Search for Novel SNPs in XPC, XPD and XPG Genes

Dissertation

Submitted in partial fulfillment of the requirements for the award of degree of

Master of Technology

In

Biotechnology

Submitted By

Sonika Chibh

(Roll No. 601404021)

Under the supervision of:

Dr. Vikas Handa

Assistant Professor

Dr. Siddharth Sharma

Assistant Professor

DEPARTMENT OF BIOTECHNOLOGY

THAPAR UNIVERSITY

PATIALA – 147004

July 2016

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled, ***"Search for Novel SNPs in XPC, XPD and XPG Genes"***, in partial fulfillment of the requirements for the award of degree of Master of Technology in *Biotechnology* submitted in Department of Biotechnology of Thapar University, Patiala, is an authentic record of my own work carried out under the supervision of *Dr. Vikas Handa* and *Dr. Siddharth Sharma* and refers other researcher's work which are duly listed in the reference section.

I have not submitted the matter embodied in this report for the award of any other degree.

Sonika Chibh

Date: 15 July, 2016

(Sonika Chibh)
601404021

This is to certify that the above statement made by the candidate is correct and true to the best of my knowledge.

15.07.2016

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology

Dr. Siddharth Sharma
Assistant Professor
Department of Biotechnology

CERTIFICATE

This is to certify that the work reported in M. Tech dissertation entitled **"Search for Novel SNPs in XPC, XPD and XPG genes"** submitted by Sonika Chibh in partial fulfillment of the requirement for the award of Degree of Masters in Technology in Biotechnology to Thapar University, Patiala is a record of student's own work carried out by her under my supervision and guidance. The report has not submitted for the award of any other degree or certificate in this or any university.

15.07.16

Dr. Vikas Handa

Assistant Professor

Department of Biotechnology

Dr. Siddharth Sharma

Assistant Professor

Department of Biotechnology

Dr. Dinesh Goyal

Head of Department

Department of Biotechnology

Dr. S.S Bhatia

Dean, Academic Affairs

Thapar University Patiala

ACKNOWLEDGEMENT

I would never have been able to finish my dissertation without the guidance of my advisor, help from friends, and support from my family.

I express my deepest gratitude towards my guide ***Dr. Vikas Handa*** *(Assistant Professor) and co-guide* ***Dr. Siddharth Sharma*** *(Assistant Professor) Department of Biotechnology, Thapar University, Patiala for his valuable guidance, support and constant encouragement. He has been very kind and patient while correcting my mistakes and clearing my doubts throughout the project. Blessings, help and guidance given by him from time to time shall carry me a long way in the journey of life on which I am about to embark.*

I would also like to extend my thankfulness towards ***Dr. Dinesh Goyal****, Head of Department and* ***Dr. Niranjan Das****, P.G coordinator Thapar University, for their support, kind cooperation and encouragement. I am really pleased to acknowledge the kind help, cooperation and moral support which I have received throughout my dissertation from of all the teaching as well as non teaching faculty members of Department of Biotechnology, which helped me a lot in completion of this work.*

I am really thankful to PhD scholars Gurpreet and Shweta Lawnia for their kind help and support. I would like to express my utmost gratitude to my parents, for their unconditional affection and support and towards my dear friends for giving me support, friendly environment and unforgettable moments in the Thapar University.

At last, I would like to thanks my Almighty God for his constant blessings without which any task would be impossible.

Sonika Chibh

Date: 15 July, 2016 — Sonika Chibh

Place: Patiala — (601404021)

TABLE OF CONTENTS

ABBREVIATIONS

Abbreviations	Full Form
A	Adenine
C	Cytosine
C5	Cytosine at 5^{th} position
CR/YG	R is a SNP followed by Cytosine and Guanine is followed by Y SNP
CpG	Phosphodiester bond between cytosine and guanine
DNA	Deoxyribonucleic Acid
DNMT	DNA Methyltransferase
dNTPs	Deoxynuceotides Triphosphates
Exp CpG	Expected frequency of CpG dinucelotides
EtBr	Ethidium Bromide
GpC	Phosphodiester bond between guanine and cytosine
G	Guanine
5mC	5 methyl Cytosine
N4	Nitrogen at 4^{th} position
N6	Nitrogen at 6^{th} position
NER	Nucleotide Excision Repair
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
T	Thymine
TpG	Phosphodiester bond between thymine and guanine
TBE	Tris Borate EDTA
TEMED	Tetramethylethylenediamine

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

SNP (Single nucleotide polymorphism) is one of the most common types of genetic variations in the DNA sequence. There are many causes of SNPs. DNA methylation is one of the significant causes of SNPs. DNA methylation is an epigenetic mechanism that involved in number of events *viz.* gene regulation, gene imprinting, and X chromosome inactivation and even in diseases like cancer. It has been found that the pattern of variation or alteration in DNA methylation of cytosine residue is one of the consistent molecular changes in human tumors or cancer. DNA methylation is mainly occurs at cytosine residue in CG containing sites in mammalian genome. CG sites act as mutational hotspot *viz.* CG/CG → TG/CA which is a phenomenon in vertebrate genome due to DNA methylation. In order to search for novel SNPs in the DNA repair genes, we have used RFLP analysis by using the tetra-nucleotide site (CCGG) of HapII restriction enzyme. This study identified one putative SNP in the region of XPD gene and also shows that there is potential role of association of CG dinucleotides with the genetic polymorphism in the genome.

Keywords Single Nucleotide Polymorphism, DNA Methylation, CpG dinucleotides, CpG Islands, XP Genes

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

1.1 Genome Polymorphism

The integrity and stability of the genome is very important for its function as if there is any alteration in the genome then it affects the downstream pathways that may lead to various diseases. It has been found that variation in the genome is a key step for cancer. So maintaining of genetic information in DNA is very important. The main function of DNA (Deoxyribonucleic Acid) is to carry the genetic information in most of the living organisms. DNA contains sequence of four nucleotides named as cytosine, guanine, thymine and adenine that are covalently linked to deoxyribose sugar in a linear polymer. As we all know, human beings differ from one another *viz.* in their physical appearance, susceptibility to disease, response to medication *etc*. To understand the basis of all these differences, it requires the understanding of genetic polymorphism that is variation in the DNA sequence of genome. Scientists have found that many of these differences are SNPs. SNP (Single nucleotide polymorphism) is the type variation in the DNA sequences that occur when there is alteration in the single nucleotide in the genome. SNPs basically arise from mutations. When an allele has frequency of greater than or equal to 1%, then polymorphism is considered to be SNPs (Brookes, 1999).

Figure 1: SNP Formation

https://biogeniq.ca/en/snp/

The occurrence of SNPs has been found throughout the human genome. The main cause of SNPs arises due to DNA replication error. The frequency of SNP has been found to be 1 per 1000 base pairs (*www.ncbi.nlm.nih.gov/SNP/*). SNPs can act as biological markers, which can be further used in the diagnosis of diseases. When SNPs occur within a gene

they may act as a cause for disease by affecting the gene's function. It has been found that in mammalian genome, DNA methylation is one of the significant causes of SNPs.

1.2 DNA Methylation

In 1948, Hotchkiss discovered the DNA methylation in calf thymus DNA. DNA methylation is one of the important epigenetic mechanisms. Any alteration in methylation mechanism can affect gene expression. DNA methylation occurs at N6 position of Adenine, N4 and C5 position of cytosine. But in eukaryotes, it occurs only at C5 position of cytosine (Hotchkiss, 1948). DNA methylation is responsible for the various processes such as genomic imprinting, inactivation of X- Chromosome, cellular differentiation *etc.* (Hermann *et al.,* 2004). It was also found that aberrant DNA methylation is responsible for the cause of cancer (Das and Singal, 2004). When cytosine at C5 position gets methylated by addition of methyl group ($-CH_3$) then form 5-Methylcytosine (5mC) in DNA, and that occurs primarily in CpG dinucelotides in the mammalian genome (Roberts *et al.,* 2003). This covalent modification by methyl group is catalyzed by the action of enzymes named as DNA Methyltransferases (DNMT) (Roberts *et al*., 2003). The cytosine residues that are modified lie immediate 5' of guanine base (CpG methylation). Methylation is found in CpG sequences across the entire genome. Approximately 60% to 90% all CpGs are methylated (Ehrlich M. *et al.,* 1982). It has been found that CGs are globally underrepresented in genomes of vertebrate species and leads uneven distribution of CGs in eukaryotic methylated genomes.

1.3 CpG Islands

The reason behind uneven distribution of CGs is due to the existence of CpG islands. The CpG islands are the regions having high number of CG dinucleotides (Handa and Jeltsch, 2005). Most of the CpG islands are usually found to be unmethylated. Two most accepted definitions of CpG islands were given by Gardiner-Garden and Frommer in 1987 and Takai and Jones in 2002. According to Gardiner-Garden and Frommer, CpG islands are 200bp long having G+C content equal to or more than 50% and Obs/Exp CpG ratio equal to or more than 0.6. According to Takai and Jones, CpG islands are 500bp long having G+C content equal to or more than 55% and Obs/Exp CpG ratio equal to more than 0.65.

1.4 Deamination

Cytosines have been found to undergo deamination giving rise to Uracil, an unnatural base in DNA. The resulting U: G mismatch is repaired by Uracil DNA glycosylase enzyme system. If deamination occurs at 5mC it forms Thymidine. The resulting T:G is not repaired specifically by any of the DNA repair pathway (Matuso *et al.*, 1993). If G/T mismatch not corrected, it may lead to the conversion of the G:C base pair into A:T. Thus CpG sites are more susceptible to mutation, which converts CG/CG to TG/CA after replication.

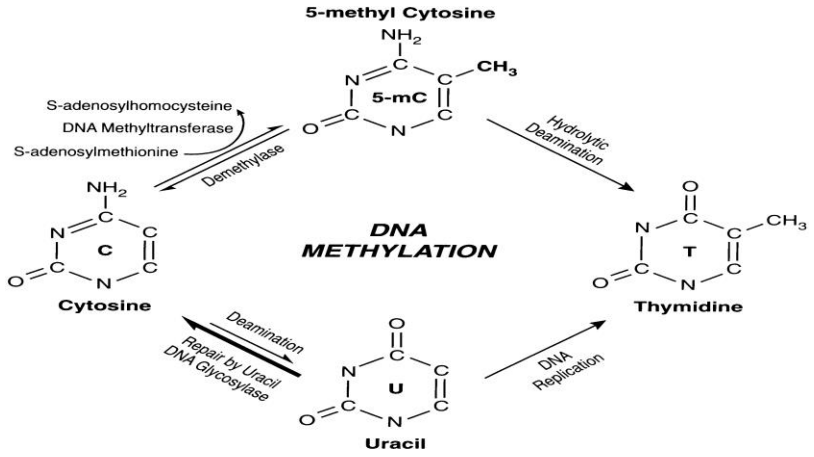

Figure 2: DNA Methylation and Deamination Reaction

(Singhal and Ginder, 1999)

1.4 SNPs due to CGs

The rate of mutation differs between sites within the human genome. These are the sites in DNA where a cytosine is separated from guanine by one phosphate. The notation of "*CpG*" is used to differentiate it from the pairing of *CG* in double stranded DNA. CpG nucleotides have higher rate of mutation than at any other sites in the human genome. A quantitative study was conducted with the genome of chimpanzee and it was found that C to T changes occurred more frequently or transitions occurred more frequently than the transversions. Transition rates were studied by measuring the difference of CpG and GpC transition rates and also see the methylation dependent transition rates *viz.* 5mC deamination rates. This difference was calculated by measuring the transition rate of number of CpG to TpG dinucleotide and transition rate of number of GpC to GpT. It was

found that the cytosine in GpC sites was not methylated. They concluded that the deamination rates of 5methylcytosine were highly dependent on CG content and also estimated that the mutation rate of CpG to TpG higher than other transitions (Jiang and Zhao, 2006).

It is a well known fact that DNA undergoes mutations which is the basis of evolution. The mutations may be caused due various reasons varying from replication error to action of mutagens. However nature tries to maintain integrity of the genetic information to great extent in all the living organisms. DNA polymerase has proofreading activity, but not well enough **(**Jansson, K *et al*., 2013). It introduces errors in about 1 in 10^7 nucleotides added, which it does not correct (Prindle *et al.*, 2012). When DNA damage remains unrepaired it can result in genomic instability that leads to mutations leading to cancer and many other diseases. There are some mechanisms that correct errors left by the replication system after action of mutagens on DNA. Since many mutations are deleterious in nature, DNA repair systems are very important for the survival of all organisms.

1.5 NER Pathway

There are many repair pathways to repair DNA damage such as Nucleotide Excision Repair Pathway, Mismatch Repair Pathway and Recombination Repair Pathway *etc.* We are focusing NER (Nucleotide Excision Repair) Pathway that is involved in vulnerability of melanoma, the most highly malignant skin cancer.

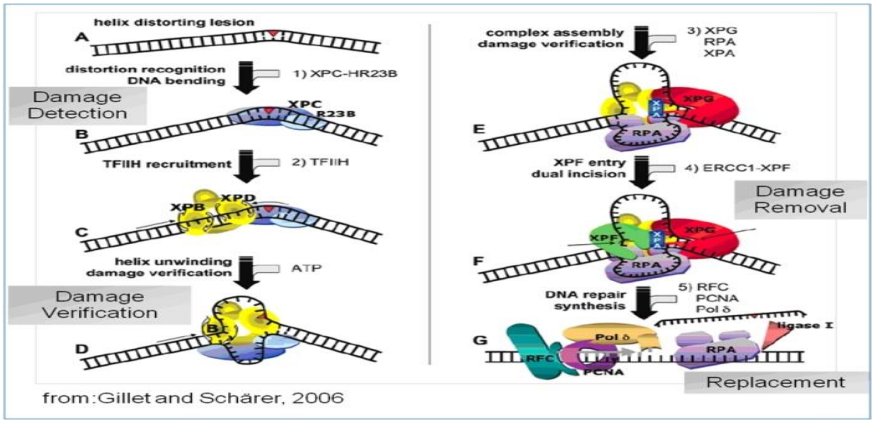

Figure 3: Nucleotide Excision Repair Pathway

1.6 Genes involved in NER Pathway

There are many genes involved in NER pathway in mammalian cells named as XPA, XPB, XPC, XPD, XPE, XPF, and XPG . We are focusing on XPC, XPD and XPG genes in this study.

XPC, XPD and XPG genes play very important role in this pathway. These genes are very important in the early steps of NER pathway *viz*. in recognition of damage, formation of open complex, removal of DNA lesions *etc.* Mutations in these genes result in Xeroderma pigmentosum, which is an autosomal recessive disorder (May N *et al.*, 2010).

NAME	DESCRIPTION	LOCATION	FUNCTION	DISEASES
XPC	Encoded by XPC gene	(3p25.1)	Recognition of damage	Xeroderma pigmentosum
XPD	Encoded by ERCC5 gene	(13q33.1)	ATP dependent helicase activity	1.Cockayne syndrome 2.Mental retardation
XPG	Encoded by ERCC2 gene	(19q13.32)	Acts as single-stranded endonuclease that makes structure-specific cleavage in DNA	1.Photosensitive Trichothiodystropy 2.Cockayne syndrome

Table 1: Brief description of XPC, XPD and XPG Genes

(May N *et al.*, 2010)

As these three genes play very important role in DNA repair pathway. If there is any mutation or variation in these genes it may lead to many diseases. Many studies have shown that there are inconsistencies in the results of variations in nucleotide excision

repair genes that may increase the risk of cancer by affecting repair efficacy. If Single-nucleotide polymorphisms (SNPs) are located in NER genes, such polymorphism can influence DNA repair capacity that may lead to development of cancer. It was found that some polymorphisms in DNA repair genes cause change in the phenotype of individual and thereby increase chances of cancer development. The lesions that distort the double helix of DNA or interfering in base pairing which blocks DNA replication are repaired by NER pathway. This pathway mainly removes the damage that caused by Ultraviolet light which is responsible for the formation of thymine dimmers (Costa *et al.,* 2003).

In this study, our main objective is to search for novel SNPs in these three genes based on RFLP (Restriction Fragment Length Polymorphism). The main purpose of RFLP analysis is to identify the variation or change in the sequence of genome that is identified by the cutting pattern of restriction enzymes. The putative sites for RFLP have been selected by excluding existing SNP site (dbSNP) and focusing on sites with higher propensity of mutations. As mentioned earlier CG/CG are hotspots for mutation. The regions consisting of high abundance of HpaII site (CCGG) have been selected as they comprise of CG dinucleotide. Methylation lead mutation in CG is expected to destroy the HpaII site (CCGG/CCGG to CTGG/CCAG). It is expected that methylation bases mutations that resulted in SNPs in human genome may be detected by RFLP after cutting the target DNA with HpaII.

CHAPTER 2

LITERATURE REVIEW

2. LITERATURE REVIEW

This chapter provides the brief knowledge of previous work done by the scientists on DNA alterations and their association with SNPs. This helps us to understand the subject in better way and acts as directive for this research work.

2.1 **SNP (Single Nucleotide Polymorphism)**

A single nucleotide polymorphism is a variation in the DNA sequence that occurs when one nucleotide is substituted by another nucleotide base. Many scientists have found that SNPs are one of the most abundant types of genetic variations. Although there are many other types of variations, but SNPs are largely the most useful and widely applied biomarkers in the genetic studies (Johnson, 2009). The SNPs do not occur randomly. These mutations are occurred once in history and then passed to further generations. Scientists have found that there are many causes of SNPs. DNA methylation is one of the significant causes of SNPs.

2.2 **DNA Methylation:**

DNA methylation is one of the epigenetic mechanisms that plays very important role in the expression of genes. DNA Methylation occurs in both prokaryotes as well as eukaryotes. Methyl group is added to the N6 position of adenine and N4 and C5 positions of cytosine but in eukaryotes, methylation occurs at C5 position of cytosine. There are specific enzymes that play very important role in methylation process known as DNA Methyltransferases. There are mainly three types of DNMT enzymes in eukaryotes *viz.* DNMT 1, 2 and 3. DNMT enzymes introduced DNA methylation into DNA after the replication process. S-adenosylL-methionine is used as the source of the methyl group by DNMT enzymes and transferred to the DNA bases. In case of prokaryotes methylation process occurs at all sites which are recognized by the enzymes DNA Methyltransferases. But in case of mammalian genome methylation is mainly occurs at CG containing sites (Hermann, 2004).

2.3 **CpG Islands**

It has been found that there is no uniform pattern of methylation in eukaryotes. There is uneven distribution of CGs in the genome which is exemplified by existence of CpG Islands. CpG Islands are mainly unmethylated. There are two definitions of CpG Islands that are given by Gardiner-Garden and Formmer and Takai and Jones, which is acceptable in most of the cases.

In 1987, Gardiner-Garden and Formmer studied that CpG Islands are the regions of DNA that contains high G+C content. CpG Islands are mainly 200 base pair or longer than 200 base pair stretch of DNA with C+G content equal to or more than 50% and observed expected CpG frequency of 0.6 or more. CpG islands are mainly found at 5' ends of all housekeeping genes and 3' ends of tissue-specific genes.

In 2002, **Takai and Jones** had done the further analysis of complete genomic sequences of human chromosomes 21 and 22 to see the properties of CpG islands by using a search algorithm. The definition of CpG Islands' was redefined as stretches of more than or equal to 500 base pairs of DNA with G+C content more than or equal to 55% and observed over expected CpG frequency of 0.65 or more. Those CpG islands which present at the promoter region play very important role in gene silencing. One of the very important methods of bioinformatics was used for the analysis of various species named as nearest neighbor method. This study suggested that the human genome showed the strongest suppression of CpG as compare to other species. This finding was compatible with the detection of 5-methylcytosine formation due to CpG methylation.

2.4 **Association of 5mC to mutations:**

The pattern of variation or alteration in DNA methylation of cytosine residue has been found consistent molecular changes in human tumors or cancer. When methyl group is added to the C5 position of cytosine it forms 5mC (5 methyl cytosine), which gets deaminated to form thymine that is a transition mutation. There is no specific DNA repair pathway for this mutation. On the other hand cytosine residues when get deaminated to form Uracil which can be repaired by one of the DNA repair enzyme named as uracil

deglycosylase. Studies conducted by the scientists show that CG sites are mutational hotspot.

In 1980, Bird had found that deficiency of CpG dinucelotides related to DNA methylation. It was recognized that in many mammals the presence of CpGs were less than would be expected composition of the base. The study was conducted by one of the method in which degree of methylation was analyzed with the help of restriction enzymes HpaII and Msp1. By running the samples on gel electrophoresis for vertebrate DNA samples, there was a comparison of molecular weights of the digests with the fraction of available CCGG that was kept uncut by Hpa II. It was noticed that organisms with the most extreme CpG deficiency had the highest levels of DNA methylation *viz.* mutation of 5mC site was relatively more frequent compared to the other dinucleotides and those genomes which were poorly methylated can't have the deficiency of CpG.

In 2006, **Jiang and Zhao** conducted quantitative analysis study, and suggested that 5methylcytosine deamination rates at CpG dinucelotides are highly dependent on local CG content and also on the length of flanking sequences of SNPs. This study was conducted with the chimpanzee genome and it was found that the C to T changes occurred more frequently than other transitions. Transition rates were analyzed by measuring the difference of CpG transition and GpC transition, and examined methylation-dependent transition rates (5mC deamination rates). This difference was calculated by measuring the CpG transition rate by measuring the number of CpG to TpG transitions and GpC transition rate by the number of GpC to GpT transitions. It was found that the cytosine in GpC sites was not methylated. It was concluded that 5Mc deamination rates were highly dependent on local GC content and also estimated that the mutation rate of ^{m}CpG to TpG to be 10–50 folds higher than other transitions.

In 2000, Ford *et al.*, showed that mutations are early events of diseases, so it might be possible that defects in DNA repair pathway probably create a risk factor for cancer. It was studied that variants found in some DNA repair genes were responsible to predispose to cancer. There was a study conducted in which single nucleotide polymorphisms

(SNPs) were identified in DNA repair genes. It was well studied that in mammalian mutations, transitions at CpG sites predominate and was over-represent in diseased human. This reflects that there is contribution of the deamination of 5-methylcytosine. When the selection of variations in human DNA repair genes occurred then about six G: C to A: T transitions were identified and out of which five occur at CpG sites.

Previous studies suggested that 5-Methylcytosine is a mutagenic site *viz*. it can undergo spontaneous deamination reaction which converts 5mC to T. It has been found that the mutation rate that is caused by deamination is further increased by many other factors viz. environmental factors *i.e.* Sunlight, UV radiations and other carcinogenic substances *viz*. tobacco *etc*. (Brueckner, 2005). The damage to DNA is occurred by both endogeneous as well as exogeneous sources that may lead to mutations. So for the maintenance of integrity of the genome, DNA repair pathways are there to repair the DNA damage. There are number of repair pathways but we are focusing on NER repair pathway as already mentioned.

2.5 **NER Pathway:**

Sunlight is one of the causes of DNA damage that includes DNA lesions which can be removed by one of the DNA repair pathway named as nucleotide-excision repair pathway. There are approximately eight NER proteins that participate in this pathway *viz.* ERCC1, XPA, XPB, XPC, XPD, XPE, XPF, and XPG and most of these identified from XP complementation groups. Genetic alterations in these genes may alter NER pathway (Chunying *et al.,* 2006). Individuals with inherited defects in nucleotide excision repair have low repair of DNA lesions which induced by Ultraviolet radiations. They are at extremely high risk of skin cancers. It has been studied that in healthy individuals, polymorphisms in DNA repair genes show association with DNA damage and susceptibility to cancer (Baccarelli *et al.,* 2004).

In 2013, Dupuy A. *et al.,* showed that presence of 5mCs in XPC gene might be the cause of alterations. There are mainly seven different genes which are involved in NER repair pathway which is used to remove the lesions that are caused by UV light. Mostly

XP patients in North African and European countries have alteration in XPC gene. There was an experiment conducted in which the treatment with a demethylating agent or the use of 5mC insensitive nuclease was used for the correction of XPC gene. This demethylating agent named as 5-aza-Dc help to remove the negative effect of CpG methylation. This enable to re-express the full-length XPC protein in the cell lines .So it might be possible that the presence of methylated cytosines (5mC) in the *XPC* gene is might be the cause of alteration.

As we all know that XPC, XPD and XPG genes play very important role in NER pathway. If Single-nucleotide polymorphisms (SNPs) located in NER genes, it can influence the DNA repair capacity resulting in implications such as cancer development. As the functional impact of all polymorphisms has not been identified, some polymorphisms in DNA repair genes are responsible to induce phenotypical changes. So our main purpose of this study to search for a novel SNP which can be further act as diagnostic marker for the diseases like cancer.

CHAPTER 3

SCOPE OF STUDY

3. SCOPE OF STUDY

DNA methylation is one of the epigenetic mechanisms. It mainly occurs at cytosine residue of DNA and form 5mC. When 5mC site get deaminated and form thymine (TpG/CpA), it will not repaired by any of DNA repair mechanism. One of the significant causes of SNPs is DNA methylation in mammalian genome. This work focuses on understanding the relationship of methylation and CG/CG→TG/CA mutation.

More than expected number of SNPs appears to be mutated due to CGs. Though large number SNPs are exist in human genome in three genes but one cannot rule out the possibility of existing of new SNPs. These genes play important role in human health. So the present study attempts to find out the novel SNPs.

A new method has been tried to detect SNPs by focusing on CG dinucleotides in non-CpG island sequence of DNA.

CHAPTER 4

OBJECTIVES

4. OBJECTIVES

1 Analysis of already existing SNPs in three genes of NER pathway (XPC, XPD and XPG)

2 Search for potential SNP sites in the three genes

3 Experimental determination SNPs at the expected sites using RFLP analysis

CHAPTER 5

MATERIALS AND METHODS

5. MATERIALS AND METHODS

The study was carried out with the help of computational approach to search for a sequence stretch consisting of large number of CCGGs for the digestion with HpaII restriction enzyme. Preferably choose those CCGGs that were not in the part of CpG Islands and also prefer the sequence not carrying any existing SNP based on CG/CG → TG/CA mutation. Hereafter such SNPs will be referred to as CG-SNP.

5.1 Data Source:

DNA sequences were downloaded from NCBI (URL: *www.ncbi.nlm.nih.gov*). To search for novel SNPs the first step of this study was to analyze the existing SNPs in dbSNP (Single Nucleotide Polymorphism Database) of the three genes, XPC, XPD and XPG.

S.NO.	SPECIES	GENE NAME	LOCUS	ACCESSION NO.
1	*Homo sapiens*	XPC	3p25.1	NC_000003.12
2	*Homo sapiens*	XPD	13q33.1	NC_000019.10
3	*Homo sapiens*	XPG	19q13.32	NC_000013.11

Table 2: DNA sequences of different genes taken from GenBank

5.2 dbSNP Database *(URL: www.ncbi.nlm.nih.gov/SNP/)*

dbSNP database is a collection of genetic variations *viz.* polymorphisms. It was used to search for already existing CG-SNPs *viz.* CR/YG in all the three genes.

5.3 CpG Island Searcher *(URL: cpgislands.usc.edu)*

CpG Island Searcher was used to search for CpG Islands in all the three genes. For searching of CpG Islands the following criteria (Takai and Jones, 2002) was used:

1) GC content – 55%
2) Obs/Exp CpG – 0.65
3) Length – 500bp
4) Gap between adjacent islands – 100bp

5.4 Sequence analysis tools

Microsoft Excel

Microsoft excel spread sheet was used for computation and statistical analysis of data. Following functions were used:

S.No.	Tools	Class	Functions
1	COUNTIF	Statistical	It is used to count the number of cells in the given range.
2	LEN	Text	It is used to return the number of characters in a text string.

Table 3: Functions used in Microsoft Excel

Notepad ++

Notepad++ was used for recording macros and running them to analyze DNA sequences of large size. DNA sequence manipulation such as converting the sequence line into single line was performed by 'line operation' (ctrl J). Further the single line sequence was broken into the length of 1500. Macro recording was applied to execute simple sequence manipulations which were required to be repeated several times.

5.5 Odds Ratio (URL: *http://statpages.info/)*

Odds Ratio and p- value was calculated with the help of 2*2 Contingency Table. Odds ratio is basically used to quantify that how the presence and absence of one property associated with the presence and absence of other property.

If OR = 1, there is no association between two properties

If OR > 1 and OR < 1, there is a possible statistical relationships between the two properties

(Westergren *et al.*, 2001)

	Condition	
	Present	Absent
Property 1	A	B
Property 2	C	D

Table 4: Odds Ratio = (a/c) / (b/d)

P-value indicates that whether the findings in a research study are statistically significant *viz.* that the results are not likely to have occurred by chance. (Forbes, 2012)

5.6 Sequence of Primers

Following primers were used for the PCR amplification of the region of XPD gene.

Forward primer – 5'CAG TCA CAC TTG CAA ACC 3'

Reverse primer - 5'AGC TCA TCT CTC CGC AGG 3'

5.7 DNA polymerase Enzyme

Taq DNA polymerase enzyme of *Geneaid* was used for PCR amplification.

5.8 Methodology

5.8.1 **Genomic sequence**

Sequence of all the three genes *i.e.* XPC, XPD and XPG was downloaded from NCBI in the FASTA format.

Gene	Start Position	End Position	Accession No.
XPC	14145147	14178672	NC_000003.12
XPD	45351391	45370646	NC_000019.10
XPG	102845841	102876001	NC_000013.11

Table 5: Brief description about the genes

5.8.2 **Search for CpG islands**

CpG Islands were searched for all the three genes by using CpG Island searcher (*cpgislands.usc.edu).*

5.8.3 **Search for Non CpG Islands**

Identification of regions in the genes which do not overlap with CpG islands *viz.* non CpG region. CpG islands usually not methylated that's why there was lower probability for novel SNPs in CpG island regions.

5.8.4 **Search for CG-SNPs**

CG-SNPs that were already existed in all the three genes were searched by using dbSNP database. The CG- SNPs were searched in the form of CR/YG. (R is the SNP in CR that is followed by C and Y is SNP in YG). If mutation occurred by 5mC then CR/YG should be

$$CG \rightarrow CA$$

Or

$$CG \rightarrow TG$$

S.No.	Start Position	End Position	Transcript	Ref ID
XPC	14180634	14144657	Minus strand	rs1870134
XPD	45350896	45352529	Minus strand	rs13181
XPG	102843853	102876496	Plus strand	rs17655

Table 6: Brief description about SNPs of all the three genes

5.8.5 **Search for CGs and CCGGs**

Total number of CGs and CCGGs in all the three genes was searched with the help of "FIND" option in Notepad++ in the CpG Island region or non CpG island region.

5.8.6 **Analysis of CG-SNPs**

CG-SNPs were searched in the SNP database of each gene by looking for:

- C/T SNP followed by G
- G/A SNP preceded by C

5.8.7 **Mapping of CG-SNPs with CCGGs and identification of the region of gene having maximum number of CCGGs not part of SNPs**

The position of CG-SNPs viz. CR/YG was searched through dbSNP database and the position of CCGGs was searched manually in notepad++. After that the positions of CG-SNPs was mapped with the positions of CCGGs and searched for the CYGG and CCRG. This study was done to exclude the CCGGs that were part of CG-SNPs and left with the CCGGs not part of CG-SNPs and look for the region of the gene that contains maximum number of CCGGs not part of SNPs.

5.8.8 **Designing of Primers** (*https://www.thermofisher.com*)

The primers were designed for the selected region of XPD gene. The melting temperature *viz.* (T_m) of all the three genes was calculated by using the T_m calculator of ThermoFisher. Then after designing of forward and reverse primers optimization of PCR was done.

5.8.9 **Simulation**

Simulation of digestion was done to identify the cutting pattern of CCGGs sites by HpaII restriction enzyme. The expected size of DNA fragments after restriction digestion was plotted on Excel sheet using inversely proportion relationship between ln of molecular mass of DNA fragment and mobility of the corresponding band.

5.8.10 **PCR Amplification**

The selected region of XPD gene was PCR amplified by using samples of Lung cancer patients and amplification was checked by running the samples on agarose gel electrophoresis.

PCR amplification *(https://www.promega.in)*

PCR amplification of 1596bp long DNA fragment of XPD gene was done in 30µl reaction mixture containing 3ul of DNA template, 1.5µl of both forward primers 5'CAG TCA CAC TTG CAA ACC3' and reverse primer 5'AGC TCA TCT CTC CGC AGG 3' with 1mM $MgCl_2$, 3ul dNTPs, 3ul of 10x PCR buffer and 0.9 ul of1U Taq DNA polymerase. PCR was performed with initial denaturation step of 1minute at 94ºC then followed by 30 cycles of 40 seconds at 94ºC, 30 seconds at 54ºC as annealing step and 2 minutes at 72ºC followed by final extension step of 5minutes at 72ºC. The amplification was confirmed by using 1.2 % agarose gel electrophoresis.

Components	Volume (30ul)
Distilled water	17.1ul
10x Reaction buffer	3 ul
Forward primer (10um)	1.5ul
Reverse primer (10um)	1.5ul
dNTPs (100mm)	3ul
Template DNA	3ul
Taq DNA polymerase (1u/ul)	0.9ul

Table 7: PCR Reaction Mixture

Touchdown PCR:

Touch-down PCR was used to increase the specificity of the reaction. PCR amplification of 1596bp long DNA fragment of XPD gene was done in 30µl reaction mixture containing 3ul of DNA template, 1.5µl of both forward primers 5'CAG TCA CAC TTG CAA ACC3' and reverse primer 5'AGC TCA TCT CTC CGC AGG 3' with 1mM $MgCl_2$, 3ul dNTPs, 3ul of 10x PCR buffer and 0.9 ul of1U Taq DNA polymerase. The first step to set the reaction of the Touchdown-PCR was to set 5 °C to 10 °C higher annealing temperatures than the melting temperature of the primers that was calculated. By end of the amplification stages, the annealing temperature was decreased to 2 °C to 5 °C below the melting temperature. PCR was performed with initial denaturation step of 1minute at 94ºC then followed by 30 cycles of 40 seconds at 94ºC, 30 seconds at 54ºC as annealing step and 2 minutes at 72ºC followed by final extension step of 5minutes at 72ºC. The amplification was confirmed by using 1.2 % agarose gel electrophoresis.

Agarose Gel Electrophoresis (1.2%) (*https://www.addgene.org)*

For 1.2% of agarose gel electrophoresis 1.2 grams of agarose powder was weighed and it was added into flask having 100 ml of TBE buffer. Then agarose powder was melt properly in oven until the solution became clear. After that agarose solution was kept for cooling and then approximately 2ul of EtBr was added. Now combs were placed into gel casting tray and melted agarose solution was poured into gel casting tray. The tray was kept for 10-15 minutes for solidification of agarose. When the agarose was solidified combs were pulled out carefully. Gel casting tray was kept into electrophoresis tank containing enough amount of TBE Buffer for running the gel. The gel was run at 60 volts till the bromophenol blue was reached at the end. Then gel picture was taken in gel documentation system. In some cases multiple bands were observed. So to cut the desired band, gel extraction method was used.

Extraction of DNA from band in agarose gel (*http://fg.cns.utexas.edu/)*

Gel electrophoresis was done to analyze the PCR amplification product. Using UV trans-illuminator the gel was exposed to UV briefly to visualize the DNA bands. The band

corresponding to the expected amplified product was cut out of the gel as a small agarose block using scalpel blade. Gel piece for each sample was separately placed in 1.5 ml microcentrifuge tubes. Then tubes were kept in freezer for overnight. After overnight incubation, gel pieces were crushed properly and resuspended in autoclaved TE buffer. Further ethanol precipitation was done to purify the DNA. The volume of sample was measured and appropriate amount of sodium acetate (having pH 5.2 and final concentrations 0.3 M) was added to the sample. Then the sample was mixed properly. 100% chilled ethanol was added around 2 to 2.5 times of the sample. Again proper mixing of sample was done. After mixing, the sample was kept at -20 degrees for 20 minutes. After 20 minutes incubation the sample was spinned at 12000 rpm for 15 to 20 minutes. Now the supernatant was discarded and kept the pellet. Pellet was washed with 1 ml 70% ethanol and air dry the pellet. The pellet was resuspended into appropriate volume of TE.

5.8.11 **RFLP Analysis**

RFLP analysis mainly used to search for polymorphism by comparing the migration of wild type samples to digested samples by electrophoresis. The PCR product of XPD gene was then subjected to restriction digestion with 1U of HpaII restriction enzyme incubated overnight at 37^{o}C. After digestion, the samples were then run on 15% polyacrylamide gel electrophoresis.

Native PAGE (*http://microbiology.ucdavis.edu)*

For running of polyacrylamide gel electrophoresis, the first step was to prepare the following reagents

Acrylamide: bisacrylamide as (29:1) (30%w/v)

Ammonium persulfate (10% w/v)

5X TBE Buffer

After preparing the reagents the apparatus was set. Plates and spacers were cleaned properly. Plates were adjusted in gel caster with the help of spacers. Gel solution was prepared with the desired polyacrylamide percentage.

Gel %	30% Acry/Bis	Water (ml)	5X TBE (ml)	10% APS (ul)	TEMED (ul)
10%	4.0 ml	5.6	2.4	200	10
12%	4.8 ml	4.8	2.4	200	10
14%	5.6 ml	4.0	2.4	200	10

Table 8: Percentage for PAGE

After addition of TEMED, immediately the gel solution was poured before it get solidify. Combs were inserted immediately after pouring the gel solution. Gel solution was allowed to polymerize for 30 to 35 minutes at room temperature. Combs were removed after the polymerization of gel. After that the gel was removed from gel caster. Gel was placed into electrophoresis tank that having 1X TBE buffer of pH 8. DNA samples were mixed with loading buffer. After that, first Molecular ladder was loaded into the first well and after that samples were loaded. Electrodes of electrophoresis tank were connected to the voltage. Then the gel was run until marker dye was reached the desired length. Now to visualize the DNA fragments, silver staining was done.

Silver Staining (*http://www.proteinchemist.com/*)

Insert the gel into fixative solution that contained (Water: methanol: Glacial Acetic Acid) (50:40:10) for 30 minutes. The gel was washed three times with distilled water. After washing, the gel was kept in silver stain solution for 20 minutes that contained 1% AgNO3 and 150ul of formaldehyde. Again washing was done with distilled water. The gel was kept into developer solution that contained 3 g sodium carbonate in 100 ml, 150ul formaldehyde and 20ul of sodium thiosulfate until bands became visible

CHAPTER 6

RESULTS

6. RESULTS

6.1 CpG Islands in all the three genes

XPC Gene

	CpG Island 1	CpG Island 2
Start to End Position	5119-5624	10972-12202
%GC	55.1	57.9
Obs/Exp CpG	0.656	0.855
Length	506	1231

XPD Gene

	CpG Island 1	CpG Island 2
Start to End Position	13720-14331	19743-21182
%GC	70.9	60
Obs/Exp CpG	0.651	0.745

Length	612	1440

XPG Gene

	CpG Island 1	**CpG Island 2**
Start to End Position	1-1179	45494-46774
%GC	58.5	56.4
Obs/Exp CpG	0.924	0.775
Length	1179	1281

Table 9: Search of CpG Islands in all the three genes by using the criteria: %GC≥55%, Obs CpG/Exp CpG≥0.65, Length≥500bp and Gap between adjacent islands≥100bp

6.2 Statistical analysis

Gene	Acc No.	Position (Length)	SNPs	CGs	CG-SNP	nCG SNP	CCGGs		CGs	CCGGs	CCGGs SNP	CCGGs non SNP	Exons	Introns
XPC	NC_000003.12	14145147-14178672 **(**33525**)**	2358	819	565	1793	44	**CGI**	110	19	13	6	5	1
								nCGI	709	25	15	10	0	10
XPD	NC_000019.10	45351391-45370646 (19255)	2275	764	596	1679	56	**CGI**	150	11	3	10	4	6
								nCGI	654	45	25	18	10	8
XPG	NC_000013.11	102845841 - 102876001 (30160)	2481	1081	515	1966	29	**CGI**	169	11	4	7	4	3
								nCGI	912	18	12	6	1	5

Table 10: Statistical Analysis of the XPC, XPD and XPG genes to search for CCGGs not part of existing SNPs

As shown in table 10. In **XPC gene** sequence total 2358 SNPs were found in dbSNP database out of which 565 were CG-SNPs. 565 CG-SNPs mapped with 819 CGs to search for CGs which are not coinciding with the existing CG-SNPs.

Further genome sequence of the gene was searched for HpaII sites (CCGG). Out of 44 HpaII sites only 16 HpaII sites (CCGG) were found not be coinciding with existing CG-SNPs. Within those 16 CCGGs, 6 CCGGs were found to be CpG Islands.

In **XPD gene** sequence total 2275 SNPs were found in dbSNP database out of which 596 were CG-SNPs. 596 CG-SNPs mapped with 764 CGs to search for CGs which are not coinciding with the existing CG-SNPs.

Further genome sequence of the gene was searched for HpaII sites (CCGG). Out of 56 HpaII sites only 28 HpaII sites (CCGG) were found not be coinciding with existing CG-SNPs. Within those 28 CCGGs, 10 CCGGs were found to be CpG Islands.

In **XPG gene** sequence total 2481 SNPs were found in dbSNP database out of which 515 were CG-SNPs. 515 CG-SNPs mapped with 1081 CGs to search for CGs which are not coinciding with the existing CG-SNPs.

Further genome sequence of the gene was searched for HpaII sites (CCGG). Out of 29 HpaII sites only 13 HpaII sites (CCGG) were found not be coinciding with existing CG-SNPs. Within those 13 CCGGs, 7 CCGGs were found to be CpG Islands.

From the above data the region of XPD gene was selected for PCR amplification having 8 CCGGs in it. Out of 8 CCGGs 3 exist as known SNPs and 5 CCGGs not exist as SNPs in dbSNP database.

6.3 Odds ratio and p- value

	XPC			XPD			XPG	
	SNP	**nSNP**		**SNP**	**nSNP**		**SNP**	**nSNP**
CG	565	254	**CG**	596	168	CG	515	566
nCG	1793	32547	**nCG**	1679	19016	nCG	1966	29597

Table 11: Statistical analysis of CG-SNPs in the three genes

XPC		XPD		XPG	
Odd Ratio	41.07	**Odd Ratio**	40.17	**Odd Ratio**	13.64
p-value	0	**p-value**	0	**p-value**	0

Table 12: Calculation of Odds ratio and p- value for all the three genes

Odds ratio for all the three genes were calculated. As shown in the table 12 odds ratio was coming out 41.07 fold increase in case of XPC gene, 40.17 fold increase in XPD gene and 13.64 fold in XPG gene. P value was almost 0 in all the three genes. This indicates that there was significantly high probability of SNPs due to CG dinucleotides *viz.* CG→ TG/CA.

6.4 Selection of a region of gene for amplification and Primer designing for that region

The mapping of CCGGs with CG-SNPs was done. All the three genes were analyzed and the region of 1596 base pairs of XPD gene was selected having 8 CCGGs in it out of which 3 were already existing CG-SNPs.

Gene	Accession number	Start position	End position
XPD	NC_000019.10	45362541	45364136

Table 13: Region of XPD gene for PCR amplification

```
   1 cagtcacact tgcaaacctg cctcccgcac gtcagccctc ccagcctggc tcctctaacc
  61 ccaggcgcct accccggctg gtcgctagct gaggatgagt ctgtcccctc cactgggcca
 121 tggagcctct acggctgtgt ccccagcatg aagcaggcac ttggggaatg tgtgttgagt
 181 gactgagatc ccctctttgc cctcacaagt ctccaatcct ggggaagatg ggcaaaccca
 241 catggaaggt ggcctcaccc acagaggcag ggcagccagg gcctttggag cccacaggca
 301 tggcttcaag ccccaactac cgggcacagt gctgtgtggc tctgggcaaa ggcttccccc
 361 ctgggagtcc cagtgtcctc atttgcaggg tggggatgct gtactggggg tgcaactgag
 421 gtaaggaaca taaagggctt agctcctcct tagtgctcag ggcagggtag gtgattggct
 481 tcgctgctct ctcattaacc caccaaagaa ccctgtgaag taaagacccg actccttggc
 541 cattttacaa aggaggaaac tgcagttcaa agaagaccca tgacttaccc aaaggaaaag
 601 ccccctcgta tcccttggtc attctctgcc tgccctgtgc ctcaagagac tgactccccc
 661 aaaatgcgcc cagactactc catggggacg cccctgccct caggtttgga agattagagg
 721 catcggtggg aagatacatg ggccggggac acctcctcca gctcccagcc tgcgaggccc
 781 tggtactggt ttctgctccc agcagaagcc cctgtccagg cccccaccag ttggttctca
 841 acacactccc tccctgccgc tgtggcctgg cttagctccc agctctcacc tcccaccctc
 901 cttggctcct cccccagccc gcccctctgt aaatggtgct cccactctct cttgggtggg
 961 ggagattctc caatccagcc aggtgttgcc tgccaggccc ctgcccgacg cacggaccca
1021 ggatcagctc caaccttgct cccggacctc ccaggtcaga ccagacaagg tcccacgtgc
1081 ctcgcccagc tcactcactc ctgatgctgc agtggtgacc tggggctaca ggcgtggagg
1141 acacggctct gcataaccgg gacctgccgg gcccccaccc cgcgcgctgt ctggggccgc
1201 acctgagggg cttgcgctgg atgcacacgc gctgggccag gccgctcagg aaggcgggcg
1261 ggctctcctg caccacatgc tgcacacgca gccgccactt cacgtactcc agcagccgcc
1321 tcaggaagcc caggaaatgc tcggccgtgc ggatggagcc aggcactgcc tctgcgagga
1381 gacgctatca gcggcgacgg ggaggcggga aagggactgg ggggcagcgg ggggtcgggg
1441 ctcaccctgc agcacttcgt cgggcagcac ggggttggcc aggtgggcgt ccgtctccgg
1501 ggcggcgctg gcctcccgca gcccctccac cagacgccgg tactcgtccc gcaggcgctg
1561 ctcgtctgtc tctttgatcc tgcggagaga tgagct
```

Figure 4: XPD Gene Sequence was taken from *Genbank* having 8 CCGGs out of which 3 three ccgg already exist as SNPs in dbSNP database and five ccgg not as SNPs

Forward Primer: 5'*cagtcacacttgcaaacc* 3'

Reverse Primer: 5' *agctcatctctccgcagg* 3'

6.5 Simulation of digestion

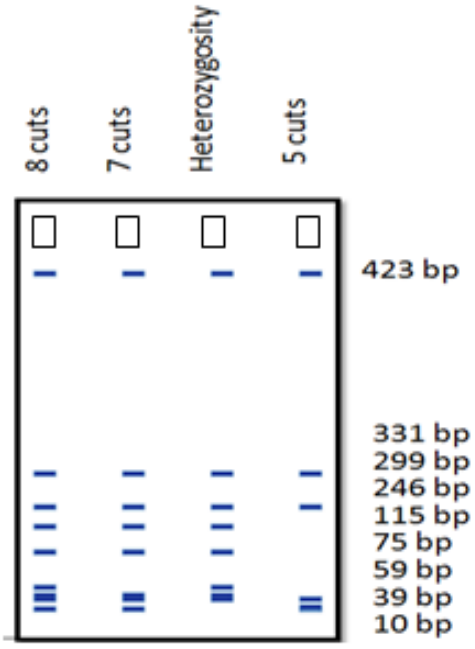

Figure 5: Simulation pattern of restriction digestion of the PCR product using HpaII enzyme

Simulation was done to identify the cutting pattern of digested samples. Total eight CCGGs were identified in the region of XPD gene, out of which three CCGGs were already part of existing SNPs. Figure 5 representing the cutting patterns of 8 CCGG sites. First lane is showing the cutting pattern if all the 8 CCGG sites will cut. Second lane is showing if 7 CCGG sites will cut. Lane 3 is showing the pattern of heterozygosity if 7 CCGG sites will cut. Last lane is showing the pattern if 5 CCGG sites will cut.

6.6 PCR amplification

Amplification of DNA samples of lung cancer patients' was done. The amplicons obtained after the amplification process were run on agarose gel (1.2%). The agarose gel was stained with EtBr which stacked between the base pairs of the DNA and is

responsible for illuminating when seen under the UV light. This was done to see whether the DNA had undergone amplification or not.

6.6.1 **Optimization of PCR**

For optimization of PCR, the reaction was done at different temperatures by using different concentration of $MgCl_2$. PCR was optimized at 54 °C with 1 mM $MgCl_2$ concentration.

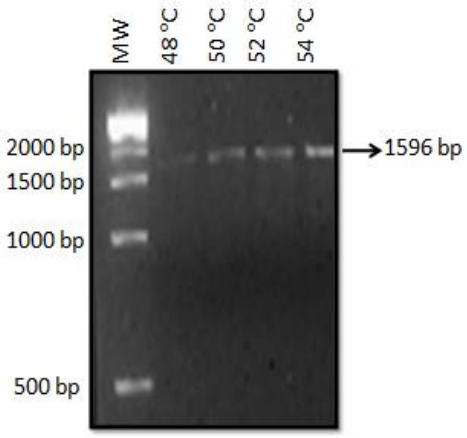

Figure 6: Representing the optimization of PCR at different temperatures

As shown in figure 6. Same sample of DNA *viz.* LC 404 (Lung cancer patients' sample) was loaded in all the four lanes for optimization of PCR. In first lane molecular weight of 500bp was loaded. In next four lanes DNA sample (LC404) was loaded and the PCR was done at different temperatures (48 °C, 50 °C, 52 °C and 54 °C). And the optimization of PCR was done at 54 °C.

After optimization of PCR reaction, DNA samples of lung cancer patients' and controls were further amplified for the RFLP analysis.

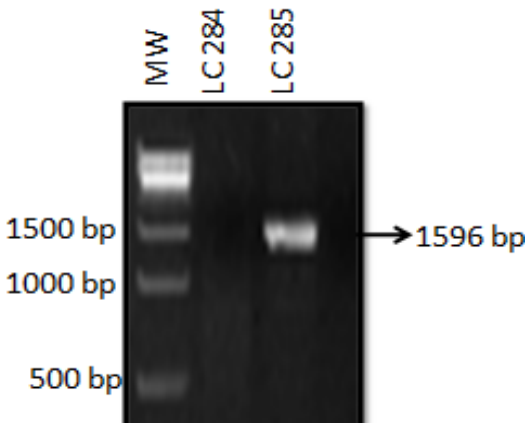

Figure 7: Representing 1.2% gel electrophoresis of PCR amplified product of 1596 bp of XPD gene

6.5.1 Cutting out of bands by gel extraction method

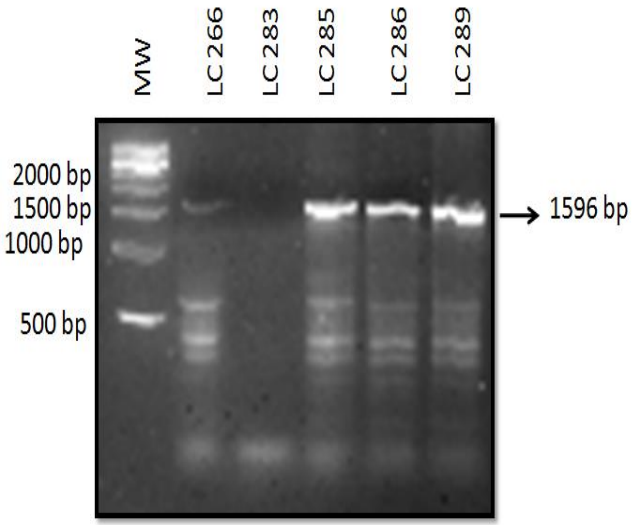

Figure 8: 1.2% of gel electrophoresis showing the presence of multiple bands after PCR amplification

As shown in figure 8. multiple bands were observed after PCR amplification. Then the gel extraction method was done to cut the bands of appropriate size with the help of scalper. Further samples were ethanol precipitated and PCR amplification of those samples was done.

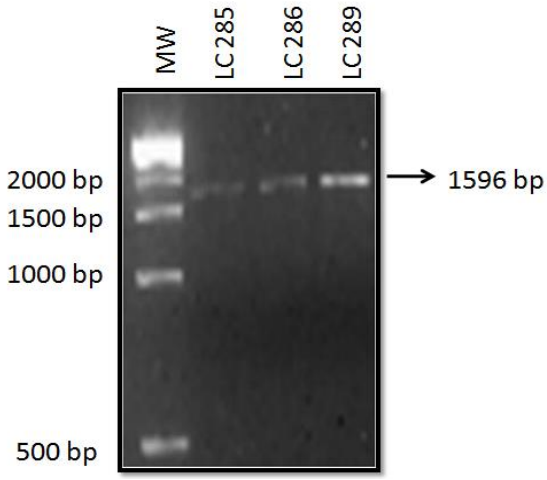

Figure 9: Representing 1.2% Gel Electrophoresis of the PCR amplified product of 1596 bp of XPD gene

6.6.3 **Polyacrylamide Gel Electrophoresis**

The PCR products were then digested with HpaII restriction enzyme capable of excising the PCR product at CCGG site which is specific to HpaII enzyme. The digestion was checked by running the digested samples on 15 % Polyacrylamide gel electrophoresis.

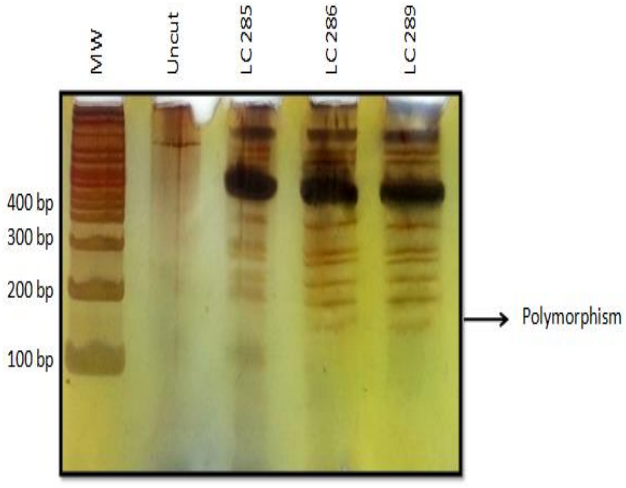

Figure10: Representing 15% PAGE for the restriction digestion of the PCR product using HpaII enzyme

As shown in figure 10, molecular weight of 100bp was loaded in the first lane and in the second lane uncut DNA sample was loaded to compare the digested samples with undigested samples. Polymorphism in the LC285 sample was observed as shown in figure 10. LC 286 and LC 289 samples showed the same digestion pattern. When comparing LC 285 sample with other two samples, it was observed that one band was missing in LC 285 sample *viz.* one polymorphic site was identified.

6.6.4 **PCR amplification**

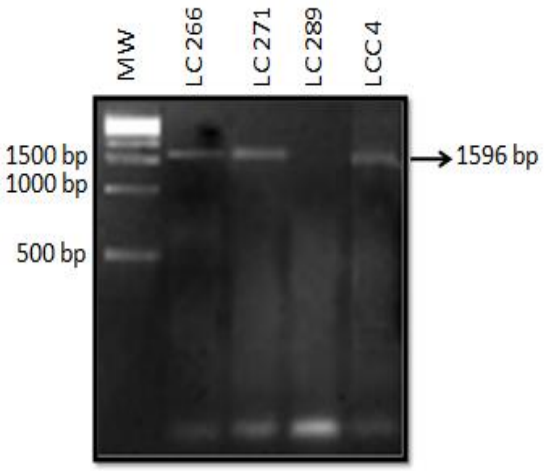

Figure 11: Representing 1.2% Gel Electrophoresis of the PCR amplified product of 1596 bp of XPD gene

As shown in figure 11. PCR amplification of DNA samples was done. Three DNA samples of Lung cancer patients' and one control sample (LCC) was used for amplification. Out of three LC samples, two samples of patients were amplified.

After running the agarose gel electrophoresis of PCR amplified samples, the restriction digestion of the PCR amplified samples was done with HpaII enzyme and then the Polyacrylamide gel Electrophoresis of 12% was run to see the digestion pattern.

6.6.5 **Polyacrylamide Gel Electrophoresis**

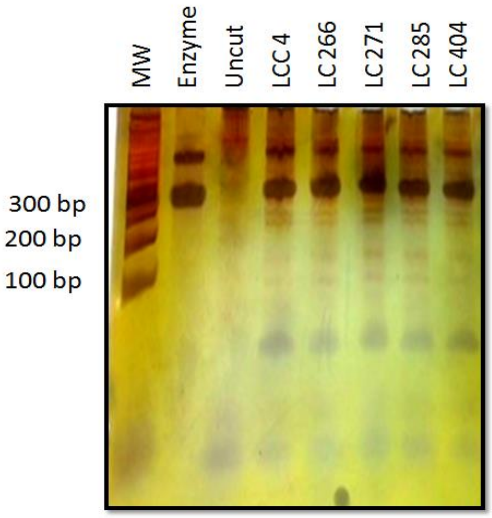

Figure 12: Representing 12% PAGE for the restriction digestion of CCGG sites by HpaII

As shown in figure 12, molecular weight of 100bp was loaded in the first lane. In second and third lane enzyme and uncut DNA sample was loaded. In the fourth lane the sample that was taken as control (LCC 4) was loaded and in the last four lanes digested samples of lung cancer patients' was loaded. As shown in the figure 12, no polymorphism was observed in the digested samples.

CHAPTER 7

DISCUSSION

7. DISCUSSION

SNP (Single nucleotide polymorphism) is the type of variation in the DNA sequences that occur when there is alteration in the single nucleotide in the genome. There are many types of genetic changes but SNPs are probably the most significant genetic changes that increase the risk of diseases. The importance of SNPs comes due to their ability for drug efficacy and their side effects, disease risk and most importantly for ancestral studies. There are huge number of SNPs exist in the database *viz.* dbSNP database. People are more focusing on searching for SNPs basically in the genes that play critical role in the development of diseases.

In 2000, Ford had conducted a study in which single nucleotide polymorphisms (SNPs) were identified in DNA repair genes. When the screening of variations in human DNA repair genes occurred then approximately six G: C to A: T transitions were identified. **In 2006, Jiang and Zhao** also concluded that C to T changes occurred more frequently than other transitions. From the various studies as it is expected that SNPs are arising from transition mutations *viz.* CG→ TG/CA, which is a phenomenon in vertebrate genome due to DNA methylation.

DNA methylation is an epigenetic mechanism that occurs both in eukaryotes as well as prokaryotes. Methyl group is added to the N6 position of adenine and N4 and C5 positions of cytosine but in eukaryotes, methylation occurs only at C5 position of cytosine with the help of specific enzymes named as DNA Methyltransferases (Dnmt). In DNA methylation mainly occurs at CpG dinucleotides in mammalian genome. When methyl group is added to cytosine at 5^{th} position then it forms 5-methyl cytosine which gets further deaminated to form thymine residue. This resulting T: G mismatch is not repaired by any of the specific DNA repair pathway. Therefore CpG dinucleotides are susceptible to mutations *viz.*CG/CG to TG/CA.

We have selected the three genes *viz.* XPC, XPD and XPG that play very important role in one of the DNA repair pathway named as Nucleotide Excision Repair Pathway. It has been found that polymorphisms in these genes play important role in predisposition of cancer. We have analyzed the already existing SNPs in dbSNP database and it is found that majority of SNPs are due to transitions and within those high proportion of SNPs caused due to CG→ TG/CA. This data is strongly favored when we have analyzed the statistical data and odds ratio is coming out of 41.07 folds increase in XPC gene, 40.17 folds increase in XPD gene and 13.64 folds increase in XPG gene and p- value is 0 in all the three genes. This indicates that there is a significantly high proportion of SNPs that arise due to CpG dinucleotides. Exploring this natural phenomenon we have attempted an approach to find the new SNPs. One of the simplest ways to find out the novel SNPs is RFLP analysis from different tetra- nucleotide sites. We have selected the restriction site of HpaII enzyme *viz.* CCGG site to perform RFLP.

To perform RFLP analysis, first we have analyze the sequence of all the three genes to look for the region which has maximum number of CCGG sites in a short stretch of sequence which can be further amplified by PCR. This analysis also includes the exclusion of existing CG-SNPs and CpG Islands from the sequence of the gene. CpG islands usually not methylated that's why there may be lower probability to find the novel CG-SNPs *viz.* CG→ TC/GA. We have also tried to avoid the intronic region of the genes because any change in the coding region *viz.* exons play very important role in gene expression as compare to non coding region *viz.* introns.

Fulfilling the above criteria, we have selected the sequence region of 1596 bp of XPD gene having eight CCGG sites in it out of which 3 CCGG sites already exists as known CG-SNPs and 5 CCGG sites does not exist in dbSNP database. We have performed the simulation of digestion to see what kind of bands pattern will observe when HpaII enzyme will cut the CCGG sites. There are possibilities of large number of restriction patterns with eight CCGG sites in the sequence of diploid genome. This approach is to use restriction pattern of each individual as a fingerprint. We have run the PCR for healthy *viz.* LCC (Lung cancer control samples) and diseased patients' *viz.* LC (Lung

Cancer patients' samples). But we have faced the problems in PCR amplification. In some reactions, DNA amplification was done and in some reactions DNA amplification of samples was not done. There was no consistency in the results of PCR. Exact reason of this problem was not known but there might be possibility of the presence of inhibitors or might be the presence of polymorphism in the sequence of primers.

We have identified one putative polymorphic site as shown in the figure 10. Further confirmation of this putative SNP will only be done by the sequencing method. The approach that we have followed is very much promising but more work is required to be done in this field.

CHAPTER 8

CONCLUSION

8. CONCLUSION

Single Nucleotide Polymorphism occurs normally throughout the person's DNA. To understand basis of the phenotypic variations among the individuals, their susceptibility to diseases, their response to medications *etc.*, requires the understanding of genetic polymorphisms. One of the significant causes of SNPs is DNA methylation. We had analyzed the existing SNPs in the genes that showed majority of SNPs were due to transitions. Within those high proportions, SNPs were due to CG→ TG/CA. This data was strongly favored by calculating the odds ratio and p-value which suggested that there was a significantly high proportion of SNPs due to CG dinucleotides. Exploring this natural phenomenon, RFLP approach was used to search for novel SNPs with HpaII enzyme having restriction site CCGG. We had selected the region of 1596 bp of XPD gene that contained 8 CCGGs in it. Out of 8 CCGGs 3 were already existed in dbSNP database. We had designed the primers for the region of XPD gene and PCR amplification was done with the DNA samples of lung cancer patients. We have identified a putative SNP but it could neither be characterized nor confirmed because the results of PCR were not consistent. The actual reasons of the inconsistency of PCR results were not known but there might be the possibility of the presence of inhibitors or presence of polymorphism in the sequence of primers. Our results show that this approach is a promising method to search for novel SNPs and suggest that methylation and its association with the genetic variations is an important area for research but more research work is required to be done in this field of study.

CHAPTER 9

REFERENCES

REFERENCES

1. Baccarelli, A., Calista, D., Minghetti, P., Marinelli, B., Albetti, B., Tseng, T., & Landi, M. T. (2004). XPD gene polymorphism and host characteristics in the association with cutaneous malignant melanoma risk. *British journal of cancer*, *90*(2), 497-502.
2. Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, *8*(7), 1499-1504.
3. Brookes, A. J. (1999). The essence of SNPs. *Gene*, *234*(2), 177-186.
4. Brueckner, B., Boy, R. G., Siedlecki, P., Musch, T., Kliem, H. C., Zielenkiewicz, P. & Lyko, F. (2005). Epigenetic reactivation of tumor suppressor genes by a novel small-molecule inhibitor of human DNA methyltransferases. *Cancer research*, *65*(14), 6305-6311.
5. Costa, R. M., Chigançcas, V., da Silva Galhardo, R., Carvalho, H., & Menck, C. F. (2003).The eukaryotic nucleotide excision repair pathway. *Biochimie*, *85*(11), 1083-1099.
6. Das, P. M., & Singal, R. (2004). DNA methylation and cancer. *Journal of clinical oncology*, *22*(22), 4632-4642.
7. Dupuy, A., Valton, J., Leduc, S., Armier, J., Galetto, R., Gouble, A., & Sarasin, A. (2013). Targeted gene therapy of xeroderma pigmentosum cells using meganuclease and TALEN™. *PLoS One*, *8*(11), e78678.
8. Ehrlich, M., Gama-Sosa, M. A., Huang, L. H., Midgett, R. M., Kuo, K. C., McCune, R. A., & Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic acids research*, *10*(8), 2709-2721.
9. Forbes, D. A. (2012). What is a p value and what does it mean? *Evidence Based Nursing*, *15*(2), 34-34.
10. Ford, B. N., Ruttan, C. C., Kyle, V. L., Brackley, M. E., & Glickman, B. W. (2000). Identification of single nucleotide polymorphisms in human DNA repair genes. *Carcinogenesis*, *21*(11), 1977-1981.
11. Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, *196*(2), 261-282.

12. Gillet, L. C., & Schärer, O. D. (2006). Molecular mechanisms of mammalian global genome nucleotide excision repair. *Chemical reviews*, *106*(2), 253-276.
13. Handa, V., & Jeltsch, A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. *Journal of molecular biology*, *348*(5), 1103-1112.
14. Hermann, A., Gowher, H., & Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases. *Cellular and Molecular Life Sciences CMLS*, *61*(19-20), 2571-2587.
15. Hotchkiss, R. D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. *Journal of Biological Chemistry*, *175*(1), 315-332.
16. Jansson, K., Alao, J. P., Viktorsson, K., Warringer, J., Lewensohn, R., & Sunnerhagen, P. (2013). A role for Myh1 in DNA repairs after treatment with strand-breaking and crosslinking chemotherapeutic agents. *Environmental and molecular mutagenesis*, *54*(5), 327-337.
17. Jiang, C. & Zhao Z. (2006). Directionality of point mutation and 5-methylcytosine deamination rates in the chimpanzee genome. *BMC genomics*, *7*(1), 1.
18. Johnson, A. D. (2009). Single-Nucleotide Polymorphism Bioinformatics A Comprehensive Review of Resources. *Circulation: Cardiovascular Genetics*, *2*(5), 530-536.
19. Le May, N., Egly, J. M., & Coin, F. (2010). True lies: the double life of the nucleotide excision repair factors in transcription and DNA repair. *Journal of nucleic acids*, *2010.*
20. Li Chunying, Zhibin Hu, Zhensheng Liu, Li-E. Wang, Sara S. Strom, Jeffrey E. Gershenwald, Jeffrey E. Lee *et al.*, (2006). Polymorphisms in the DNA repair genes XPC, XPD, and XPG and risk of cutaneous melanoma: a case-control analysis. *Cancer Epidemiology Biomarkers & Prevention*, *15*(12), 2526-2532.
21. Matsuo, K., Clay, O., Takahashi, T., Silke, J., & Schaffner, W. (1993). Evidence for erosion of mouse CpG islands during mammalian evolution. *Somatic cell and molecular genetics*, *19*(6), 543-555.

22. Prindle, M. J., & Loeb, L. A. (2012). DNA polymerase delta in DNA replication and genome maintenance. *Environmental and molecular mutagenesis*, *53*(9), 666-682.
23. Roberts, R. J., Belfort, M., Bestor, T., Bhagwat, A. S., Bickle, T. A., Bitinaite, J., & Firman, K. (2003). A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic acids research*, *31*(7), 1805-1812.
24. Singal, R., & Ginder, G. D. (1999). DNA methylation. *Blood*, *93*(12), 4059-4070.
25. Takai, D., & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*, *99*(6), 3740-3745.
26. Westergren A., Karlsson S., Andersson P., Ohlsson O., Hallberg I.R. (2001) Eating dif®culties, need for assisted eating, nutritional status and pressure ulcers in patients 2 admitted for stroke rehabilitation. *Journal of Clinical Nursing* 10, 257±269