDA	rs12233719, rs12305038, rs12468274, rs12468485 rs12658397, rs12762549, rs1277441, rs12948783
30	CDH13	rs13401281
31	CES1	rs138385713, rs139368788
32	CES1P1	rs139544515, rs141213385
33	CES2	rs141531882
34	CLCC1	rs1418553
35	CMPK1	rs143414470
36	COL1A2	rs144470777
37	COMT	rs146644707, rs146898897, rs1523127, rs1523130 rs1551285, rs1617640, rs165728, rs167769, rs168107, rs1695
38	COMT; TXNRD2	rs17110453
39	CRYBG2	rs17216177
40	CYP2C8	rs17287570, rs1736557, rs17376848, rs174699, rs17583889, rs17822471
41	CYP2J2	rs17822931
42	CYP3A4	rs17863783, rs17868320, rs17868323
43	CYP3A5	rs1799782
44	CYP4F2	rs1799971
45	DAPK1	rs1800460
46	DPYD	rs1800469, rs1800629, rs1800871, rs1800896, rs1801030, rs1801131, rs1801133, rs1801158, rs1801159, rs1801160, rs1801265, rs183205964, rs185217050, rs185346775, rs1872328, rs187805828, rs1885301, rs191934521, rs2010963 rs2011404
47	DPYS	rs2019604, rs2020870, rs2032582
48	EGFR	rs2069762, rs2069835, rs2070474, rs2072671
49	EPO	rs2075252
50	ERCC1	rs2075507, rs2108623
51	FAT1	rs2160652
52	FDPS	rs2180314
53	FMO2	rs2227983
54	FMO3	rs2228100
55	GABBR2	rs2228145, rs2228171
56	GPR35	rs2231137
57	GPX3	rs2231142
58	GSTA2	rs2232228
59	GSTA5	rs2233302
60	GSTP1	rs2235013
61	HAS3	rs2235047
62	HNMT	rs2236168

63	HTR2A	rs2239393
64	HTR3E	rs2242480
65	IL10	rs2244613, rs2244614
66	IL2	rs2273697, rs2290271
67	IL6	rs2291767
68	IL6R	rs2293348
69	KCNQ1	rs2294950
70	KCNQ5	rs2297480
71	LARP1B	rs2297595
72	LMNTD1	rs2304389
73	LRP2	rs2305364, rs2306283
74	MAN1A1	rs2425886
75	MIR2054	rs2459693
76	MIR27A	rs25489, rs2600834
77	MTHFR	rs2622604, rs2669429
78	MUC16	rs2677760
79	NCOA7	rs2699905
80	NFE2L2	rs2740574
81	NR1I2	rs2779562, rs2804402, rs2811178
82	NSUN3	rs2959023
83	NT5C2	rs3024971
84	NT5C3A	rs316019
85	OPRK1	rs3212986
86	OPRM1	rs3397
87	OTOS	rs35599367
88	PDE3A	rs36024412
89	PHC1	rs367619008
90	PIK3CA	rs3730089, rs3740066, rs3750117, rs3755319
91	PIK3R1	rs3760091
92	POM121L2	rs3813627
93	RHBDF2	rs3813628
94	SERPINA6	rs3887137
95	SIRPA	rs3918290
96	SLC10A2	rs4124874, rs41269255
97	SLC13A3	rs4148323
98	SLC15A1	rs4148350
99	SLC16A5	rs4148808
100	SLC1A1	rs4149056
101	SLC22A17	rs4149178
102	SLC22A2	rs42524
103	SLC22A7	rs4261716
104	SLC25A13	rs4407290
105	SLC28A1	rs45589337, rs4646316
106	SLC28A3	rs4655226, rs4680, rs4715354
107	SLCO1B1	rs4788863, rs4818
108	SLCO1B3	rs4834232, rs4877847, rs4880

109	SLCO4C1	rs4982753
110	SLCO6A1	rs532545
111	SOD2	rs553668, rs554344
112	SPG7	rs55886062
113	SPRY2	rs56038477, rs56276561
114	SSU72	rs56293913
115	STAT6	rs5744168, rs5746849
116	SULT1A1	rs580253, rs602950, rs60369023
117	SULT1A1; SULT1A2	rs62298861
118	SULT2B1	rs6265, rs6269
119	TACR1	rs6311
120	TAOK3	rs6431558, rs6443624
121	TENM4	rs6443950
122	TGFB1	rs6668296
123	TGFB2	rs6690069
124	TLR5	rs6721961
125	TMEM131L	rs67376798
126	TNF	rs6755571
127	TNFRSF1B	rs6759892
128	TOMM40L	rs6785049, rs7016778
129	TPMT	rs712829, rs717620, rs7187684
130	TYMS	rs7194667
132	UGT1A1	rs72549307, rs7287550, rs729147, rs7319981
131	UGT1A10; UGT1A6; UGT1A7; UGT1A8; UGT1A9	rs73420732, rs737866, rs740603, rs7439366, rs750155
132	UGT1A4	rs75017182, rs75267292, rs7586110, rs7668258, rs768172
133	UGT1A8	rs770063251, rs77475703, rs7754103, rs776746
134	UGT1A9	rs7853758, rs795484
135	UGT2B7	rs7977213, rs8001466, rs8056100, rs8187710
136	UPB1	rs8192924
137	VEGFA	rs833061, rs871514
138	VPS13D	rs885004
139	XDH	rs895819, rs9024
140	XRCC1	rs9332377, rs9351963
141	ZMIZ1	rs9393888
142	ZNF165	rs9514091
143	ZNF568	rs9657362

Data Filtration: The collected SNPs were then filtered out to separate variants and their related genes from various other information such as p-value, and the duplicate values are separated where a total of 269 values of SNPs along with INDELs with their rs-id and associated genes

are present. These values are further filtered to remove the INDELs from SNPs and in total 265 SNPs were considered for further analysis with 143 genes (table 1).

3.2 Distribution pattern analysis of major and minor SNP alleles

Various SNPs are well distributed around the globe with different effects on different populations which form the basis of the population genetic. In our study we used the Ensemble browser which is one of several well-known genome browsers for the retrieval of genomic information which is used here to study population genetics of World population. Information about genes, transcripts and further annotation can be retrieved at the genome, gene and protein level. Allele frequencies in different populations are shown graphically, and in tabular format. The pie chart represents the distribution percentage of Minor and Major allele in all phase as well as sub-populations. They are represented by three-letter population codes analyzed by the 1000 Genomes project, which can be hovered over to know what they mean. The human related distribution pattern of different SNPs was studied by typing either the name of the gene, or the related rs-id in the search box wherein the information displayed were noted down with the major and minor alleles well distributed in African, East Asian, American, South Asian and European populations along with their sub population distribution.

The IGVdb portal encompasses the IGVBrowser that houses genotype data of samples that were recruited in the IGVC project. The IGVC data provide a basal level variation data in Indian population to study genetic diseases and pharmacology. The reported population distribution pattern of the SNPs (major and minor allele frequency) of an Indian Population is analyzed and noted.

3.3 Annotation of the SNPs

Location of the SNPs as biomarker is of utmost importance in terms of the population genetics. SNPnexus tool was used to annotate the SNPs which is a web-based variant annotation tool designed to simplify and assist in the selection and prioritization of known and novel genomic alterations. SNPnexus allows single queries using dbSNP identifiers or chromosomal regions for annotating known variants and also allows batch queries using dbSNP identifiers or genomic coordinates. It is a tool for Genomic Mapping by providing genomic coordinates for the queried SNP and for known SNPs, the tool provides related genotypes and allele frequency for population data. Currently SNPnexus supports the two most recent human genome assemblies GRCh38/hg38, GRCh37/hg19 and NCBI36/hg18. When the batch query of the SNPs for annotation are fed into the tool the table containing the information of all the dbSNPs,

their chromosome position, REF allele, ALT allele, Minor Allele, contig, contig position, band position i.e., the SNP's cytogenetic location were generated which were later used to create interactive matrix of pairwise linkage Analysis.

3.4 Pairwise linkage analysis

We have used LD matrix tool to create an interactive heatmap matrix of pairwise linkage disequilibrium statistics using the the LDlink suite which is a suite of web-based applications designed to effectively and easily interrogate linkage disequilibrium in population groups. Each included application in the suite is specialized for interrogating and displaying unique aspects of linkage disequilibrium. LD throughout the genome reflects the population history, the breeding system and the pattern of geographic subdivision, whereas LD in each genomic region reflects the history of natural selection, gene conversion, mutation and other forces that cause gene-frequency evolution (Slatkin, 2008). Information annotated from SNP nexus was used to gain generate the matrix. We looked for the correlation value ≤ 0.7 which was found in chromosome numbers 1, 2, 3, 4, 10, 11 and 12 and was absent in Chromosome 5, 6, 7, 8, 9, 13, 14, 16, 17, 19, 20 and 21.

3.5 Selection of Non-Synonymous SNPs and Site Directed Mutagenesis

The non-synonymous SNPs were separated from the list of total SNPs and a SNP having ID rs1695 was selected to study which is a non-synonymous polymorphism in exon 5 of GSTP1 gene. It is also called GSTP1*B with nucleotide change at 313 with A>G, which resulted into the coding SNP change at position 105 from I>V. The 105Val protein is associated with lower enzyme activity than 105Ile (Watson *et al.*, 1998). Further 3-dimensional structure of GSTP1 protein complex has been retrieved from the PDB database having PDBID: 6LLX. And site-directed mutagenesis was performed by changing the amino acid residue of 104 Isoleucine to 104 Valine through UCSF chimera, to get the structure of GSTP1 variant, as its 3D crystal structure is not reported.

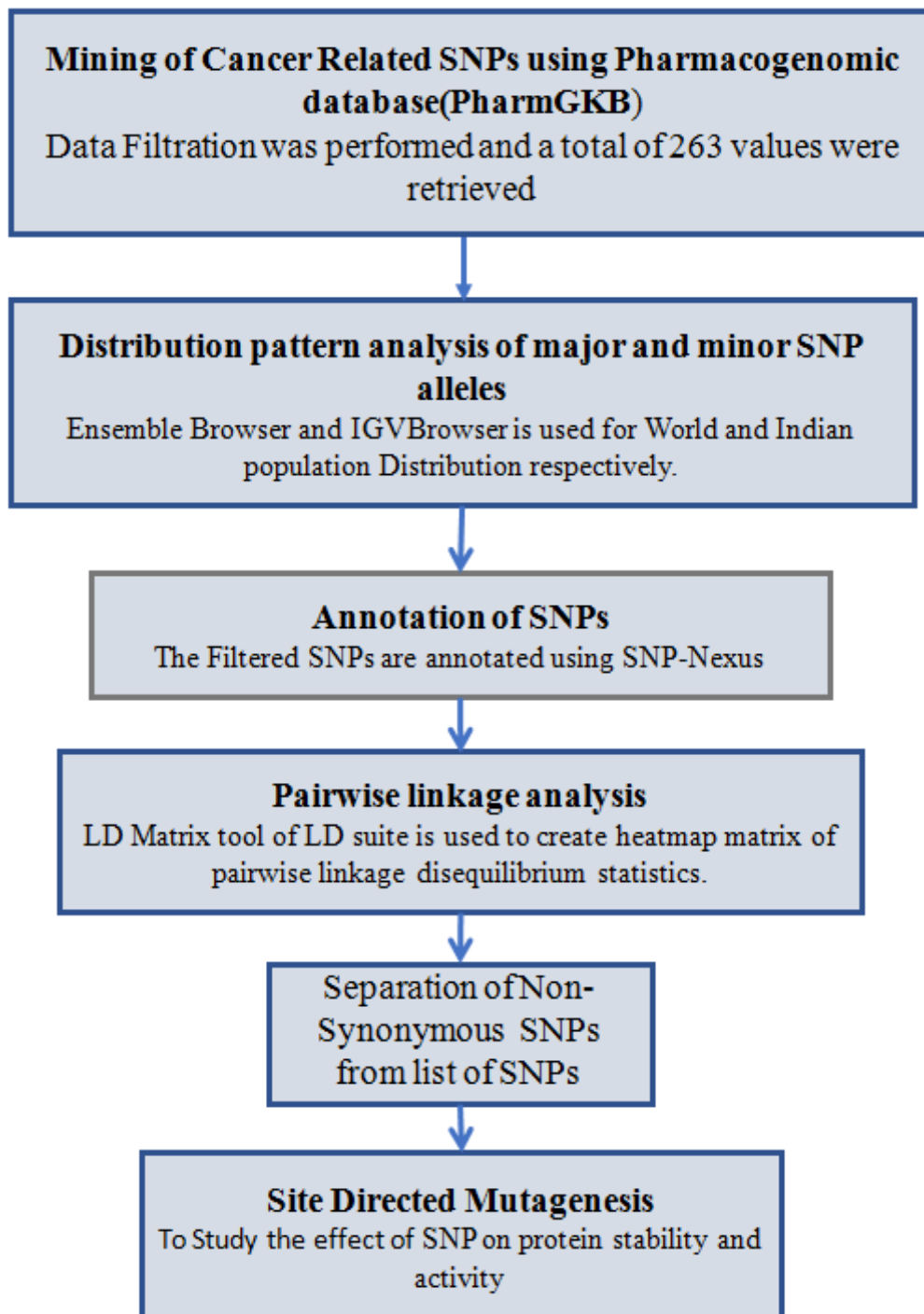


Figure 4: Schematic representation of work flow

4. RESULTS AND DISCUSSION

4.1 SNP Distribution analysis of Indian and World Population

Mining of SNPs from Indian and World Population suggested a wide range of distribution pattern. On the basis of availability and non-availability of observed SNPs and major and minor alleles in different population following categories are considered namely i) Indian population specific ii) South Asian specific iii) African related population iv) American v) East Asian vi) European and vii) other mix pattern population. Out of all the observed SNPs only 23 of them are reported in Indian population. Out of total of 265 values 23 of them have major allele frequency of 100% in overall world population. SNP with rs1042028, rs1801030, rs187805828, rs367619008, rs770063251 are specifically reported in Ashkenazi Jewish, Finnish, Non-Finnish European and other Populations. Out of 265, 31 of the SNPs are also reported in Indian Population and are encoded by variety of genes listed in the table below (Table 2)

Table 2: SNPs Reported in Indian Population1

S. No	Genes	Variant
1	ABCA1	rs1048977
2	ABCC2	rs10499563
3	BDNF	rs10509681
4	CASP1	rs11572080
5	CDA	rs167769
6	CES1	rs1799782
7	COMT	rs1800460, rs1800469
8	CYP2C8	rs1800629, rs1800871
9	CYP3A4	rs1801131
10	HTR2A	rs1801133, rs2010963
11	IL10	rs2069762
12	IL2	rs2069835
13	IL6	rs2228145, rs2244613
14	IL6R	rs2740574
15	MTHFR	rs4149056, rs4149178
16	SLC22A7	rs580253
17	SLCO1B1	rs6265
18	STAT6	rs6269
19	TAOK3	rs6311
20	TGFB1	rs717620
21	TNF	rs740603
22	TOMM40L	rs795484
23	TPMT	rs6311
24	UGT1A1	rs4148323
25	VEGFA	rs3813628
26	XRCC1	rs3887137

4.2 Linkage disequilibrium

In population genetics linkage disequilibrium (LD) is the non-random association of alleles at different loci in a given population. To observe the distribution of pairwise linkage disequilibrium among the given SNP variants, we looked for both R^2 and D' respectively. Both R^2 and D' are the two most widely used measures of Linkage disequilibrium and provides a non-random association of alleles at two or more loci and they are used together for mapping purposes because of their significant advantages over each other. In this study we analysed LD for all SNP variants, and considered only those of them as LD pairs whose values fall in considerable range for both D' and R^2 parameters, i.e., ≥ 0.7 . Out of 22 chromosomes eight of them show values greater than the given 0.7 (table 3). The range under which they fall are within the range of 1 to 10 with chromosome 1 showing the highest range of frequency 10 out of total of 47 values and lowest for chromosome number four, eleven and twelve with values as 1. Contrastingly chromosomes five, six, eight, nine and thirteen to sixteen do not fall above the given correlation range (figure 5).

Table 3: Chromosome numbers showing significant correlations

Chromosome number	Total Values	Value ≥ 0.7
Chromosome 1	47	10
Chromosome 2	30	8
Chromosome 3	11	2
Chromosome 4	13	1
Chromosome 5	4	0
Chromosome 6	19	0
Chromosome 7	22	0
Chromosome 8	5	0
Chromosome 9	8	0
Chromosome 10	20	3
Chromosome 11	6	1
Chromosome 12	14	1
Chromosome 13	7	0
Chromosome 14	2	0
Chromosome 15	2	0
Chromosome 16	18	0
Chromosome 17	4	0
Chromosome 18	1	0
Chromosome 19	12	0
Chromosome 20	2	0
Chromosome 21	2	0
Chromosome 22	14	4

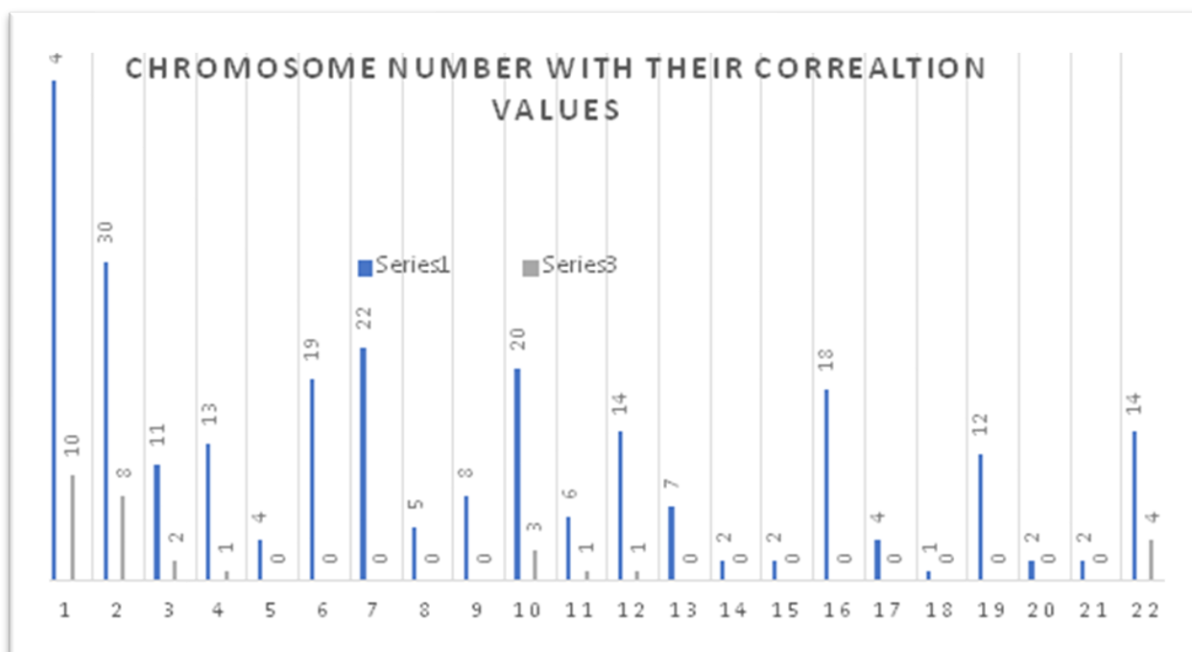


Figure 5: Chromosome showing correlation values greater than the reference value of 0.7

The graph above shows the range of values of different chromosome showing correlation values greater than the reference value of 0.7(Figure 5). Here chromosome 1 shows the highest peak with value 10 and chromosomes 4, 11, 12, and 18 shows lowest peak of value of 1.

A total of 31 SNPs showed significant Linkage Disequilibria. The highest value is the value 1 which is given by five SNPs and the lowest value of 0.704 is given by one of the SNP (Table 4).

Table 4: Details of SNPs showing significant linkage disequilibrium (LD)

S.No	rs_x	rs_y	Values
1	rs75017182	rs56038477	1
2	rs115349832	rs56038477	1
3	rs3813628	rs3813627	0.994
4	rs602950	rs532545	0.95
5	rs2072671	rs532545	0.854
6	rs2072671	rs602950	0.902
7	rs56276561	rs56038477	0.83
8	rs56276561	rs75017182	0.83
9	rs56276561	rs115349832	0.83
10	rs3813628	rs3813627	0.994
12	rs4261716	rs7586110	0.761

13	rs4261716	rs11692021	0.761
14	rs6759892	rs7586110	0.746
15	rs6759892	rs11692021	0.746
16	rs6759892	rs4261716	0.98
17	rs4124874	rs871514	0.876
18	rs3755319	rs871514	0.775
19	rs3755319	rs4124874	0.857
20	rs1523127	rs1523130	0.87
21	rs2459693	rs2677760	0.739
22	rs7439366	rs7668258	1
23	rs2804402	rs1885301	0.873
24	rs8187710	rs17216177	0.929
25	rs11572080	rs10509681	1
26	rs580253	rs554344	1
27	rs1277441	rs795484	0.778
28	rs740603	rs5746849	0.826
29	rs2239393	rs6269	0.949
30	rs4818	rs2239393	0.704
31	rs165728	rs174699	0.818

4.3 SNP Annotation

To get more information about the SNPs we performed the annotation of these SNPs to observe their possible impact. It was observed that two of the SNPs with rs_id rs17868323 and rs11572103 and coding for UGT1A10; UGT1A6; UGT1A7; UGT1A8; UGT1A9, CYP2C8 and NR1I2 genes respectively which are involved in cancer also show significant role in other diseases such as HIV infection and drugs irinotecan, oxaliplatin, s1 (combination), paclitaxel, ibuprofen, amodiaquine, artesunate and paclitaxel, carboplatin, efavirenz are respectively used for the treatment. Another SNP with rs id rs1800469 is coded by gene CYP2C8 is involved in multiple sclerosis and drug used against which is irinotecan. There is a pattern of relation between HIV infection and cancer general population, people who are infected with HIV are currently about 500 times more i.e., compared with the likely to be diagnosed with Kaposi sarcoma, 12 times more likely to be diagnosed with non-Hodgkin lymphoma, and, among women, 3 times more likely to be diagnosed with cervical cancer (Ramírez *et al.* 2017). Also, HIV-infected people with a range of cancer types are more likely to die of their cancer than HIV-uninfected people with these cancers (Coghill *et al.* 2015). Other SNPs rs_id rs10509681, rs1058930, rs1113129, rs11572080, rs17110453 are also encoded by similar gene CYP2C8, and are have tendency to cause other diseases like Peripheral nervous system disease, neurotoxicity syndrome and stroke.

Table 5: Top 20 SNPs with their Clinical Significance

S no.	Gene	rs_id	Disease	Drugs
1	UGT1A10; UGT1A6; UGT1A7; UGT1A8; UGT1A9	rs17868323	hyperbilirubinemia, HIV infection, kidney transplantation	irinotecan, oxaliplatin, s 1 (combination)
2	TGFB1	rs1800469	Multiple Sclerosis	irinotecan
3	CYP2C8	rs11572103	HIV Infection, Hypertension, Portal	paclitaxel, ibuprofen, amodiaquine, artesunate
4	NR1I2	rs1523127	Thrombocytopenia, HIV infections	paclitaxel, carboplatin, efavirenz
5	UGT1A8	rs1042597	menopause, lung transplantation, kidney transplantation, diarrhoea, Schizophrenia	Tamoxifen, acetaminophen, ABT-751, tapentadol, desmethylnaproxen
6	CYP2C8	rs10509681	Peripheral nervous system disease, coronary artery disease	rosiglitazone, paclitaxel, ibuprofen, pioglitazone, tacrolimus
7	CYP2C8	rs1113129	Anaemia, Neurotoxicity syndrome	Paclitaxel
8	TPMT	rs1142345	Sjogren's syndrome, Colitis, Ulcerative, Lupus Erythematosus, Systemic	Mercaptopurine, thioguanine, azathioprine
9	TGFB2	rs1418553	cognitive dysfunction	Opioids
10	NSUN3	rs14447077	Hand-foot syndrome	Capecitabine
11	SSU72	rs14689889	Hand-foot syndrome	Capecitabine
12	ABCG2	rs2231137	cessation, mucositis, Epilepsy, Arthritis, Rheumatoid	methotrexate, granisetron, palonosetron, valganciclovir
13	ABCG2	rs2231142	Gout, CNS infection	rosuvastatin, atorvastatin, simvastatin, imatinib, sunitinib

14	ABCC2	rs2273697	CNS infection, glomerular disease, Epilepsies, Partial	pravastatin, carbamazepine, oxcarbazepine
15	SLCO1B1	rs4149056	coronary stenosis, Dyslipidaemia, etc	atorvastatin, hmg coa reductase inhibitors, pravastatin, simvastatin
16	ADRA2A	rs553668	cognitive dysfunction, sedation	doxazosin, phenoxybenzamine, opioids
17	TNFRSF1B	rs3397	cognitive dysfunction	Tumour necrosis factor alpha (TNF-alpha) inhibitors
18	ABCB1	rs2032582	CNS depression in infants, Myasthenia Gravis, Cholelithiasis	tacrolimus, fluoxetine
19	IL10	rs1800896	Drug Hypersensitivity, Arthritis, Rheumatoid	cyclosporine, mycophenolate mofetil
20	TPMT	rs1142345	Sjogren's syndrome, Colitis, Ulcerative, Lupus Erythematosus, Systemic	mercaptopurine, thioguanine, azathioprine

4.4 Site directed Mutagenesis

SNP-hits were observed using PharmGKB database which categorized them as UTR, Synonymous, non-Synonymous, intronic missense, etc. Among these the Protein GSTP1 with SNP-id (rs1695) was considered for site directed mutagenesis. Apart from causing cancer, the gene also caused other disease like syncope, Lupus erythematosus, Lupus Nephritis and tuberculosis. The drugs that are used against these are namely cisplatin, fluorouracil, oxaliplatin, cyclophosphamide, doxorubicin. The clinical significance of the same is mentioned in the table (table 6). It is not reported for Indian population, it shows major and minor allele A and G with 65% and 35% frequencies respectively in the overall World Population. This gene is mostly reported in East Asian Population as A being the major allele showing a frequency of 82% in the population.

Table 6: Description of Clinical Significance of GSTP1 Protein

1	Protein Name	GSTP1
2	rs_id	rs1695
3	Indian Population	Not reported
4	Major	A
5	Minor	G
6	Classification	Non-synonymous /missense
7	Disease	Syncope, Lupus erythematosus, Lupus Nephritis, tuberculosis
8	Drug	cisplatin, fluorouracil, oxaliplatin, cyclophosphamide, doxorubicin

The figure below (figure 6) shows the allele frequencies of GSTP1 protein based on the data from 1000 genome project. The Major and Minor alleles i.e. A and G show the overall percentage of 64.74% and 35.26% respectively. The highest is for South Asian Population and Lowest is for African Population.


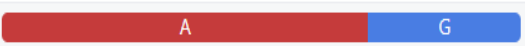
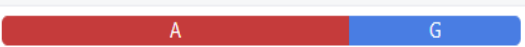



>	AGGREGATED POPULATIONS ⓘ		DISTRIBUTION	A ⬆	G ⬆
✓	All populations	n=5,008		64.74%	35.26%
>	SAS	n=978		70.55%	29.45%
>	EUR	n=1,006		66.90%	33.10%
>	EAS	n=1,008		82.14%	17.86%
>	AMR	n=694		52.45%	47.55%
>	AFR	n=1,322		51.97%	48.03%

Figure 6: Alleles Frequencies of protein GSTP1 based on data from the 1000 genome project, phase III.

The PDB structure for protein GSTP1 used for mutagenesis is 6LLX (pdb_id). It has a resolution of 1.581 Å (figure 7). The two ligands attached to the structure are GSH, MES with protein Glutathione S-transferase P and the sequence length of 215 amino acids (Figure 8). It has a total of 9 pockets with pocket one has the highest score of 20.55 and pocket 9 has the lowest score of 1.24 which was calculated from prank web tool. It has two chains A and B. The

wild type structure of the protein is shown in (figure 9) and after site directed mutagenesis performed through UCSF chimera, the mutated structure is obtained (figure 10).

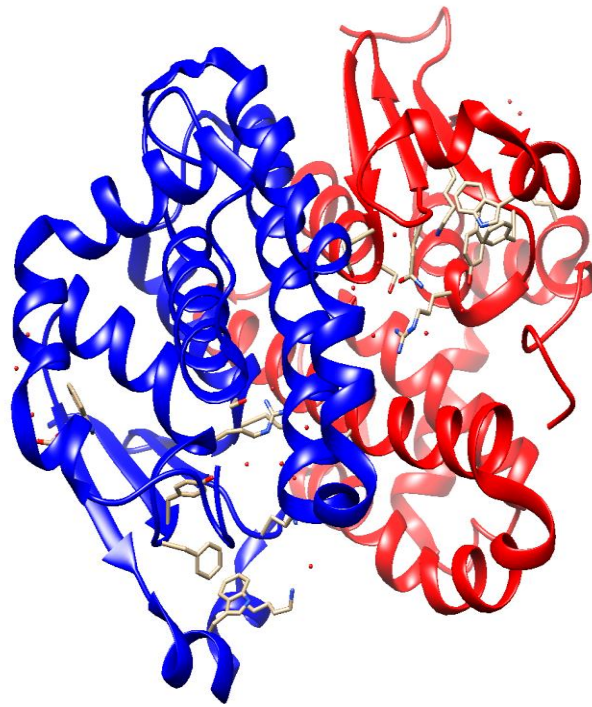


Figure 7: Ligand-free structure of GSTP1 (PDB ID 6LLX)

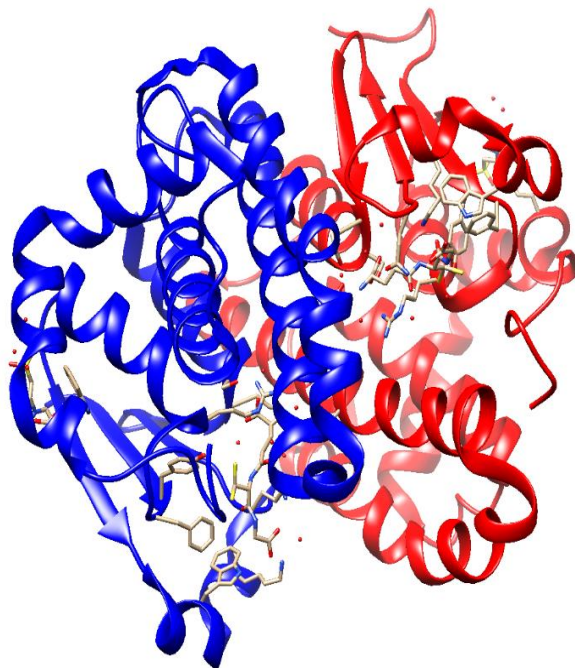


Figure 8: Structure of GSTP1 (6LLX) with complex MES and GSH

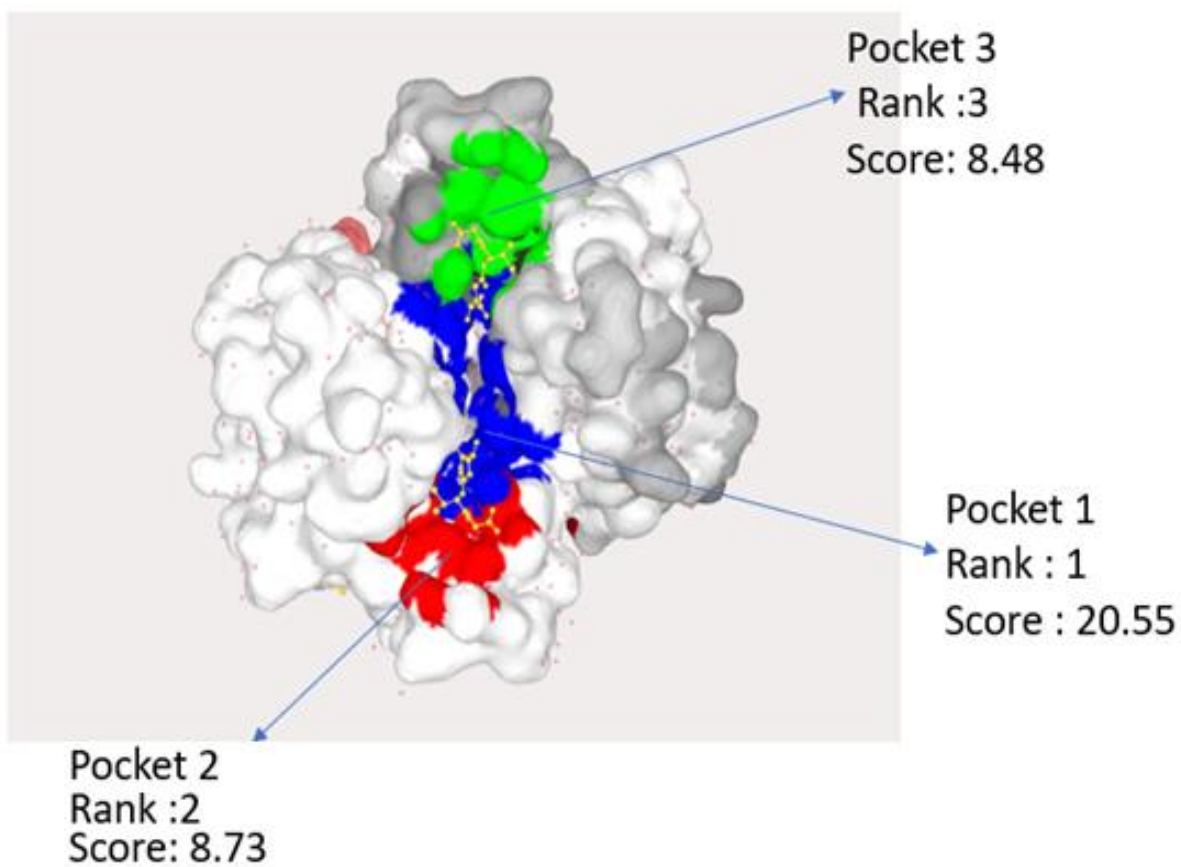


Figure 9: Snapshot of pocket rank one, two and three along with their score taken from prank web tool.

The above diagram (figure 9) shows the possible binding pocket of GSTP1 with PDB ID (6LLX). The pocket ranked as 1 has the highest score of 20 with residues with residue id A_101 CYS, A_102 LYS A_104 ILE A_13 ARG A_49 TYR A_51 GLN A_52 LEU A_53 PRO A_64 GLN A_65 SER A_66 ASN A_94 ASP A_97 GLU A_98 ASP B_101 CYS B_102 LYS B_13 ARG B_49 TYR B_51 GLN B_52 LEU B_53 B_64 GLN B_65 SER B_66 B_94 ASP B_97 GLU B_98 ASP respectively. The second rank is followed by pocket 2 with a score 8.73 of and finally pocket 3 with score 8.43.

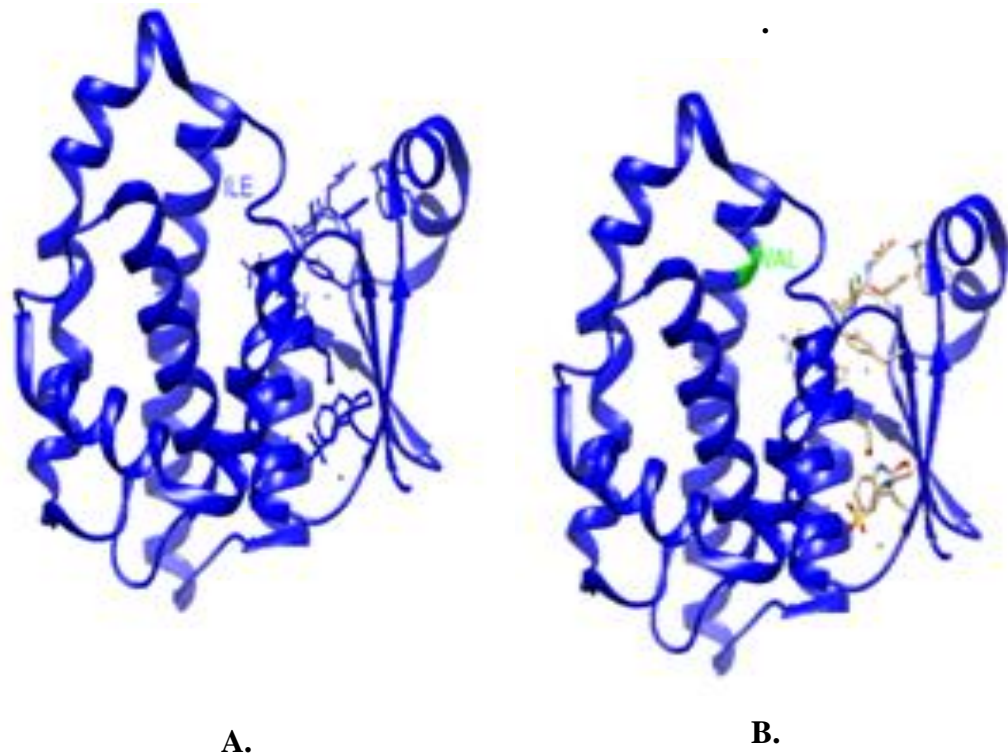


Figure 10 A. Wild Type Structure B. Mutant Structure of GSTP1

Structure of mutant type (dbSNP ID: rs1695) was produced creating mutagenesis in GSTP1 protein by replacing single amino acid residue i.e., isoleucine to valine at position 104 using UCF Chimera. Following the same a change in different energy is observed. To observe the impact of selected SNPs on protein, bonded, non-bonded, electrostatic, torsions, improper and total energy changes were computed for wild type as well as mutant protein using SPDBV_4.10. Results are tabulated in the table 7.

Table 7: Results of energy difference calculation of wild type and mutant type protein (GSTP1)

	Bonds (KJ/mol)	Angles (KJ/mol)	Torsion (KJ/mol)	Improper (KJ/mol)	Nonbonded (KJ/mol)	Electrostatic (KJ/mol)	Total (KJ/mol)
Wild type	408.949	1067.410	972..924	279.717	-6049.69	-5843.31	-9164.002
Mutant Type	409.213	1067.254	972.019	279.738	-6046.07	-5843.31	-9161.156

The above table shows clear difference in total energy i.e., the energy change of -2.846 KJ/mol is observed. The difference in energy for bonded, non-bonded, angle, torsion, improper and electrostatic energies (KJ/mol) are 0.264, -3.62, -0.156, -0.905, 0.021, 0 respectively. In wild type at position 104 residue isoleucine contribute in bond, angles, torsions, improper, non-bonded, electrostatic and total energy of each residue (E) as 1.437, 1.584, 2.087, 0.860, -6.22, -18.46 and -2.154 (KJ/mol) respectively whereas mutant residue valine 104 contribute in the bond, angles, torsions, improper, non-bonded, electrostatic as 1.701, 1.428, 1.182, 0.881, -7.59, -18.46 respectively. Since is the overall energy of mutant type is reduced hence the effect of SNP makes the protein more stable than the wild type. This means the over all activity of the protein is affected.

5. SUMMARY AND CONCLUSION

This study identifies the importance of SNPs in cancer related studies as well as describes the vital role they play in population genetics. It was observed that some of Indian-population specific SNPs also shared a range of distribution pattern among the East Asians., South Asians, African and American population. Annotation of the cancer related SNPs showed that they are also involved in the crucial physiological process of clinical significance. Observed patterns and annotation results, will prove to be beneficial in future for the selection of potential SNPs as that can be used as a marker for various purposes including- population-specific pathophysiological and clinical studies. To observe the effect of non-synonymous SNPs on the structural properties of wild type protein structure, site directed mutagenesis was also performed on non-synonymous protein which concluded an considerable change in energy from wild to mutant which means the gene is affected and such alterations in affinity may lead to variable rate of drug metabolism in different population. Their proteins also showed significant roles in ADME process which means that the study could be potentially valuable in pharmacogenomics mainly for population-specific clinical trials and to avoid genetic variant based adverse drug reactions to different individuals in the given world as well as Indian Population.

REFERENCES:

- Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: A map of human genome variation from population-scale sequencing. *Nature*. (2010), 467: 1061-1073.
- Cancer pharmacogenomics: strategies and challenges Heather E. Wheeler, Michael L. Maitland, M. Eileen Dolan, Nancy J. Cox, and Mark J. Ratain Published online 2012 Nov 27.
- Cascorbi I, Bruhn O, Werk AN (May 2013). "Challenges in pharmacogenetics". *European Journal of Clinical Pharmacology*. 69 Suppl 1: 17–23.
- Chuang JH, Li H: Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol*. (2004)
- Cockram J, White J, Zuluaga DL, Smith D, Comadran J, Macaulay M, et al. (December 2010). "Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome". *Proceedings of the National Academy of Sciences of the United States of America*. **107** (50): 21611
- Coghill AE, Shiels MS, Suneja G, Engels EA. Elevated cancer-specific mortality among HIV-infected patients in the United States. *Journal of Clinical Oncology* 2015; 33(21):2376-2383.
- Concetta Crisafulli, Concetta; Romeo, Petronilla Daniela; Calabrò, Marco; Epasto, Ludovica Martina; Alberti, Saverio (2019). "Pharmacogenetic and pharmacogenomic discovery strategies". *Cancer Drug Resistance*. 2 (2): 225–241
- El-Deiry WS, Goldberg RM, Lenz HJ, Shields AF, Gibney GT, Tan AR, et al. (July 2019). "The current state of molecular testing in the treatment of patients with solid tumors, 2019". *CA: A Cancer Journal for Clinicians*. 69 (4): 305–343.
- Erichsen, H., Chanock, S. SNPs in cancer research and treatment. *Br J Cancer* **90**, 747–751 (2004).
- Evans WE, Relling MV (October 1999). "Pharmacogenomics: translating functional genomics into rational therapeutics" *Science*. 286 (5439
- Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. Global Cancer Observatory: Cancer Today. Lyon: International Agency for Research on Cancer; (2020)
- Ferlay J, Colombet M, Soerjomataram I, et al: Estimating the global cancer incidence and mortality in (2018): GLOBOCAN sources and methods. *Int J Cancer*

- Hernández-Ramírez RU, Shiels MS, Dubrow R, Engels EA. Cancer risk in HIV-infected people in the USA from 1996 to 2012: a population-based, registry-linkage study. *Lancet HIV* 2017 Aug 10.
- Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Res.* 2002;30(1):163–5.
- Housman, G., Byler, S., Heerboth, S., Lapinska, K., Longacre, M., Snyder, N., & Sarkar, S. (2014). Drug resistance in cancer: an overview. *Cancers*, 6(3), 1769–1792.
- Human single-stranded DNA binding proteins are essential for maintaining genomic stability
Nicholas W Ashton 1, Emma Bolderson, Liza Cubeddu, Kenneth J O'Byrne, Derek J Richard
- IGVBrowser—a genomic variation resource from diverse Indian populations
Ankita Narang, Rishi Das Roy, Amit Chaurasia, Arijit Mukhopadhyay, Mitali Mukerji, Indian Genome Variation Consortium, and Debasis Dash
- Krishna, R., & Mayer, L. D. (2000). *Multidrug resistance (MDR) in cancer. European Journal of Pharmaceutical Sciences*, 11(4), 265–283.
- Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, Watkins PB, Daly A, Wrighton SA, Hall SD, Maurel P, Relling M, Brimer C, Yasuda K, Venkataramanan R, Strom S, Thummel K, Boguski MS, Schuetz E. Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nat Genetics*. 2001;27:383–391.
- Li ZW, Dalton WS. Tumor microenvironment and drug resistance in hematologic malignancies. *Blood Rev.* 2006;20(6):333–42.
- Lindblad-Toh K, Winchester E, Daly MJ, Wang DG, Hirschhorn JN, Lavolette J-P, Ardlie K, Reich DE, Robinson E, Sklar P, Shah N, Thomas D, Fan J-B, Gingeras T, Warrington J, Patil N, Hudson TJ, Lander ES: Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat Genet.* 2000, 24: 381-386.
- Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 46, 3–26 (2001).
- Longley DB, Johnston PG. Molecular mechanisms of drug resistance. *J Pathol.* 2005;205(2):275–92.

- Mahdi KM, Nassiri MR, Nasiri K. Hereditary genes and SNPs associated with breast cancer. *Asian Pac J Cancer Prev.* 2013;14(6):3403-9.
- Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S., & Baradaran, B. (2017). The Different Mechanisms of Cancer Drug Resistance: A Brief Review. *Advanced pharmaceutical bulletin*, 7(3), 339–348
- MedlinePlus Bethesda (MD): National Library of Medicine (US); cited 2020 Jul 1
- Nat Rev Genet. 2013 Jan; 14(1): 23–34. Published online 2012 Nov
doi: 10.1038/nrg3352 PMCID: PMC3668552 NIHMSID: NIHMS467214 PMID: 23183705
Cancer pharmacogenomics: strategies and challenges Heather E. Wheeler, Michael L. Maitland, M. Eileen Dolan, Nancy J. Cox and Mark J. Ratain
- Norton P. Peet and Philippe Bey Pharmacogenomics: challenges and opportunities. *Drug Discovery Today.* 2001;6:495–498
- Office of the Registrar General and Census Commissioner, India, Ministry of Home Affairs, Government of India: Causes of death statistics.
- Pharmacogenetics and pharmacogenomics 11 October 2017 By Victoria Rollinson, Richard M. Turner & Munir Pirmohamed Corresponding author Victoria Rollinson
- Rahner, N. & Steinke, V. Hereditary cancer syndromes. *Deutsch. Arztebl. Int.* 105, 706–714 (2008).
- Schmidt, A. & Weber, O. F. In memoriam of Rudolf Virchow: a historical retrospective including aspects of inflammation, infection and neoplasia. *Contrib. Microbiol.* 13, 1–15 (2006).
- Shen TH, Carlson CS, Tarczy-Hornoch P (August 2009). "SNPit: a federated data integration system for the purpose of functional SNP annotation". *Computer Methods and Programs in Biomedicine*
- Slatkin M. (2008). Linkage disequilibrium--understanding the evolutionary past and mapping the medical future. *Nature reviews. Genetics*, 9(6), 477–485
- Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* 458, 719–724 (2009).

Szakács, G., Paterson, J. K., Ludwig, J. A., Booth-Genthe, C., & Gottesman, M. M. (2006). *Targeting multidrug resistance in cancer. Nature Reviews Drug Discovery*, 5(3), 219–234.

The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications Yuanyuan Zhang & Zemin Zhang 01 July 2020

Tomasetti C, Li L, Vogelstein B: Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* 355:1330-1334,201

Translational Research and Onco-Omics Applications in the Era of Cancer Personal Genomics. *Advances in Experimental Medicine and Biology*

Vermeersch, K. A., & Styczynski, M. P. (2013). Applications of metabolomics in cancer research. *Journal of carcinogenesis*, 12, 9.

Vogel F. Moderne problem der humangenetik. *Ergeb Inn Med U Kinderheilk* 1959;12:52–125.

Vogelstein, B. & Kinzler, K. W. The multistep nature of cancer. *Trends Genet.* 9, 138–141 (1993).

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. & Kinzler, K. W. Cancer genome landscapes. *Science* 339, 1546–1558 (2013).

Watson, M. (1998). Human glutathione S-transferase P1 polymorphisms: relationship to lung tissue enzyme activity and population frequency distribution. *Carcinogenesis*, 19(2), 275–280.

Witzmann, F., Grant, R. Pharmacoproteomic in drug development. *Pharmacogenomics J* 3, 69–76 (2003).

Yadav, A., & Katara, P. (2018). *In-silico mining of SNP-effects on structural properties of CYP2C9 and their consequences. 2018 International Conference on Bioinformatics and Systems Biology (BSB).*

Document Information

Analyzed document	ANAM.docx (D142174336)
Submitted	2022-07-18 12:36:00
Submitted by	pramod
Submitter email	pmkatara@gmail.com
Similarity	4%
Analysis address	pmkatara.alld@analysis.orkund.com

Sources included in the report

W	URL: https://en.wikipedia.org/wiki/Cancer_Pharmacogenomics Fetched: 2022-07-18 12:37:12		11
SA	University of Allahabad / for Plagiarism.pdf Document for Plagiarism.pdf (D136281645) Submitted by: pmkatara@gmail.com Receiver: pmkatara.alld@analysis.orkund.com		4
W	URL: https://oaepublishstorage.blob.core.windows.net/4a874645-f42e-413c-bb05-b0d961272afe/3039.pdf Fetched: 2022-07-18 12:37:14		1
SA	Karolinska Institutet / Pharmacol Rev-2011-Ma-437-59.pdf Document Pharmacol Rev-2011-Ma-437-59.pdf (D12615113) Submitted by: pingpong-orkund+anonymous.93@ki.se Receiver: pingpong-orkund.ki@analys.orkund.se		2
SA	Dibrugarh University, Dibrugarh / chapter 7 Dr Bhaskar Mazumder (1).docx Document chapter 7 Dr Bhaskar Mazumder (1).docx (D25715368) Submitted by: bhmaz@dibru.ac.in Receiver: bhmaz.dibru@analysis.orkund.com		1

Entire Document

A DISSERTATION ON
 In silico analysis of variants of cancer pharmacogenomic importance
 SUBMITTED TO THE
 DEPARTMENT OF BIOENGINEERING
 FACULTY OF ENGINEERING
 INTEGRAL UNIVERSITY, LUCKNOW
 IN PARTIAL FULFILMENT
 FOR THE
 DEGREE OF MASTER OF TECHNOLOGY
 IN BIOINFORMATICS
 BY
 Anam Tanveer