In silico analysis of single nucleotide polymorphisms of XPD gene in lung cancer patients

A Dissertation submitted in partial fulfillment of the requirement for the award of degree of

Master of Science

In

Biotechnology

Submitted by

Neha Garg

Roll No.: 301701018

Under the guidance of

Dr. Manoj Baranwal

Associate Professor

Dr. Siddharth Sharma

Associate Professor

DEPARTMENT OF BIOTECHNOLOGY

THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY

PATIALA-147004

July 2019

CERTIFICATE

This is to certify that the dissertation entitled "***In silico* analysis of single nucleotide polymorphisms of XPD gene in lung cancer patients**" submitted by degree of Master of Science in the subjects of Biotechnology, Thapar Institute of Engineering and Technology (TIET), Patiala is a bonafied work carried out by Neha Garg under the supervision of Dr. Manoj Baranwal, Associate Professor and Dr. Siddharth Sharma, Ph.D., Associate Professor, Department of Biotechnology, Thapar Institute of Engineering and Technology (TIET), Patiala and that no part of this work has been submitted for any other degree.

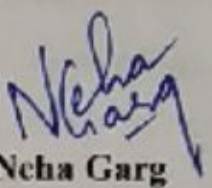
The assistance and help received during the course of investigation has been fully acknowledged.

Dr. Manoj
Baranwal, Associate
Professor, Department of
Biotechnology,
Thapar Institute of Engineering and Technology,Patiala.

Dr. Siddharth
Sharma, Associate Professor,
Department of
Biotechnology,
Thapar Institute of Engineering and Technology, Patiala.

DECLARATION

I, the undersigned, hereby declare that the work presented in the M.Sc. dissertation entitled "***In silico* analysis of single nucleotide polymorphisms of XPD gene in lung cancer patients**" has been carried out be me under the supervision and guidance of Dr. Manoj Baranwal, Associate Professor and Dr. Siddharth Sharma, Associate Professor, Department of Biotechnology, Thapar institute of Biotechnology, Patiala. Further, I declare that no part of this dissertation has been submitted for a degree or any other qualification of any other university or examining body in India/elsewhere.



Neha Garg

(Roll no. 301701018)

ACKNOWLEDGEMENT

Firstly, I would like to express my special thanks of my gratitude to **Dr. Manoj Baranwal**, Associate Professor and **Dr. Siddharth Sharma**, Associate Professor, Department of Biotechnology, Thapar Institute of Engineering and Technology for guiding me through the entire project. His constant motivation and brilliant supervision gave me an opportunity to learn a lot.

I express my special thanks to all the faculty members of Department of Biotechnology, Thapar Institute of Engineering and Technology, for their constant support, help and encouragement.

A special thanks to **Ms. Yogita Kautish** Ph.D. scholars, **Ms. Neha Srivastav, Mr. Sahil Jain**, Department of Biotechnolgy, Thapar Institute of Engineering and Technology, who always help me in every situation and their valuable guidance, expedient advice and affection during my work was valuable.

Above all I remember with gratitude the grace of Almighty for the successful completion of this work.

Neha Garg

Roll no. 3017101018

ABBREVIATIONS

XPD – Xeroderma pigmentosum group D

SNP – Single nucleotide polymorphisms

SIFT – Sorting Intolerant from Tolerant

PROVEAN – Protein Variation Effect Analyzer

POLYPHEN2 – Polymorphism Phenotyping v2

PHD-SNP – Predictor of human Deleterious Single Nucleotide

RMSD – Root- mean -square deviation

SCLC- Small cell lung cancer

NSCLC – Non- small cell lung cancer

CT scan – Computed tomography scan

CBC – Complete blood count

WHO – World Health Organization

ERCC2 – ERCC Excision Repair 2

NER – Nucleotide excision repair

TFIIH – Transcription Factor II Human

RI – Reliability Index

PDB – Protein Data Bank

CS – Cockayne Syndrome

TTD – Trichothiodystrophy

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Lung cancer is one the leading cause of cancer death in both men and women. 80-85% of the lung cancer is of non-small cell lung cancer (NSCLC) and rest is contributed by small cell lung cancer (SCLC). Genetic variation in DNA repair genes has been reported to affect the chemotherapeutic response among non-small cell lung cancer patients. XPD gene play an important role in Nucleotide Excision repair pathway respectively. In the present study, deleterious SNPs of XPD gene was predicted based on consensus prediction tools and then deleterious SNPs was analyzed by different computational tools to know their effects of structure and function of XPD protein. Nineteen out of 174 clinically relevant SNPs were commonly found to be deleterious by three prediction tools. Fourteen SNPs were present inside of protein domain and five were outside of protein coding region. Sixteen SNPs commonly found to decrease the stability of protein by two tools (I-mutant and MuPro). Seventeen SNPs were prediction to be disease related by Phd-SNP and SNP&GO. RMSD value of mutant model generated by considering change due to 19 SNPs was found to be in range of 0.970 to 0.979 representing not much protein structural variation. Hence, it is suggested that these 19 SNPs of XPD gene may be associated with cancer development.

Keywords: NSCLC, SCLC, XPD, SNP, RMSD, PYMOL, SNP&GO, Phd-SNP.

Chapter

Introduction

Majorly lung cancer is caused by tobacco smoking. Tobacco smoking is associated with the hazard of cancer of larynx, mouth, urinary bladder and kidney [Gazdar et al. 2018]. In most countries, lung cancer is most prevalent in men. In 2012 lung cancer accounts for 17% in men excluding non-melanoma skin cancer and for 9% in women [Ferlay et al. 2015].

XPD is a protein participate in transcription-coupled nucleotide excision repair [Rudolf et al. 2008]. XPD protein has 760 amino acids. XPD gene consists of 22 exons and 21 introns. XPD gene is associated with three disorders viz,: Cockayne syndrome, Xeroderma pigmentosum, trichothiodystrophy which are distinguish by cancer proneness, mental and growth retardation, high ultraviolet-light hypersensitivity [Sarasin et al. 2002].

The two variant allele of XPD exon 23 and exon 10 have been bridge with the high risk of lung cancer [Johnson et al. 2008]. It was studied that the risk of lung cancer in heterozygous carriers of Lys751Gln were at 1.64-fold and homozygous carriers were at 2.01-fold. Carriers of Asp312Asn were also at higher risk than homozygous carriers of wild type allele.

Single nucleotide polymorphisms prevalently called SNPs. In people SNPs are the most common type of genetic variation. SNPs also give the powerful tools for a variety of medical genetic studies [Abdelraheem et al. 2016]. Many missense substitutions are identified in single nucleotide polymorphisms. There is recent development in bioinformatics tools in analyzing SNPs for their association in different disease with special focus on cancer.

In the current study, a different prediction tools were used to find the SNPs which are deleterious in nature. Further, these SNPs were analyzed by different computation tools for their effect on structure and function of XPD protein.

Chapter 2

Review of literature

2.1 Lung cancer

2.1.1 Introduction

In industrialized countries lung cancer is the most lethal forms of cancer. In men it was second and in women it was fourth most common cancer in Finland [Minna et al. 2011]. In United states, major cause of the lung cancer is death [John Abbotts et al. 1984]. Immunotherapy is the recent field for the treatment of lung cancer. Immune- related adverse events (irAEs) are the classical auto-immune side effects for the treatment of lung cancer, which can also affects multiple organs [Inthasot et al. 2019]. Lung cancer found to be the most common cause of mortality worldwide. Major symptoms of lung cancer include weight loss, chest, discomfort, loss of appetite, coughing, shortness of breath and drastic change in voice etc [Loeb et al. 1984].

Types of cancer

It is of two main types of lung cancer-**Small cell** and **Non-small cell**.

Small cell lung cancer (SCLC)

Only 15% accounts SCLC of all lung cancers. They mostly originate in neuroendocrine cells. It is of major classes. One is small cell carcinoma (oat cell cancer) along with second is combined cell carcinoma. Middle-age patients who were initially detected by diabetes mellitus with no sign of lung cancer. By SCLC the patients were finally diagnosed and it was confirmed by bronchoscopic biopsy [Wang et al. 2019].

Non-Small cell lung cancer (NSCLC)

Only 85% accounts NSCLC of all lung cancers [Sushama et al. 2011]. In lung cancer it is most numerous type- NSCLC is grouped into three types: adenocarcinoma, squamous cell carcinoma and large-cell (undifferentiated) as shown in figure2.1. For all lung cancer- Squamous cell carcinoma accounts of 25 to 30%. It mainly occurs in centre of chest area in bronchi [Tod et al. 2008] and Large cell tumor occurs from the epithelial cells of the lung [Larsen et al 2011]. Adenocarcinoma is the most common type of NSCLC, it is associated with the smokers as well as non smokers.

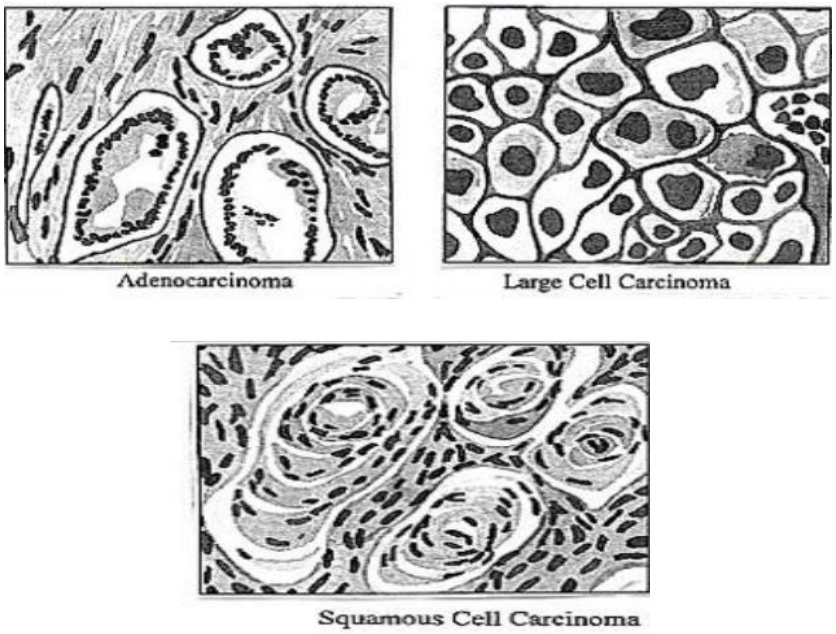


Figure 2.1: Representing microscopic images of different histology of lung cancer

(Source:http://www.immunerecovery.net/lung_cancer.html)

Epidemiological factor

Smoking

One of the major cause of lung cancer is smoking tobacco, especially cigarettes . Low-tar and low-nicotine cigarettes, , herbal cigarettes, chewing tobacco are the types of tobacco by-products. It also cause cancer [Loeb et al. 1984] 4000 active chemical compounds are present in Cigarette smoking and around 60 compounds are carcinogensis. The first paper were established in the landmark 1955 [Gaborieau et al. 2001] to indicate the risk of bidi smoking to lung cancer. Other major risk factors of lung cancer are drinking alcohol, coffee as well as consumption of red meat [B Ganesh et al. 2011].

Occupational exposure to chemical carcinogens

Occupational risk of lung cancer are Asbestos, Nickel Refinery workers, Haematie mining, Hard rock mining, Radiations, Chromium, Ethers and Mustard gas, Arsenic, chloromethyl, oils and coke exposures etc [Swaen et al. 1997].

Symptom of lung cancer [Minna et al. 2011]

Major clinical sign and symptom which may implicate occurrence of lung cancer are:

- Cough

Diagnosis of lung cancer [Allmark et al. 2008]

In order to detect lung cancer and to determine stage of cancer, several tests are recommended which includes:

- CT scan for lung cancer staging
- CBC, chest X-ray
- Physical examination

Trends of lung cancer worldwide

In the year 2012, there were 14.1 million cancer cases around the world and out of these 7.4 million cases were in men and 6.7 million in women. 6.9% of all new cancer cases diagnosed in India. In 2012 1.8 million cases were diagnosed. The assumption cause of lung cancer is smoking and it is thought to be responsible for 85% of all types of this cancer.

In several countries in the Americas, smoking frequency is around 20-30% in men and 10-20% in women while in Central America, frequency is lower mostly in women. In many South American and Caribbean countries the tobacco epidemic initially a little decades behind than in Northern America. Either Northern European countries have been able to considerably reduce smoking frequency in both men and women [Dikshit et al. 2015]. For men the smoking frequency did not decrease in many countries in South-Eastern Asia, Western Asia and Central Asia. In African men the smoking frequency is lower than in men from other section of the world [Jemal et al. 2015]. Trends in mortality rates among males for selected countries as shown in figure 2.2. Trends in mortality rates among females for selected countries as shown in figure 2.3.

For women, Denmark had the highest rate of lung cancer followed by USA and Canada. In less developed countries 54% of lung cancer cases occured. More cases of lung cancer arised in Northen America and in Oceania and less in Africa, Latin America and Caribbean.

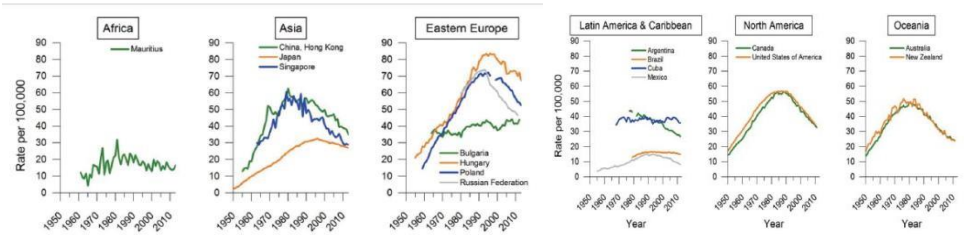


Figure 2.2 Trends in mortality rates among males for selected countries (1950-2013).

(Source: WHO IARC Cancer Mortality Database)

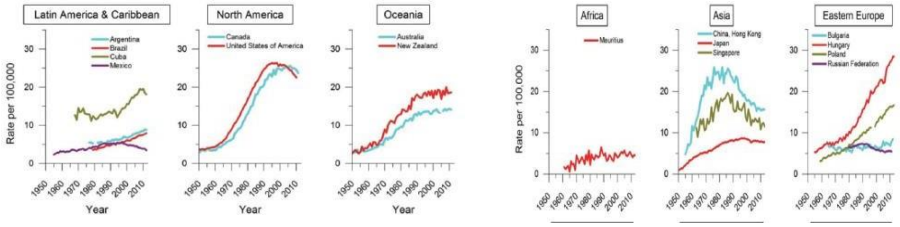


Figure2.3 Trends in mortality rates among females for selected countries (1950-2013)

(Source: WHO IARC Cancer Mortality Database)

Lung cancer stages

There are four stages of lung cancer. The size of tumor is 3cm at stage I and there is no metastasis. The size of tumor is 6cm at stage II then metastasis begins. At the stage of III, size of tumor is become more than 6cm and metastasis mainly reach in the lymph node. Last the tumor migrates to other organs of the body at stage IV (Source: www.beaconhospital.com.my/lung-cancer). At stage I there is highest survival rates (71%). Survival rate is much lower at stage IV disease (14%).

Cisplatin

Lung cancer is the most frequent cancer now a day's worldwide. The two most common platinum based chemotherapy for NSCLC is cisplatin and carboplatin. Cisplatin is most commonly used treatment as compared to carboplatin because it has an excellent anti-tumor activity. Although it has many side-effects like vomiting and nausea etc. at stage II and III of non-small cell lung cancer [Florea et al. 2011].

2.2 XPD

2.2.1 Introduction

XPD is also known as ERCC2. It belongs to the RAD3/XPD subfamily of helicases. It is a part of transcriptional factor IIH as well as helicase XPB in eukaryotes. TFIIH is necessary for the Nucleotide Excision Repair (NER) pathway. For transcription initiation, helicase activity of XPD is not necessary. XPD gene is associated with three disorders: Cockayne syndrome, Xeroderma pigmentosum, trichothiodystrophy which are distinguish by cancer proneness, mental and growth retardation, high ultraviolet-light hypersensitivity [Sarasin et al. 2002].

Helicase c2, DEAD_2, DUF1227 are the three domains of XPD. In helicase c2 domain, larger numbers of deleterious SNPs are found in it. There are five residues selected by XP-causing mutations [Johnson et al. 2008].

Protein structure

XPD gene is located on human chromosome 19q13.32. It has 22 exons and 2.3-kb long coding region. XPD protein inclusive of 760 amino acids. It joins with Xeroderma Pigmentosum complementation group D and forms the ERCC2-XPD complex as shown in figure 2.4. In the XP-D group, several different mutations have been detected [Stary et al. 2002]. XPD requires to be not only available but also enzymatically active for its DNA repair function. Structure of XPD Protein as shown in figure 2.4. It plays role in sustaining the stability of the TFIIH complex and also in nucleotide excision repair [Spitz et al. 2001].

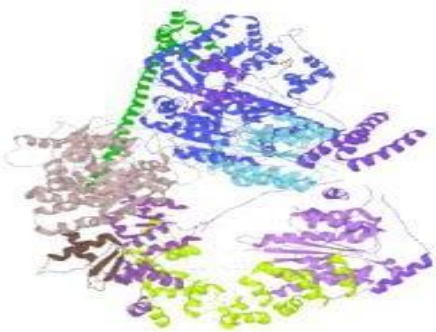


Figure2.4: Structure of XPD Protein

(Source:https://www.uniprot.org/uniprot/P18074)

Role of XPD within the TFIIH Complex

At first sight, the role of TFIIH in transcriptional to be completely equivalent.. In transcription, at the promoter site it is essential for opening up of the DNA. DNA repair is based on the helicase activity of XPD while transcription is not [Amos et al. 2001].

NER pathway

Nucleotide excision repair pathway repairs the non-specific DNA damage which are caused by chemical inter- or intra- strand additional compound formation, gamma or ultra violet radiation and cross-linking. NER removes the structurally unrelated DNA damage. Cyclobutane pyrimidine dimer and photoproducts contributes two most common types of damage in DNA which is corrected with the help of NER pathway. Mechanism of NER pathway as shown in figure 2.5. There are six genes which encodes for the proteins that has an important role in NER pathway and these genes XPD, ERCC1,ERCC3, ERCC4, ERCC [Friedberg et al. 2001].

NER

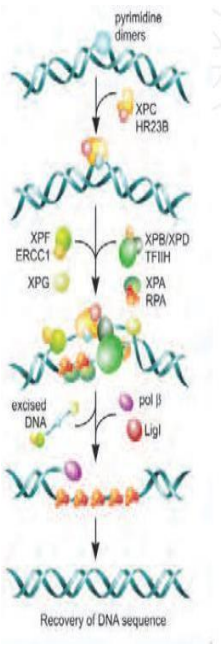


Figure 2.5: Mechanism of NER pathway

(Source: Mladenov& Iliakis,2011)

Relation of XPD with Lung cancer

In the XPD gene two single nucleotide polymorphisms (SNP) have been studied. In exon 10 XPD Asp321Asn causes an amino acid substitution in a sustain region of XPD. In exon 23 XPD Lys751Gln causes an amino acid substitution in the C-terminal section of the protein. The existence of the variant allele XPD exon23 and exon10 has been connected with high risk of lung cancer [Wang et al. 2001]. The two variant alleles in XPD Asp312Asn and Lys751Gln are also connected with low DNA repair capacity [Spitz et al. 2001]. It was studied and established that the risk of lung cancer in heterozygous carriers of Lys751Gln were at 1.64- fold and in homozygous carriers were at 2.01-fold. Carriers of variant allele Asp312Asn were also at higher risk of lung cancer than homozygous carriers of the wild type allele.

Single nucleotide polymorphisms (SNPs)

Sequence variation in genomes is the type of single nucleotide polymorphisms (SNPs). Single nucleotide polymorphism is defined as change in base of the DNA whether adenine, guanine, cytosine or thymine. In single nucleotide polymorphism (SNP) data many missense substitutions are identified [Abdelmoneim et al. 2019]. SNP may exist within coding region. Non-synonymous SNP changes the phenotypic effect of the genes. So to check or predict whether the SNP lies within coding regions , non-coding region or among coding and non-coding region, whether SNP causes any phenotypic change in the gene or whether the SNP contributes to benign or malignant tumor, various bioinformatics tools are there.

Bioinformatics tools

Bioinformatics tools helps to research and determine the SNPs for drug discovery [Abdelhameed et al. 2019]. Sequence and structure based method are used in SNP prediction tools. Bioinformatics tools are SIFT, PROVEAN, I-MUTANT, POLYPHEN2, MuPro, Phd-SNP, SNP&GO, ELASPIC, Mutation 3D, Genemania, PDBsum, SWISS-model and PYMOL.

SIFT- It stands for Separating Intolerant from the Tolerant. SIFT mainly predicts the phenotypic effect of the genes. It is sequence based software tool. Its input is a query sequence and it uses

that query sequence on the basis of multiple alignments and then it gives a score. This score mainly predict whether the SNP has deleterious or not. If the score is equals to or less than 0.05 then the SNP is consider as deleterious and if the score is more than 0.05 then the SNP is tolerated. SIFT is present as an online tool (http://sift-dna.org)[Muthusamy et al. 2016].

PROVEAN- It stands for protein variation effect analyser. Score predicts whether SNP has functional effect or not. PROVEAN tool is available online (http://provean.jcvi.org.)[Mustafa et al. 2019].

POLYPHEN2- It stands for polymorphism phenotyping v2. This bioinformatics tool uses comparative and structural evolutionary relationships and predicts whether stability and function of the protein affected or not. It is based on the protein sequence alignment, 3-D structure of the protein and several machine learning programmes. Multiple sequence alignment is based on position specific independent counts and calculates the profile matrix (score), which is the likelihood log ratio of one amino acid at a particular position to another amino acid occuring at same position. It is analyzing the data in large volume. Its tool is available online (http://genetics.bwh.harvard.edu/pph2/)[Alla et al. 2016].

I-MUTANT- It is depend on the support vector machine and predicts the stability of protein when SNP occur. Reliability index (measure of consistency of score) should be 0<RI>10 and also to calculate Gibbs free energy change (DDG). If the value of DDG is lower than zero then the protein stability decreases and if it is more than zero then protein stability increases and if it is equal to zero then it is neutral [Muthusamy et al. 2016].

MuPro- It is based on support vector machine as well as neutral network. It is more accurate and this predicts the stability of proteins and it does not require tertiary structures to predict stability of proteins. If the value of confidence score is lower than zero then the protein stability decreases and if the value of confidence score is more than zero then the protein stability increases. It is available at (http://mupro.proteomics.ics.uci.edu/)[Elgemaabi et al. 2016].

PhD-SNP- It is based upon support vector machine. This predicts whether the mutation is disease related or neutral. Reliability index (measure of consistency of score) should be 0<RI>10 and calculated from the output of support vector machine(O). It is available at (http://snps.uib.es/phd-snp/phd-snp.html) [Hassan et al. 2016].

SNP&GO- This tool is also based upon the support vector machine and also predicts whether the mutation is disease related or neutral [Ismail et al. 2016].

ELASPIC- It is based on the esemble machine learning programme. It predicts whether protein-protein interaction and protein folding gets affected by the change in nucleotide. This predicts with the help of homology model of domains and domain-domain interactions. It is available at (http://elaspic.kimlab.org/many/.)[Nagamani et al. 2016].

Mutation 3D- It is helpful in the visualization and functionality prediction of protein when there is any change in the nucleotide. Visualisation involves the arrangement of amino acid substitution spatially on protein model. It is available at (http://mutation3d.org/index.shtml.) [Osman et al. 2016].

Genemania- This predicts the set of data which includes co-localization, protein interaction, genetic interaction, co-expression, pathways and similarity in protein domain etc. related to one gene or set of genes and also predicts genes which are related to input gene. It is available at (http://www.genemania.org/.)[Abdelraheem et al. 2016].

PDBsum- It is a web-based database. It predicts the graphic summary on each macromolecular structure placed at the Protein Data Bank (PDB). It is available at (http://www.biochem.ucl.ac.uk/bsm/pdbsum)[Laskowski et al. 2001].

SWISS-MODEL- It is an automated comparative protein modeling server. This tool automatically generated protein modeling of three-dimensional (3D) protein structures. It is available at (http://swissmodel.expasy.org)[Schwede et al. 2003].

PYMOL-It is helpful in the visualization and pymol generate 3D structures of small molecules and biological macromolecules such as proteins. This tool also predicts the value of RMSD. It is available at (https://pymol.org/buy)[Danielson et al. 2011

Chapter 3

Objectives

- Identification of deleterious single nucleotide polymorphisms (SNPs) in XPD gene involved in lung cancer based on consensus prediction algorithms.
- Computational analysis of deleterious (SNPs) for their effect on structure and function of XPD protein.

Chapter 4

Material and Methods

4.1 Bioinformatics methods

Computational analysis of harmful single nucleotide polymorphisms (SNPs) in human XPD gene involved in lung cancer using multiple softwares.
SNPs were downloaded from the NCBI database (https://www.ncbi.nlm.nih.gov/snp).
SIFT, PROVEAN, I-MUTANT, POLYPHEN2, MuPro, Phd-SNP, SNP&GO, ELASPIC, Mutation 3D, Genemania, PDBsum, SWISS-model and PYMOL, these tools were used to predict the deleterious SNPs of XPD.

4.1.1 Prediction of structural impact of nsSNP on protein.

This was performed with the help of SIFT software
FOR XPD rs 41556519
Input: rsID (rs41556519)

Figure 4.1 Input of SIFT software for XPD (rs 41556519)

(Source :http://sift.jcvi.org)

4.2 Prediction of functional effect of amino acid substitution.

For this provean software was used

Input: Protein ID, Position at which amino acid changes and amino acid changes into which new residue (from SIFT software output)

For XPD rs 41556519

Input: rsID (rs41556519)

Figure 4.2 Input of PROVEAN software for XPD(rs41556519)

(Source:http://provean.jcvi.org)

4.3 Prediction of deleterious SNP by POLYPHEN2 software.

For XPD rs 41556519

Input: rsID (rs 41556519)

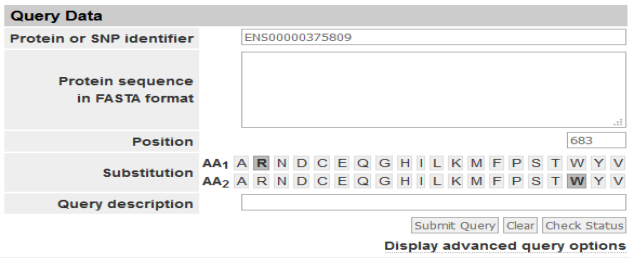


Figure 4.3 Input of POLYPHEN2 software for XPD (rs 41556519)

(Source:http://genetics.bwh.harvard.edu/pph2/)

4.4 To predict nsSNP impact on stability of protein with the help of I-Mutant software.

INPUT: Protein sequence in FASTA format (retrieved from Swissprot), Position at which amino acid changes and amino acid changes into which new residue (from SIFT software output), temperature and pH.

Protein sequence of XPD in FASTA format

>sp|P18074|XPD_HUMAN RecName: Full= Xeroderma pigmentosum group D; Short= XPD

MKLNVDGLLVYFPYDYIYPEQFSYMRELKRTLDAKGHGVLEMPSGTGKTVSLLALIMA
YQRAYPLEVTKLIYCSRTVPEIEKVIEELRKLLNFYEKQEGEKLPFLGLALSSRKNLCIHP
EVTPLRFGKDVDGKCHSLTASYVRAQYQHDTSLPHCRFYEEFDAHGREVPLPAGIYNLD
DLKALGRRQGWCPYFLARYSILHANVVVYSYHYLLDPKIADLVSKELARKAVVVFDEA
HNIDNVCIDSMSVNLTRRTLDRCQGNLETLQKTVLRIKETDEQRLRDEYRRLVEGLREA
SAARETDAHLANPVLPDEVLQEAVPGSIRTAEHFLGFLRRLLEYVKWRLRVQHVVQESP
PAFLSGLAQRVCIQRKPLRFCAERLRSLLHTLEITDLADFSPLTLLANFATLVSTYAKGFT
IIIEPFDDRTPTIANPILHFSCMDASLAIKPVFERFQSVIITSGTLSPLDIYPKILDFHPVTMAT
FTMTLARVCLCPMIIGRGNDQVAISSKFETREDIAVIRNYGNLLLEMSAVVPDGIVAFFTS
YQYMESTVASWYEQGILENIQRNKLLFIETQDGAETSVALEKYQEACENGRGAILLSVA
RGKVSEGIDFVHHYGRAVIMFGVPYVYTQSRILKARLEYLRDQFQIRENDFLTFDAMRH
AAQCVGRAIRGKTDYGLMVFADKRFARGDKRGKLPRWIQEHLTDANLNLTVDEGVQV
AKYFLRQMAQPFHREDQLGLSLLSLEQLESEETLKRIEQIAQQL

(Source:https://www.ncbi.nlm.nih.gov/protein/)

For XPD rs 121913018

Figure 4.4 Input of I-Mutant software for XPD (rs 121913018)

(Source:http://folding.biofold.org/cgi-bin/i-mutant2.0.cgi)

4.5 Prediction of stability of protein on the basis of change in energy by MuPro software.

INPUT: Mutation position, original amino acid, mutated amino acid, protein sequence in FASTA format. Mutation position, original amino acid and mutated amino acid was fetched back from SIFT software output and protein sequence was retrieved from Swissprot.

For XPD rs201392911

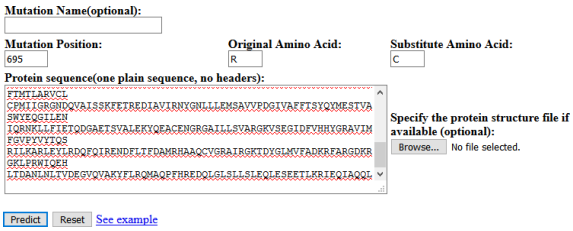


Figure 4.5 Input of MuPro software for XPD (rs201392911)

(Source:http://www.ics.uci.edu/~baldig/mutation.html)

4.6 Prediction of SNPs to know that the mutation is disease related or neutral.

This was performed by PhD-SNP and SNP&GO software

INPUT: Protein sequence in FASTA format, Swiss-prot Code for protein, mutation position, mutated amino acid and basis of prediction. Protein sequence and code for protein was retrieved from Swissprot. Mutation position, mutated amino acid was fetched back from SIFT software output.

For XPD rs201392911

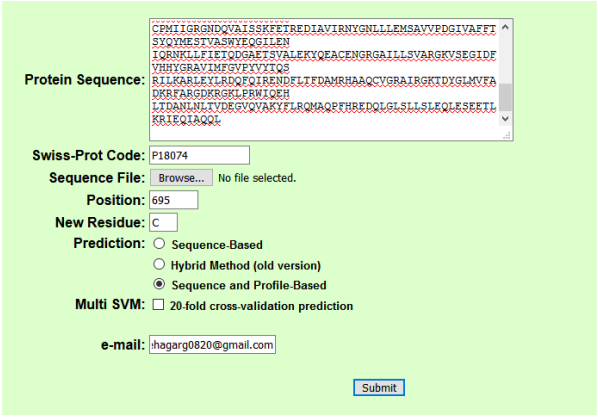


Figure 4.6 Input of Phd-SNP software for XPD(rs201392911)

(Source:http://snps.biofold.org/phd-snp/phd-snp.html)

INPUT: UNIPROT accession Number for protein, Mutation position, original amino acid, Mutated amino acid. Accession number was fetch from Swissprot website and mutation position, original amino acid, mutated amino acid was retrieved from SIFT software output.

FOR XPD rs201392911

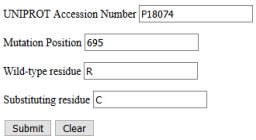


Figure 4.7 Input of SNP&GO software for XPD (rs201392911)

(Source:https://snps-and-gp.biocomp.unibo.it/snps-and-go/)

4.6 Prediction of effects of mutations on folding of protein and protein-protein interactions by ELASPIC software.

INPUT: name of gene_organism_original amino acid changes at which position into what kind of new amino acid

For XPD rs201392911

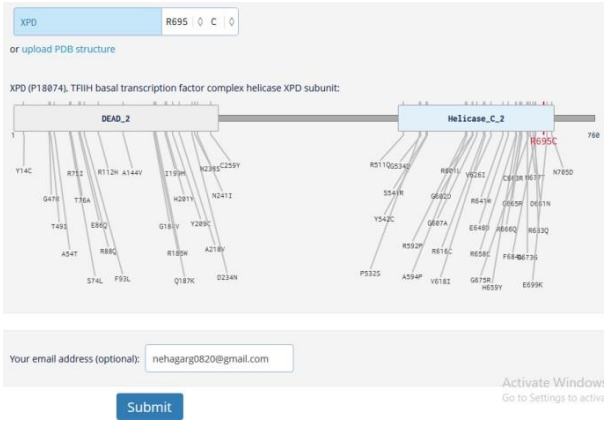


Figure 4.8 Input of ELASPIC software for XPD(rs201392911)

(Source:http://elaspic.kimlab.org/many/)

4.7 Protein 3D modeling by SWISS-model.

INPUT: Protein sequence in FASTA format and was retrieved from Swissprot

Protein sequence of XPD in FASTA format

MKLNVDGLLVYFPYDYIYPEQFSYMRELKRTLDAKGHGVLEMPSGTGKTVSLLALIMA
YQRAYPLEVTKLIYCSRTVPEIEKVIEELRKLLNFYEKQEGEKLPFLGLALSSRKNLCIHP
EVTPLRFGKDVDGKCHSLTASYVRAQYQHDTSLPHCRFYEEFDAHGREVPLPAGIYNLD
DLKALGRRQGWCPYFLARYSILHANVVVYSYHYLLDPKIADLVSKELARKAVVVFDEA
HNIDNVCIDSMSVNLTRRTLDRCQGNLETLQKTVLRIKETDEQRLRDEYRRLVEGLREA
SAARETDAHLANPVLPDEVLQEAVPGSIRTAEHFLGFLRRLLEYVKWRLRVQHVVQESP
PAFLSGLAQRVCIQRKPLRFCAERLRSLLHTLEITDLADFSPLTLLANFATLVSTYAKGFT
IIIEPFDDRTPTIANPILHFSCMDASLAIKPVFERFQSVIITSGTLSPLDIYPKILDFHPVTMAT
FTMTLARVCLCPMIIGRGNDQVAISSKFETREDIAVIRNYGNLLLEMSAVVPDGIVAFFTS
YQYMESTVASWYEQGILENIQRNKLLFIETQDGAETSVALEKYQEACENGRGAILLSVA
RGKVSEGIDFVHHYGRAVIMFGVPYVYTQSRILKARLEYLRDQFQIRENDFLTFDAMRH
AAQCVGRAIRGKTDYGLMVFADKRFARGDKRGKLPRWIQEHLTDANLNLTVDEGVQV
AKYFLRQMAQPFHREDQLGLSLLSLEQLESEETLKRIEQIAQQL

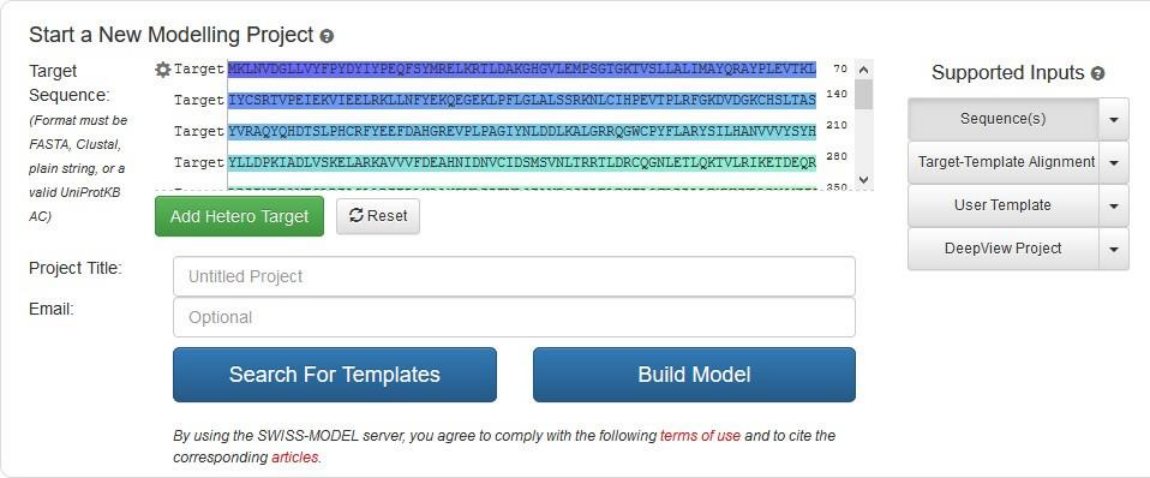


Figure 4.9 Input of SWISS-model for XPD

(Source:https://swissmodel.expasy.org)

4.8 Prediction of spatial arrangement of amino acid substitution on protein structures and domains with the help of Mutation 3D software.

INPUT:

Protein ID: retrieved from UNIPROT

Mutation: |wild type amino acid||position of mutation||mutated amino acid| For XPD (rs201392911)

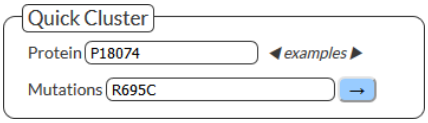


Figure 4.10 Input of Mutation 3D software for XPD (rs201392911)

(Source:http://mutation3d.org/index.shtml)

4.9 Determination of RMSD value by PYMOL software.

INPUT:

Mutated model aligned on wild type model. For XPD (rs201392911)

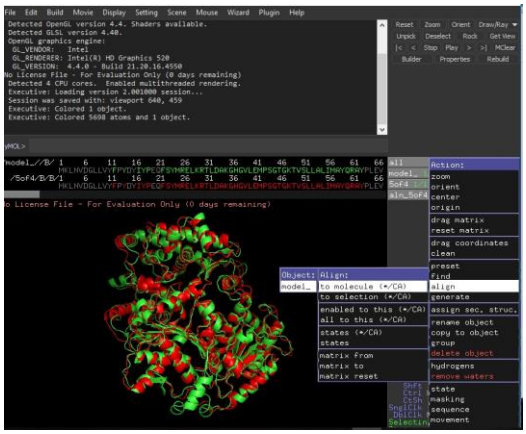


Figure 4.11 Input of PYMOL for XPD(rs201292911)

(Source:https://pymol.org/buy)

4.10 RAMCHANDRAN Plot by PDBsum software.

INPUT:

Email ID

Model: retrieved from SWISS-Model (PDB-format)

For XPD (rs376556895)

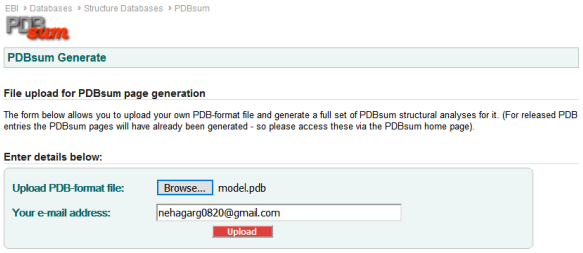


Figure 4.12 Input of PDBsum software for XPD(rs376556895)

(Source:http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/generate.html)

4.11 Prediction of interaction of gene with other genes with the help of GENEMANIA software.

INPUT:

For XPD

For all networks

Figure 4.13 Input of GENEMANIA software for XPD

(Source:http://www.genemania.org/)

Chapter 5

Results

5.1 Selection of SNPs.

There were 8242 SNPs for XPD present in NCBI SNP database and the data of 174 SNPs for clinical significance is present which was selected for further computational analysis (Table 5.1).

Table 5.1 Clinical significance of XPD gene

Benign	68
Likely benign	7
Likely pathogenic	3
Pathogenic	16
Uncertain significance	39
Untested	41

5.2 Identification of deleterious SNPs

Three predictions tools (SIFT, PROVEAN and Polyphen2) for identifying the SNPs which are deleterious in nature. 19 and 155 SNPs were reported to be were deleterious and tolerated by SIFT. SIFT identified 19 deleterious SNPs are mentioned in the (Table 5.2).

PROVEAN

Among 19 SNPs of the XPD were deleterious and for rest SNPs was neutral. Deleterious SNPs as shown in table 5.2.

Table 5.2 Prediction of deleterious SNP by using SIFT and PROVEAN

Sr. No.	rs ID	Amino acid change	Score	SIFT Prediction	Score	PROVEAN Prediction
1	rs140522180	R601Q	0	Damaging	-3.81	Deleterious
2	rs147224585	R592H	0	Damaging	-4.69	Deleterious
3	rs199738290	R497C	0	Damaging	-6.61	Deleterious
4	rs200665173	Q629H	0.031	Damaging	-3.28	Deleterious
5	rs201370106	L744P	0	Damaging	-4.48	Deleterious
6	rs201392911	R695C	0.003	Damaging	-4.77	Deleterious

7	rs137910235	R227C	0.001	Damaging	-5.21	Deleterious
8	rs200895828	V231M	0	Damaging	-2.76	Deleterious
9	rs370454709	C259Y	0	Damaging	-3.69	Deleterious
10	rs372176415	M247T	0	Damaging	-4.63	Deleterious
11	rs41556519	R683W	0	Damaging	-7.31	Deleterious
12	rs121913018	A725P	0.001	Damaging	-4.1	Deleterious
13	rs121913019	S541R	0.001	Damaging	-4.76	Deleterious
14	rs121913020	R112H	0	Damaging	-4.53	Deleterious
15	rs121913021	R658C	0	Damaging	-7.29	Deleterious
16	rs121913023	D681N	0.001	Damaging	-4.67	Deleterious
17	rs121913024	R616W	0	Damaging	-7.52	Deleterious
18	rs121913026	R722W	0	Damaging	-6.66	Deleterious
19	rs376556895	R616P	0	Damaging	-6.53	Deleterious

Polyphen2

Among 8242 SNP of XPD there were only 174 SNP whose data was present on Polyphen2 software and for rest SNP no data was available. Among 14 SNPs were probably damaging, 5 SNPs were Possibly damaging and for rest SNPs was benign. Probably damaging and possibly damaging SNPs as shown in (Table 5.3).

Table 5.3 Prediction of deleterious SNP by using Polyphen2

Sr No.	rs ID	Score	Sensitivity	Specificity	Prediction
1	rs140522180	1	0	1	Probably damaging
2	rs147224585	1	0	1	Probably damaging
3	rs199738290	0.740	0.85	0.92	Possibly damaging
4	rs200665173	0.738	0.85	0.92	Possibly damaging
5	rs201370106	0.998	0.27	0.99	Probably damaging
6	rs201392911	0.995	0.68	0.97	Probably damaging
7	rs137910235	0.991	0.71	0.97	Probably damaging
8	rs200895828	0.999	0.14	0.99	Probably damaging
9	rs370454709	0.881	0.82	0.94	Possibly damaging
10	rs372176415	0.067	0.94	0.84	Possibly damaging

11	rs41556519	1	0	0	Probably damaging
12	rs121913018	0.997	0.41	0.98	Probably damaging
13	rs121913019	0.934	0.80	0.94	Possibly damaging
14	rs121913020	1	0	1	Probably damaging
15	rs121913021	1	0	1	Probably damaging
16	rs121913023	0.999	0.14	0.99	Probably damaging
17	rs121913024	1	0	1	Probably damaging
18	rs121913026	0.999	0.14	0.99	Probably damaging
19	rs376556895	1	0	1	Probably damaging

5.3 Effect of SNPs on the protein stability

I- Mutant and MuPro prediction tools were used to assess the effect of SNPs on the stability of protein.

Among 19 Deleterious SNP of XPD gene, all the SNPs were observed within reliability index and DDG (Difference in delta G of mutant genotype and wild genotype) value in which 18 SNPs had decreased protein stability and only 1 SNPs had increased protein stability by I- Mutant as shown in (Table5.4).

All the 19 SNPs were observed within confidence score in MuPro Prediction where 16 SNPs had decreased protein stability and 3 SNPs had increased protein stability as shown in (Table 5.4).

Table 5.4 Prediction of protein stability by I-mutant and MuPro

			I-Mutant Prediction			MuPro Prediction	
Sr no	rs ID	Amino acid change	RI	DDG	Stability	Confidence score	Stability
1	rs140522180	R601Q	9	-1.16	Decrease	-1	Decrease
2	rs147224585	R592H	9	-1.29	Decrease	-0.93536413	Decrease
3	rs199738290	R497C	5	-1.07	Decrease	-0.80065765	Decrease
4	rs200665173	Q629H	9	-1.09	Decrease	0.24732523	Increase
5	rs201370106	L744P	2	-0.96	Decrease	-0.97390074	Decrease
6	rs201392911	R695C	6	-0.87	Decrease	-0.68175079	Decrease

7	rs137910235	R227C	6	-0.82	Decrease	-1	Decrease
8	rs200895828	V231M	7	-0.86	Decrease	-0.28364455	Decrease
9	rs370454709	C259Y	0	-0.18	Decrease	-0.63827701	Decrease
10	rs372176415	M247T	8	-1.62	Decrease	-1	Decrease
11	rs41556519	R683W	6	-0.46	Decrease	-0.94580477	Decrease
12	rs121913018	A725P	2	-0.32	Decrease	0.63096563	Increase
13	rs121913019	S541R	4	0.08	Increase	0.47600521	Increase
14	rs121913020	R112H	9	-1.44	Decrease	-1	Decrease
15	rs121913021	R658C	4	-0.71	Decrease	-1	Decrease
16	rs121913023	D681N	5	-1.07	Decrease	-1	Decrease
17	rs121913024	R616W	4	-0.40	Decrease	-0.54819736	Decrease
18	rs121913026	R722W	4	-0.15	Decrease	-0.61064172	Decrease
19	rs376556895	R616P	5	-0.72	Decrease	-0.69210554	Decrease

5.4 Determination of SNPs to be disease related

PhD-SNP and SNP&GO software were used to determine whether SNPs were disease related or neutral. Among 19 Deleterious SNPs of XPD gene by PhD-SNP , all the SNPs were observed within reliability index in which the 17 SNPs were found to be disease related and two SNPs were neutral as shown in (Table 5.5**)**.

Among 19 Deleterious SNPs of XPD gene by SNP&GO, all the SNPs were observed within reliability index in which the 19 SNPs were found to be disease related as shown in (Table 5.5).

Table 5.5 Prediction of deleterious effect of SNP by using PhD-SNP and SNP&GO

			PhD-SNP Prediction		SNP&GO Prediction	
Sr no	rs ID	Amino acid change	RI	Prediction	RI	Prediction
1	rs140522180	R601Q	8	Disease	10	Disease
2	rs147224585	R592H	1	Disease	9	Disease
3	rs199738290	R497C	7	Disease	10	Disease
4	rs200665173	Q629H	5	Neutral	7	Disease
5	rs201370106	L744P	7	Disease	10	Disease
6	rs201392911	R695C	2	Disease	9	Disease
7	rs137910235	R227C	2	Neutral	9	Disease
8	rs200895828	V231M	6	Disease	9	Disease
9	rs370454709	C259Y	1	Disease	9	Disease
10	rs372176415	M247T	5	Disease	8	Disease
11	rs41556519	R683W	7	Disease	10	Disease
12	rs121913018	A725P	9	Disease	10	Disease
13	rs121913019	S541R	9	Disease	10	Disease

14	rs121913020	R112H	9	Disease	10	Disease
15	rs121913021	R658C	6	Disease	10	Disease
16	rs121913023	D681N	8	Disease	10	Disease
17	rs121913024	R616W	6	Disease	10	Disease
18	rs121913026	R722W	6	Disease	10	Disease
19	rs376556895	R616P	6	Disease	10	Disease

5.5 Determination of location of SNPs in protein domain

Elaspic and Mutation 3-D were to find the position of SNPs in protein domain. It was observed that 14 SNPs were within the protein domain and five were outside of protein domain by both softwares as shown in (Table 5.6) and Figure 5.3.

Table 5.6 Prediction of changes in protein structure in SNP by Elaspic

Sr no	rs ID	Amino acid change	DDG	Prediction
1	rs140522180	R601Q	0.411	Mutation falls inside of protein domain
2	rs147224585	R592H	0.720	Mutation falls inside of protein domain
3	rs199738290	R497C	-	Mutation falls outside of protein domain
4	rs200665173	Q629H	-0.114	Mutation falls inside of protein domain
5	rs201370106	L744P	-	Mutation falls outside of protein domain
6	rs201392911	R695C	0.472	Mutation falls inside of protein domain
7	rs137910235	R227C	0.453	Mutation falls inside of protein domain
8	rs200895828	V231M	-1.444	Mutation falls inside of protein domain
9	rs370454709	C259Y	-	Mutation falls outside of protein domain
10	rs372176415	M247T	-0.546	Mutation falls inside of protein domain
11	rs41556519	R683W	0.058	Mutation falls inside of protein domain
12	rs121913018	A725P	-	Mutation falls outside of protein domain
13	rs121913019	S541R	-0.673	Mutation falls inside of protein domain
14	rs121913020	R112H	-0.122	Mutation falls inside of protein domain

15	rs121913021	R658C	-0.016	Mutation falls inside of protein domain
16	rs121913023	D681N	0.570	Mutation falls inside of protein domain
17	rs121913024	R616W	-0.470	Mutation falls inside of protein domain
18	rs121913026	R722W	-	Mutation falls outside of protein domain
19	rs376556895	R616P	0.798	Mutation falls inside of protein domain

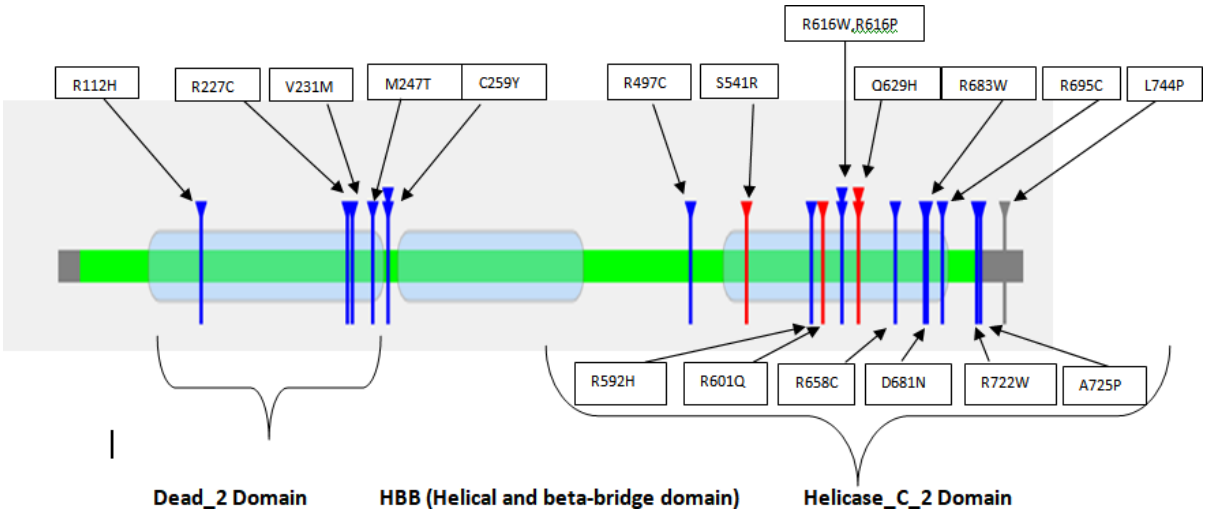


Figure 5.1 Position of mutations in protein domains of XPD

5.6 Generation of protein model and calculation of RMSD

The protein sequence of SNPs was changed as per each SNPs and then structure was generated by swiss model for each SNPs. Wild type and mutated model are represented in figure 5.1 and 5.2.

Among 19 Deleterious SNPs of XPD gene whose homology protein modeling was present on SWISS-model. Wild type protein modeling of XPD protein and mutated type protein modeling as shown in figures. RMSD value of each model was calculated by superimposing the model of each SNP with wild type model. It was observed that the RMSD value was found to be in range of 0.970 to 0.979 which represent that the structure is not much deviated from original structure.

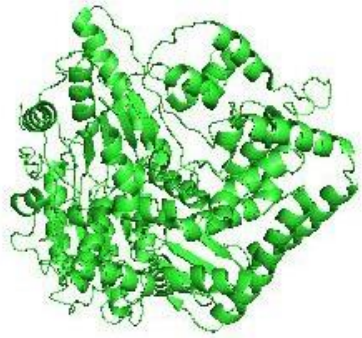
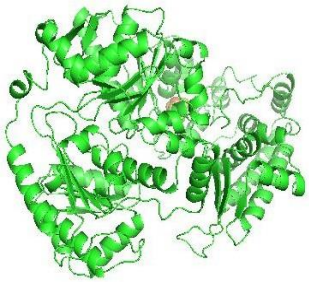


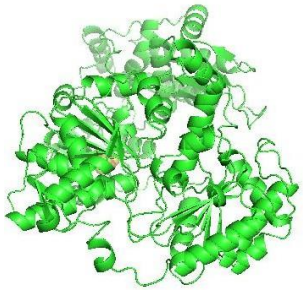
Figure 5.2 Protein model of Wild type genotype of XPD protein

Figure 5.3 Protein models of mutated type genotype of XPD gene

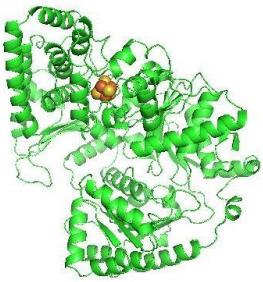
C259Y



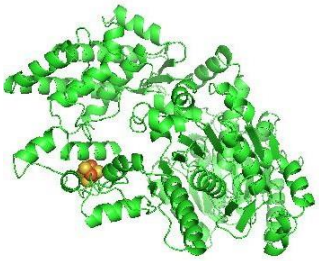
A725P



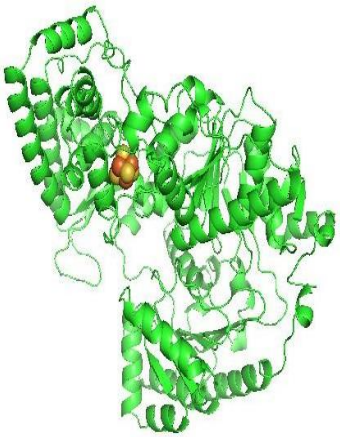
R683W



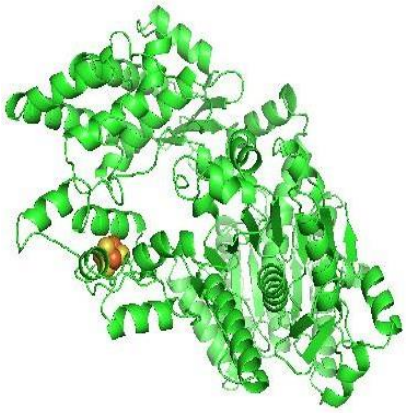
S541R



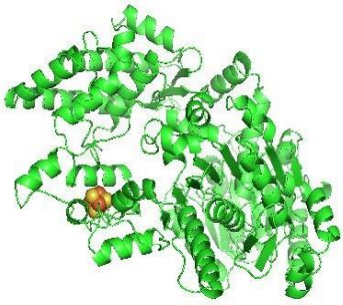
R658C



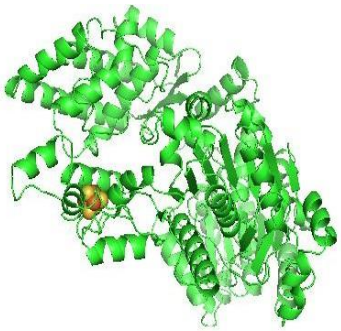
D681N



R616W

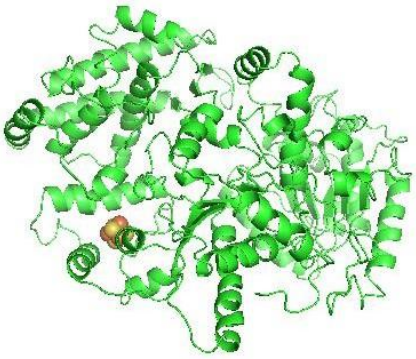


R722W

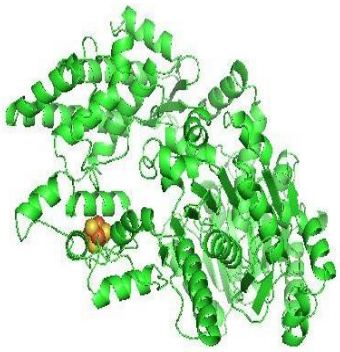


R616P

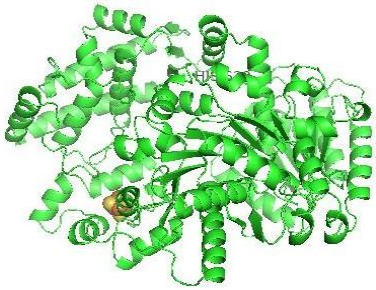
R227C



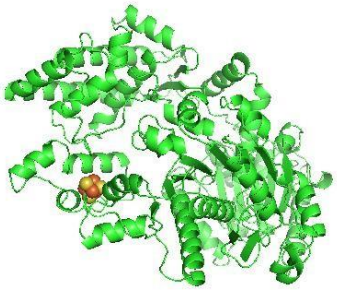
Q629H



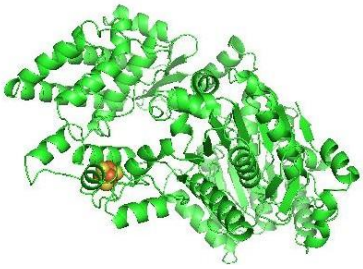
V231M



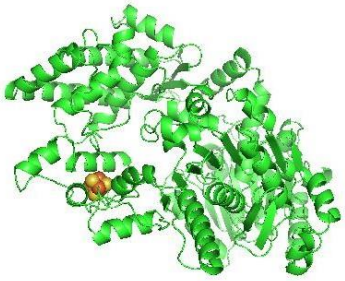
M247T



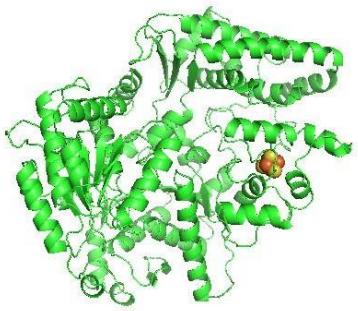
R601Q



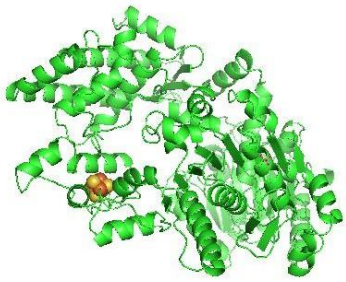
R592H



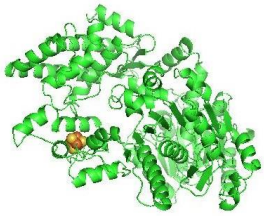
R497C



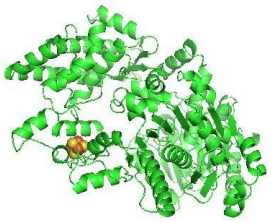
L744P



R695C



R112H



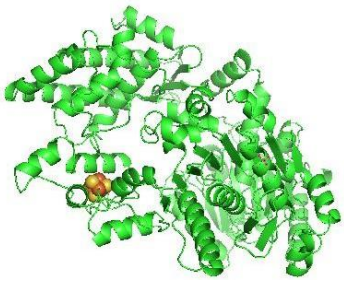


Table 5.7 Prediction of RMSD value of SNP by PyMOL software

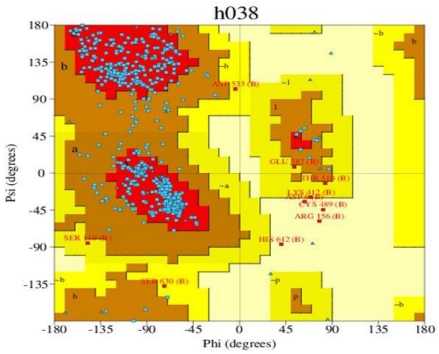
Sr no	rs ID	Amino acid change	RMSD(A˚)
1	rs140522180	R601Q	0.974
2	rs147224585	R592H	0.970
3	rs199738290	R497C	0.974
4	rs200665173	Q629H	0.979
5	rs201370106	L744P	0.975
6	rs201392911	R695C	0.970
7	rs137910235	R227C	0.970
8	rs200895828	V231M	0.974
9	rs370454709	C259Y	0.974
10	rs372176415	M247T	0.974
11	rs41556519	R683W	0.970
12	rs121913018	A725P	0.974
13	rs121913019	S541R	0.974
14	rs121913020	R112H	0.973

15	rs121913021	R658C	0.973
16	rs121913023	D681N	0.970
17	rs121913024	R616W	0.973
18	rs121913026	R722W	0.974
19	rs376556895	R616P	0.976

5.7 Determination of quality of mutated model by Ramachandran Plot

PDB sum was used to plot the Ramachandran plot of each SNP model as shown in (Table 5.5) and figure (5.4 and 5.5). There is not much variation in allowed regions of all 19 mutant model compared to actual PDB protein structure indicating that model generated was of good quality.

Figure 5.4 Wild type Ramachandran plot of XPD protein



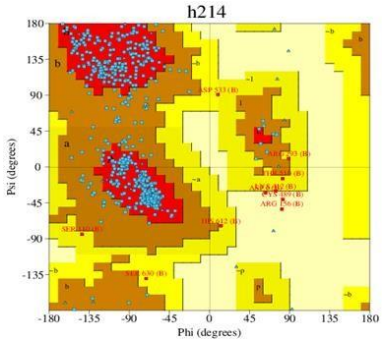
Most favoured regions 612(89.1%) residues

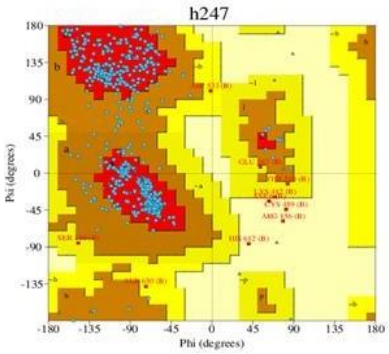
Additional allowed regions 65(9.5%) residues

Generously allowed regions 6(0.9%) residues

Disallowed regions 4(0.6%) residues

Figure 5.5 Ramachandran plot of XPD protein with R112H, C259Y, A725P, S541R



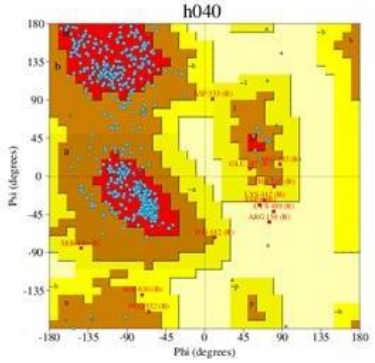


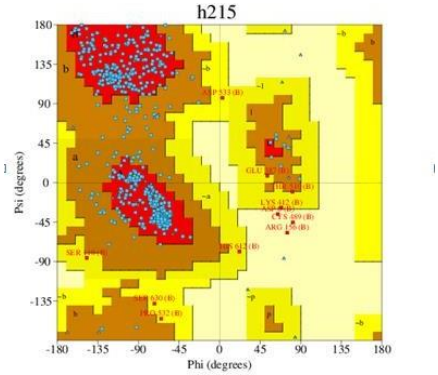
R112H

Most favoured regions 605(88.1%) residues
Additional allowed regions 72 (10.5%) residues
Generously allowed regions 6(0.9%) residues
Disallowed regions 4(0.6%) residues

C259Y

Most favoured regions 612 (89.1%) residues
Additional allowed regions 65(9.5%)residues
Generously allowed regions 6(0.9%) residues
Disallowed regions 4(0.6%) residues





A725P

Most favoured regions 607(88.5%) residues
Additional allowed regions 68(9.9%) residues
Generously allowed regions 8(1.2%) residues
Disallowed regions 3(0.4%) residues

S541R

Most favoured region 612(89.1%) residues
Additional allowed region 65(9.5%) residues
Generously allowed regions 7(1%) residues
Disallowed regions 3(0.4%) residues

Remaining mutated positions of SNPs are shown in table 5.8.

Table 5.8. Prediction of residues of SNPs by PDBsum.

Sr no	rs ID	Amino acid change	Most favoured regions residues	Additional allowed regions residues	Generously allowed regions residues	Disallowed regions residues
Wild type XPD protein			612(89.1%)	65(9.5%)	6(0.9%)	4(0.6%)
1	rs140522180	R601Q	608(88.5%)	68(9.9%)	8(1.2%)	3(0.4%)
2	rs147224585	R592H	604(87.9%)	74(10.8%)	5(0.7%)	4(0.6%)
3	rs199738290	R497C	613(89.2%)	64(9.3%)	7(1%)	3(0.4%)
4	rs200665173	Q629H	611(88.9%)	66(9.6%)	5(0.7%)	5(0.7%)
5	rs201370106	L744P	609(88.8%)	67(9.8%)	6(0.9%)	4(0.6%)
6	rs201392911	R695C	600(87.3%)	77(11.2%)	7(1%)	3(0.4%)
7	rs137910235	R227C	609(88.6%)	67(9.8%)	8(1.2%)	3(0.4%)

8	rs200895828	V231M	599(87.2%)	78(11.4%)	7(1%)	3(0.4%)
9	rs372176415	M247T	611(88.9%)	66(9.6%)	6(0.9%)	4(0.6%)
10	rs41556519	R683W	604(87.9%)	77(9.5%)	6(0.8%)	4(0.6%)
11	rs121913021	R658C	607(88.4%)	70(10.2%)	7(1%)	3(0.4%)
12	rs121913023	D681N	604(87.9%)	74(10.8%)	6(0.9%)	3(0.4%)
13	rs121913024	R616W	610(88.8%)	67(9.8%)	6(0.9%)	4(0.6%)
14	rs121913026	R722W	612(87.6%)	79(10.3%)	8(1.4%)	5(0.7%)
15	rs376556895	R616P	603(89.1%)	75(11.1%)	6(0.8%)	3(0.4%)

5.8 Determination of interaction of other genes from XPD

Genemaina show that the XPD has many crucial functions; nucleotide-excision repair, DNA damage removal, transcription-coupled nucleotide-excision repair, DNA-dependent ATPase activity, nuclear transcription factor complex, RNA polymerase complex, RNA capping, Purine nucleotide monophosphate catabolic process, ATP metabolic process and response to radiation. The genes co-expressed with, share similar protein domain, or participate to achieve similar function was summarised in (Table 5.9) and shown in figure 5.6

Table 5.9 Interaction of XPD gene with other genes

Sr no	Network	Physical Interaction	Co-expression	Predicted	Co-localization	Pathway	Shared Protein Domain
1	ERCC2	ERCC2	ERCC2	ERCC2	-	ERCC2	ERCC2
2	GTF2H2	GTF2H2	GTF2H2	GTF2H2	GTF2H2	GTF2H2	-
3	GTF2H3	GTF2H3	GTF2H3	GTF2H3	GTF2H3	GTF2H3	-
4	CDK7	CDK7	CDK7	CDK7	-	CDK7	-
5	CCNH	CCNH	CCNH	CCNH	-	CCNH	-
6	MNAT1	MNAT1	MNAT1	MNAT1	-	MNAT1	-
7	GTF2H4	GTF2H4	GTF2H4	GTF2H4	GTF2H4	GTF2H4	-
8	ERCC3	ERCC3	ERCC3	ERCC3	-	ERCC3	ERCC3
9	GTF2H5	GTF2H5	-	GTF2H5	-	-	-
10	RTEL1	RTEL1	RTEL1	RTEL1	-	-	RTEL1
11	RAD52	RAD52	RAD52	RAD52	-	-	RAD52
12	KYAT3	KYAT3	KYAT3	KYAT3	-	-	KYAT3
13	GPT2	GPT2	-	GPT2	-	-	GPT2
14	KYAT1	-	KYAT1	KYAT1	-	-	KYAT1
15	TOPBP1	TOPBP1	TOPBP1	-	-	-	-
16	MMS19	MMS19	MMS19	MMS19	-	-	-
17	ERCC6	ERCC6	ERCC6	ERCC6	-	ERCC6	ERCC6
18	ERCC5	ERCC5	-	ERCC5	-	ERCC5	-
19	PSMC4	-	PSMC4	PSMC4	-	-	-
20	POR	-	-	POR	-	-	-
21	RTEL1-TNFRSF6B	-	-	-	-	-	RTEL1-TNFRSF6B

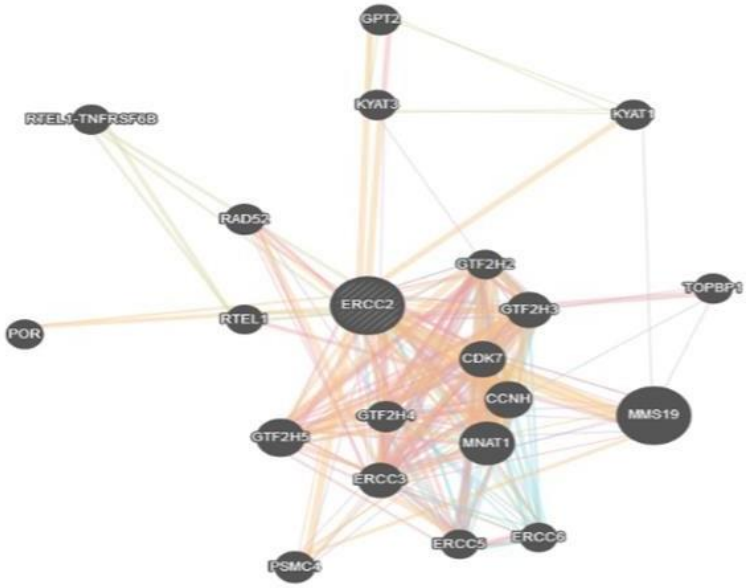


Figure 5.6 XPD gene interactions with other genes

Chapter 6

Discussion

Computational analysis was performed for the first time in this study to investigate whether the polymorphism in XPD gene contribute to any change in structure and function of the protein. Various softwares like SIFT, Polyphen2, I-mutant, Provean, MuPro, Mutation 3D, Genemania, Phd-SNP, PDBsum, SNP&GO, SWISS-model, ELASPIC, PYMOL were utilized for SNPs analysis [Osman et al. 2016]. Three different softwares (SIFT, Polyphen2, Provean) used for finding deleterious SNPs and only common were selected for further analysis. The reason of this consensus approach to increase the reliability of prediction as each tools is based on different algorithms and parameters. Total 19 SNPs from 174 SNPs were predicted as deleterious consensus approach.

Due to gene polymorphisms in XPD gene, one SNP (rs140522180) contributed to deleterious type and according to Elaspic software this mutation was also exist within protein coding region (protein domain). If any mutation observed outside of protein domain then this does not able to affect the structure and functioning of the protein [Hassan et al. 2016].

The SNP (rs140522180) was shown unstable according to I-mutant and MuPro softwares. However the reliability indexes of this mutation were calculated 8 and 10 by Phd-SNP and SNP&GO softwares respectively and it was categorized as disease related type [Ibrahim et al. 2016].

The XPD protein three-dimensional (3D) model with mutated SNP position was predicted by SWISS-model. The RMSD value was calculated by superimposing native and mutated protein molecules; this deviation value affects the function of protein [Hwang et al. 2017]. RMSD value of predicted model was 0.974˚ evaluated by PyMOL with respect to original protein molecule (PDB ID-P18074). The resulted deviation value is not much higher.

Ramachandran plot contributes to determine functional properties of protein by distribution of residues in Favoured or disallowed region as well phi and psi dihedral angles [Kwangsung Park et al. 2017]. Ramachandran plot was generated by PDBsum. Out of 760 amino acids of XPD

protein the favoured regions residues has been found 88.5% (608 residues), additional allowed regions residues 9.9% (68 residues), generously allowed regions residues 1.2% (8 residues) and disallowed regions residues 0.4% (3 residues) for SNP (rs140522180). For a good protein structure more than 90% residues is required in the favoured region.

Hence the computational analysis predicts the effect of mutation on protein coding region and that is why the analysis is not sufficient to study the overall effect of mutation on protein coding region, there is need of other analysis for investigate the overall effect of mutation on protein coding region like *in-vitro* analysis, *in-vivo* analysis and statistical analysis.

Chapter7

Conclusion

In conclusion 19 deleterious SNPs of XPD gene were found to be deleterious. and most of these SNPs were either affecting the structure as they were predicted to be decrease in stability of protein. Seventeen SNPs were identified as disease related by two different tools indicating it's the relevance of these SNPs in the cancer diagnostic and prognostic. Further to confirm its role in cancer development, there is need to study the polymorphism study in different population to establish their association in cancer development.

References

1. Inthasot, V., Bruyneel, M., Muylle, I., & Ninane, V. (2019). Severe pulmonary infections complicating nivolumab treatment for lung cancer: a report of two cases. *Acta Clinica Belgica*, 1-3.
2. Gazdar, A. F., & Zhou, C. (2018). Lung cancer in never-smokers: a different disease. In *IASLC Thoracic Oncology* (pp. 23-29). Content Repository Only!.
3. Hung, R. J., Christiani, D. C., Risch, A., Popanda, O., Haugen, A., Zienolddiny, S., ... & Wichmann, H. E. (2008). International Lung Cancer Consortium: pooled analysis of sequence variants in DNA repair and cell cycle pathways. *Cancer Epidemiology and Prevention Biomarkers*, *17*(11), 3081-3089.
4. Loeb, L. A., Emster, V. L., Warner, K. E., Abbotts, J., & Laszlo, J. (1984). Smoking and lung cancer: an overview. *Cancer Research*, *44*(12 Part 1), 5940-5958.
5. Gupta, D., Boffetta, P., Gaborieau, V., & Jindal, S. K. (2001). Risk factors of lung cancer in Chandigarh, India. *Indian Journal of Medical Research*, *113*, 142.
6. Larsen, J. E., & Minna, J. D. (2011). Molecular biology of lung cancer: clinical implications. *Clinics in chest medicine*, *32*(4), 703-740.
7. Zhou, T., Wang, Y., Zhao, X., Liu, Y., Wang, Y. X., Gang, X. K., & Wang, G. X. (2019). Small cell lung cancer starting with diabetes mellitus: Two case reports and literature review. *World Journal of Clinical Cases*, *7*(10), 1213.
8. Ganesh, B., Sushama, S., Monika, S., & Suvarna, P. (2011). A case-control study of risk factors for lung cancer in Mumbai, India. *Asian Pac J Cancer Prev*, *12*(2), 357-62.
9. Van Loon, A. J., Kant, I. J., Swaen, G. M., Goldbohm, R. A., Kremer, A. M., & van den Brandt, P. A. (1997). Occupational exposure to carcinogens and risk of lung cancer: results from The Netherlands cohort study. *Occupational and Environmental Medicine*, *54*(11), 817-824.
10. Tod, A. M., Craven, J., & Allmark, P. (2008). Diagnostic delay in lung cancer: a qualitative study. *Journal of advanced nursing*, *61*(3), 336-343.
11. Islami, F., Torre, L. A., & Jemal, A. (2015). Global trends of lung cancer mortality and smoking prevalence. *Translational lung cancer research*, *4*(4), 327.

12. Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., ... & Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer*, *136*(5), E359-E386.
13. Florea, A. M., & Büsselberg, D. (2011). Cisplatin as an anti-tumor drug: cellular mechanisms of activity, drug resistance and induced side effects. *Cancers*, *3*(1), 1351- 1371.
14. Vega, F. M., & Ridley, A. J. (2008). Rho GTPases in cancer cell biology. *FEBS letters*, *582*(14), 2093-2101.
15. Friedberg, E. C. (2001). How nucleotide excision repair protects against cancer. *Nature Reviews Cancer*, *1*(1), 22.
16. Liu, H., Rudolf, J., Johnson, K. A., McMahon, S. A., Oke, M., Carter, L., ... & White, M. F. (2008). Structure of the DNA repair helicase XPD. *Cell*, *133*(5), 801-812.
17. Spitz, M. R., Wu, X., Wang, Y., Wang, L. E., Shete, S., Amos, C. I., ... & Wei, Q. (2001). Modulation of nucleotide excision repair capacity by XPD polymorphisms in lung cancer patients. *Cancer research*, *61*(4), 1354-1357.
18. Buscemi, G. (2019). DNA repair and genome integrity. *Oxford Textbook of Cancer Biology*, 13.
19. Xie, C., Zhao, J., Hua, W., Tan, P., Chen, Y., Rui, J., ... & Yang, X. (2019). Effect of XPC polymorphisms on the response to platinum-based chemotherapy: a meta-analysis. *OncoTargets and therapy*, *12*, 3839.
20. Yan, S., Sun, R., Wu, S., Jin, T., Zhang, S., Niu, F., ... & Chen, M. (2019). Single nucleotide polymorphism in the 3′ untranslated region of LPP is a risk factor for lung cancer: a case-control study. *BMC cancer*, *19*(1), 35.
21. Arguello, J. R., Laurent, S., & Clark, A. G. (2019). Demographic history of the human commensal Drosophila melanogaster. *Genome biology and evolution*, *11*(3), 844-854.
22. Mustafa, M. I., Abdelhameed, T. A., Abdelrhman, F. A., Osman, S. A., & Hassan, M. A. (2019). Novel deleterious nsSNPs within MEFV gene that could be used as Diagnostic Markers to Predict Hereditary Familial Mediterranean Fever: Using bioinformatics analysis. *Advances in Bioinformatics*, *2019*.
23. Abdelmoneim, A. H., Mustafa, M. I., Mahmoud, T. A., Murshed, N. S., & Hassan, M. A. (2019). In Silico Analysis and Modeling of Novel Pathogenic Single Nucleotide Polymorphisms (SNPs) in Human CD40LG Gene. *bioRxiv*, 552596.
24. Akhoon, B. A., Tiwari, H., & Nargotra, A. (2019). In Silico Drug Design Methods for Drug

Repurposing. In *In Silico Drug Design* (pp. 47-84). Academic Press.

25. Yousefi, T., Mir, S. M., Asadi, J., Tourani, M., Karimian, A., Maniati, M., & Tehrani, S. S. (2019). In silico analysis of non-synonymous single nucleotide polymorphism in a human KLK-2 gene associated with prostate cancer. *Meta Gene*, *21*, 100578.
26. Amir, M., Kumar, V., Mohammad, T., Dohare, R., Rehman, M. T., Alajmi, M. F., ... & Hassan, M. I. (2019). Structural and functional impact of non-synonymous SNPs in the CST complex subunit TEN1: Structural genomics approach. *Bioscience reports*, *39*(5), BSR20190312.
27. Ford, C. T., Nodzak, C. M., Uppal, A., & Shi, X. (2019). Prediction of the Effect of Naturally Occurring Missense Mutations on Cellular N-Acetyl-Glucosaminidase Enzymatic Activity. *BioRxiv*, 598870.
28. Franczak, C., Kandathil, S. M., Gilson, P., Husson, M., Rouyer, M., Demange, J., ... & Harlé, A. (2019). Uncommon mutational profiles of metastatic colorectal cancer detected during routine genotyping using next generation sequencing. *Scientific reports*, *9*(1), 7083.
29. Castellana, S., & Mazza, T. (2013). Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Briefings in bioinformatics*, *14*(4), 448-459.
30. Nagamani, S., Singh, K. D., & Muthusamy, K. (2016). Combined sequence and sequence-structure based methods for analyzing FGF23, CYP24A1 and VDR genes. *Meta gene*, *9*, 26-36.
31. . Abdelraheem, N. E., Osman, M. M., Elgemaabi, O. M., Alla, A. A. F., Ismail, M. M., Osman, S. A., ... & Hassan, M. A. (2016). Computational Analysis of Deleterious Single Nucleotide Polymorphisms (SNPs) in Human MutS Homolog6 (MSH6) Gene. *American Journal of Bioinformatics Research*, *6*(2), 56-97.
32. Laskowski, R. A. (2001). PDBsum: summaries and analyses of PDB structures. *Nucleic acids research*, *29*(1), 221-222.
33. Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research*, *31*(13), 3381- 3385.
34. Lill, M. A., & Danielson, M. L. (2011). Computer-aided drug design platform using PyMOL. *Journal of computer-aided molecular design*, *25*(1), 13-19.
35. Sobti, R. C., Sharma, S., Joshi, A., Jindal, S. K., & Janmeja, A. (2004). Genetic polymorphism of the CYP1A1, CYP2E1, GSTM1 and GSTT1 genes and lung cancer susceptibility in a north Indian population. *Molecular and cellular biochemistry*, *266*(1- 2), 1-9.

36. Stary, A., & Sarasin, A. (2002). The genetics of the hereditary xeroderma pigmentosum syndrome. *Biochimie, 84*(1), 49-60.
37. Chandrasekaran, G., Hwang, E. C., Kang, T. W., Kwon, D. D., Park, K., Lee, J. J., & Lakshmanan, V. K. (2017). Computational Modeling of complete HOXB13 protein for predicting the functional effect of SNPs and the associated role in hereditary prostate cancer. *Scientific reports, 7*, 43830.

In silico analysis of single nucleotide polymorphisms of XPD gene in lung cancer patients

ORIGINALITY REPORT

10%	6%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.toxsmmu.com Internet Source	1%
2	www.jabonline.in Internet Source	1%
3	Mujahed I. Mustafa, Mohamed A. Hassan. "Unmasking of novel mutations within XK gene that could be used as Diagnostic Markers to Predict McLeod syndrome: Using in silico analysis", Cold Spring Harbor Laboratory, 2019 Publication	1%
4	Ernie Zuraida Ali, Yuslina Zakaria, Mohd Amran Mohd Radzi, Lock Hock Ngu, Siti Azma Jusoh. " Mutation Study of Malaysian Patients with Ornithine Transcarbamylase Deficiency: Clinical, Molecular, and Bioinformatics Analyses of Two Novel Missense Mutations of the Gene ", BioMed Research International, 2018 Publication	1%
5	Heuer, K.. "Structure of a Helically Extended	

	SH3 Domain of the T Cell Adapter Protein ADAP", Structure, 200404 Publication	1%
6	NA. "Supplement 2, Proceedings of the 14th World Conference on Lung Cancer, Book 2", Journal of Thoracic Oncology, 2011 Publication	<1%
7	www.healthscout.com Internet Source	<1%
8	Yunhui Peng, Lexuan Sun, Zhe Jia, Lin Li, Emil Alexov. "Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver", Bioinformatics, 2018 Publication	<1%
9	www.dailymail.co.uk Internet Source	<1%
10	Shreya M. Patel, Prakash G. Koringa, Bhaskar B. Reddy, Neelam M. Nathani, Chaitanya G. Joshi. "In silico analysis of consequences of non-synonymous SNPs of Slc11a2 gene in Indian bovines", Genomics Data, 2015 Publication	<1%
11	www.science.gov Internet Source	<1%
12	Ulla Vogel, Kim Overvad, Håkan Wallin, Anne	

	Tjønneland, Bjørn A. Nexø, Ole Raaschou-Nielsen. "Combinations of polymorphisms in XPD, XPC and XPA in relation to risk of lung cancer", Cancer Letters, 2005 Publication	<1%
13	dhsprogram.com Internet Source	<1%
14	www.lung-mesothelioma-cancer.com Internet Source	<1%
15	Mansi Desai, Jenabhai Chauhan. "In silico analysis of nsSNPs in human methyl CpG binding protein 2", Meta Gene, 2016 Publication	<1%
16	V. Shanthi, R. Rajasekaran, K. Ramanathan. "Computational Identification of Significant Missense Mutations in AKT1 Gene", Cell Biochemistry and Biophysics, 2014 Publication	<1%
17	ar.iiarjournals.org Internet Source	<1%
18	ijirse.in Internet Source	<1%
19	Neeraj Chandra Pant, Ravinder Singh, Aditi Chauhan, Vijaya Gupta, Ravimohan S. Mavuduru, Vijay Prabha, Prince Sharma. "Contraceptive Sperm Agglutinating Proteins	<1%

Identified in Staphylococcus warneri, Natural Microflora of an Infertile Woman", Indian Journal of Microbiology, 2018
Publication

---

20 Muhammad Naveed, Sana Tehreem, Shamsa Mubeen, Fareeha Nadeem, Fatima Zafar, Muhammad Irshad. "In-silico analysis of non-synonymous-SNPs of STEAP2: To provoke the progression of prostate cancer", Open Life Sciences, 2016 <1%
Publication

---

21 G. Divanshu, M. Lekshmi, V. Shanthi. "In silico studies of deleterious non-synonymous single nucleotide polymorphisms (nsSNPs) of NRL gene", Network Modeling Analysis in Health Informatics and Bioinformatics, 2014 <1%
Publication

---

22 C. Rudge, R. Matesanz, F.L. Delmonico, J. Chapman. "International practices of organ donation", British Journal of Anaesthesia, 2012 <1%
Publication

---

23 Vogel, U.. "Two regions in chromosome 19q13.2-3 are associated with risk of lung cancer", Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis, 20040226 <1%
Publication

24	hitachisoft.jp Internet Source	<1%
25	jvi.asm.org Internet Source	<1%
26	zdna2.umbi.umd.edu Internet Source	<1%
27	Daniel C. Ihde. "Primary prevention of lung cancer in the United States", Lung Cancer, 1997 Publication	<1%
28	Hassan, Mohamed M., Sofia B. Mohamed, Mohamed A. Hussain, and Amar A. Dowd. "Deleterious Nonsynonymous SNP Found within *HLA-DRB1* Gene Involved in Allograft Rejection in Sudanese Family: Using DNA Sequencing and Bioinformatics Methods", Open Journal of Immunology, 2015. Publication	<1%
29	eprints.bbk.ac.uk Internet Source	<1%
30	academic.oup.com Internet Source	<1%
31	epublications.uef.fi Internet Source	<1%
32	www.nature.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		

In silico analysis of single nucleotide polymorphisms of XPD gene in lung cancer patients

ORIGINALITY REPORT

10%	6%	9%	%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.toxsmmu.com Internet Source	1%
2	www.jabonline.in Internet Source	1%
3	Mujahed I. Mustafa, Mohamed A. Hassan. "Unmasking of novel mutations within XK gene that could be used as Diagnostic Markers to Predict McLeod syndrome: Using in silico analysis", Cold Spring Harbor Laboratory, 2019 Publication	1%
4	Ernie Zuraida Ali, Yuslina Zakaria, Mohd Amran Mohd Radzi, Lock Hock Ngu, Siti Azma Jusoh. " Mutation Study of Malaysian Patients with Ornithine Transcarbamylase Deficiency: Clinical, Molecular, and Bioinformatics Analyses of Two Novel Missense Mutations of the Gene ", BioMed Research International, 2018 Publication	1%
5	Heuer, K.. "Structure of a Helically Extended	

	SH3 Domain of the T Cell Adapter Protein ADAP", Structure, 200404 Publication	1%
6	NA. "Supplement 2, Proceedings of the 14th World Conference on Lung Cancer, Book 2", Journal of Thoracic Oncology, 2011 Publication	<1%
7	www.healthscout.com Internet Source	<1%
8	Yunhui Peng, Lexuan Sun, Zhe Jia, Lin Li, Emil Alexov. "Predicting protein–DNA binding free energy change upon missense mutations using modified MM/PBSA approach: SAMPDI webserver", Bioinformatics, 2018 Publication	<1%
9	www.dailymail.co.uk Internet Source	<1%
10	Shreya M. Patel, Prakash G. Koringa, Bhaskar B. Reddy, Neelam M. Nathani, Chaitanya G. Joshi. "In silico analysis of consequences of non-synonymous SNPs of Slc11a2 gene in Indian bovines", Genomics Data, 2015 Publication	<1%
11	www.science.gov Internet Source	<1%
12	Ulla Vogel, Kim Overvad, Håkan Wallin, Anne	

	Tjønneland, Bjørn A. Nexø, Ole Raaschou-Nielsen. "Combinations of polymorphisms in XPD, XPC and XPA in relation to risk of lung cancer", Cancer Letters, 2005 Publication	<1%
13	dhsprogram.com Internet Source	<1%
14	www.lung-mesothelioma-cancer.com Internet Source	<1%
15	Mansi Desai, Jenabhai Chauhan. "In silico analysis of nsSNPs in human methyl CpG binding protein 2", Meta Gene, 2016 Publication	<1%
16	V. Shanthi, R. Rajasekaran, K. Ramanathan. "Computational Identification of Significant Missense Mutations in AKT1 Gene", Cell Biochemistry and Biophysics, 2014 Publication	<1%
17	ar.iiarjournals.org Internet Source	<1%
18	ijirse.in Internet Source	<1%
19	Neeraj Chandra Pant, Ravinder Singh, Aditi Chauhan, Vijaya Gupta, Ravimohan S. Mavuduru, Vijay Prabha, Prince Sharma. "Contraceptive Sperm Agglutinating Proteins	<1%

Identified in Staphylococcus warneri, Natural Microflora of an Infertile Woman", Indian Journal of Microbiology, 2018
Publication

---

20 Muhammad Naveed, Sana Tehreem, Shamsa Mubeen, Fareeha Nadeem, Fatima Zafar, Muhammad Irshad. "In-silico analysis of non-synonymous-SNPs of STEAP2: To provoke the progression of prostate cancer", Open Life Sciences, 2016 <1%
Publication

---

21 G. Divanshu, M. Lekshmi, V. Shanthi. "In silico studies of deleterious non-synonymous single nucleotide polymorphisms (nsSNPs) of NRL gene", Network Modeling Analysis in Health Informatics and Bioinformatics, 2014 <1%
Publication

---

22 C. Rudge, R. Matesanz, F.L. Delmonico, J. Chapman. "International practices of organ donation", British Journal of Anaesthesia, 2012 <1%
Publication

---

23 Vogel, U.. "Two regions in chromosome 19q13.2-3 are associated with risk of lung cancer", Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis, 20040226 <1%
Publication

24	hitachisoft.jp Internet Source	<1%
25	jvi.asm.org Internet Source	<1%
26	zdna2.umbi.umd.edu Internet Source	<1%
27	Daniel C. Ihde. "Primary prevention of lung cancer in the United States", Lung Cancer, 1997 Publication	<1%
28	Hassan, Mohamed M., Sofia B. Mohamed, Mohamed A. Hussain, and Amar A. Dowd. "Deleterious Nonsynonymous SNP Found within *HLA-DRB1* Gene Involved in Allograft Rejection in Sudanese Family: Using DNA Sequencing and Bioinformatics Methods", Open Journal of Immunology, 2015. Publication	<1%
29	eprints.bbk.ac.uk Internet Source	<1%
30	academic.oup.com Internet Source	<1%
31	epublications.uef.fi Internet Source	<1%
32	www.nature.com Internet Source	<1%

Exclude quotes	On	Exclude matches	< 10 words
Exclude bibliography	On		