Single nucleotide polymorphism in mismatch repair pathway genes, and its computational analysis in lung Cancer patients

A

Thesis submitted

in partial fulfillment of the requirement for the award of the degree of

DOCTOR OF PHILOSOPHY

IN

BIOTECHNOLOGY

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Submitted by

SIDHARTHA SINGH

(Reg. No. 901700006)

Under the supervision of

Dr. Siddharth Sharma
Associate Professor

Dr. Manoj Baranwal
Professor

Department of Biotechnology
Thapar Institute of Engineering & Technology
(Deemed to be University)
Patiala-147004, Punjab, India

August 2022

Dedicated to my family.....

CERTIFICATE

Certified that the thesis entitled "Single nucleotide polymorphism in mismatch repair pathway genes, and its computational analysis in lung cancer patients" submitted by Mr. Sidhartha Singh, Reg. no. 901700006 in the partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in the Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, Punjab, is a record of candidate's own independent and original research work carried out by herself under my supervision and guidance. The material embodied in this thesis has not been submitted in part or whole to any other University or institute for the award of any degree.

Dr. Siddharth Sharma
(Associate Professor)
Department of Biotechnology
Thapar Institute of Engineering
& Technology
Patiala, Punjab

Dr. Manoj Baranwal
(Professor)
Department of Biotechnology
Thapar Institute of Engineering
& Technology
Patiala, Punjab

CANDIDATE DECLARATION

I hereby declare that the work being presented in the thesis entitled "**Single nucleotide polymorphism in mismatch repair pathway genes, and its computational analysis in lung cancer patients**" in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy, Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, Punjab is an authentic record of my work during the period from July 2017 to October 2021, under the supervision of Dr. Siddhartha Sharma and Dr. Manoj Baranwal, Associate Professor, Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, Punjab. This report has not been submitted for the award of any other degree or certificate in this or any other university.

Place: Patiala, Punjab

Date: 16-9-22

Siddhartha

Sidhartha Singh

(Reg.No. 901700006)

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

Date:

Dr. Siddharth Sharma
(Associate Professor)
Department of Biotechnology
Thapar Institute of Engineering
& Technology
Patiala, Punjab

Dr. Manoj Baranwal
(Professor)
Department of Biotechnology
Thapar Institute of Engineering
& Technology
Patiala, Punjab

ACKNOWLEDGMENT

In pursuit of this academic endeavor, I feel that I have been singularly fortunate because inspiration, guidance, direction, cooperation, love, and care – all came in my way in abundance, and it seems almost an impossible task for me to acknowledge the same in good term.

With a profound sense of indebtedness and heartfelt gratitude, I thank **Dr. Siddharth Sharma,** Associate Professor and **Dr. Manoj Baranwal**, Professor, Department of Biotechnology (DBT), Thapar Institute of Engineering & Technology (TIET), Patiala, Punjab (Pb), who guided and inspired me in pursuance of this work. Their association with this endeavor of mine will remain a beacon of light to me throughout my life. I sincerely thank them for making me a better person, guiding me, and improving the siliest of my mistakes. Their deep insights helped me at various stages of my research. I also remain indebted for his understanding and support of me during the times when I was really down. Their determined and enthusiastic nature has always inspired me to be consistent in my work. Words would be less to express my thanks to them.

My sincere gratitude to the visionary director of TIET, **Prof. Dr. Prakash Gopalan,** for providing advanced infrastructure and an excellent research environment. I thank **Dr. Inderveer Chana**, Dean, Student Affairs, and **Prof. Ajay Batish**, Dean Partnership, and accreditations, TIET, Patiala, for their precious guidance and constant motivation.

Constructive suggestions and critical remarks of **Dr. M.S. Reddy**, Head DBT, **Dr. Sanjai Saxena**, Professor, DBT, and **Dr. Atul Kumar Upadhyay**, Assistant Professor, DBT, towards amelioration of this work have been of immense help in successfully achieving the objectives. I am highly obliged to them for their suggestions and constant encouragement.

A special thanks to Mr. Ram Naval, Mr. Surinder Pal, Mr. Prabhat, and other lab staff members of DBT, TIET; without whoms' help, the experiments would not have been possible. I am especially indebted to my seniors, **Dr. Charu Bahl** and **Dr. Amrita Singh,** for their constant love, cooperation, and help. I would also like to thank my lab mates, **Ms. Harleen Kaur Walia** and **Ms. Parul Sharma,** who helped in my work.

I am also indebted to my friends Davinder Sillu, Yadveer Kaur, Dr. Vagish Dwibedi, Sheetal, Ramandeep Kaur, Gurleen, Tarun Saini and Manish Sarin for being there to listen when I needed an ear, who became my strength during tough times and keeping faith in me.

I like to express my deep joy to work with all M.Sc and M.tech students who gave me a chance to learn more about my field. Their presence kept my enthusiasm high, and my passion for this field grew during all these years. Prabhjot Kaur, Prerna, Isha were eager to learn more and shared a good bond with them.

Nonetheless, friends and well-wishers know my thick and thin, and there are no ways to express any form of thanking them. Words cannot express my feelings for my parents and my in-laws for their constant unconditional support- both emotionally and financially. I would not be here if it were not for them.

There are many times when you are on the verge of falling into pieces, but the presence of one single inspiration can always keep you going, such as my spouse **Dr. Shweta Singh**. She has been a constant source of strength and inspiration. I can honestly say that only her determination and constant encouragement during my Ph.D. ultimately made it possible for me to see this thesis through to the end.

I would also like to thank my grandfather, **Late Mr. Moti Ram,** and grandmother, **Late Mrs. Jayanti Devi**, who blessed all of us and always taught us the exemplary virtues of living. Thanks to the almighty for bestowing blessing to me.

Date: 16-9-22

Place: Patiala

Sidhartha

Sidhartha Singh

LIST OF PUBLICATIONS

1. Singh, S., Singh, N., Baranwal, M., & Sharma, S. (2022). Association of MLH1-93G> A polymorphisms toward lung cancer susceptibility and its association with clinical outcome in North Indian patients treated with platinum-based chemotherapy. DOI: 10.4103/jcrt.jcrt_465_21
2. Singh, S., Sharma, S., & Baranwal, M. (2021). Identification of SNPs in hMSH3/MSH6 interaction domain affecting the structure and function of MSH2 protein. *Biotechnology and Applied Biochemistry*. https://doi.org/10.1002/bab.2295
3. Singh, S., Singh, N., Baranwal, M., & Sharma, S. (2022). Polymorphisms in the MSH2 gene predict poor survival of North Indian lung cancer patients undergoing chemotherapy. *Biomarkers in Medicine*, *16*(2), 69-82. https://doi.org/10.2217/bmm-2021-0565

WORKSHOPS AND CONFERENCES

- Attended the workshop on Intellectual Property Rights and patenting at TIET on February 17, 2018.

Abstract

Background: Lung cancer is the leading cause of cancer mortality globally, and the critical risk factors are smoking and occupational exposure. Molecular alterations in the repair pathway genes may lead to improper repair and, ultimately, carcinogenesis.

Objective: To evaluate the role of single nucleotide polymorphic variants of DNA mismatch repair (MMR) genes *i.e. MLH1 (rs1800734), MSH3 (rs26279), MSH6 (rs3136228, rs1800932, rs1042821) and MSH2 (rs63749993, rs2303425, rs2303426, rs4987188, rs2303428,* and *rs17217772)* towards lung cancer susceptibility.

Methodology: The study was designed to find out any association between genetic polymorphism of mismatch repair pathway genes and risk of developing lung cancer. This study recruited 500 lung cancer patients from the Department of Pulmonary Medicine, Postgraduate Institute of Medical Education and Research (PGIMER), and 500 healthy controls. These cancer patients were further subdivided in to smaller subgroups based on histology of lung cancer, gender, smoking status (Yes/No and Heavy smokers/ Light smokers). Genomic DNA from lung cancer subjects was genotyped using PCR-RFLP (polymerase chain reaction-restriction fragment length polymorphism) for each polymorphic site under study. Logistic regression was used to analyze subjects to determine any association towards development of lung cancer. Classification and Regression tree (CART) analysis was carried to understand SNP-SNP (single nucleotide polymorphism) interaction. Overall survival analysis was carried out using Kaplan-Meier survival analysis and Cox-regression analysis. Toxicity associated with different chemotherapeutic regimen was also evaluated. Further different computational tools were used to analyze the nsSNPs to evaluate their effect on structure and stability of the protein.

Results: For *MLH1* polymorphism, reduced risk of developing Adenocarcinoma (ADCC) in patients harboring variant (**p= 0.0007**) and combined type genotype (**p=0.008**) was reported. Further, the heterozygous type and combined type genotype of heavy smokers reported a 2-fold (p=0.001) and 1.8-fold increased risk of lung cancer development (**p=0.007**). For *MSH3* polymorphism, females, patients harboring heterozygous type (*GA*) genotype have a 2-fold increased risk (**p=0.04**) for the development of lung cancer. For *MSH6 rs3136228* polymorphism, decreased risk of developing lung cancer in the combined genotype (**p= 0.03**) and 1.4-fold increased risk of developing lung cancer was reported in individuals carrying mutant type genotype (**p=0.03**). Segregation based on histological subtypes, subjects carrying the heterozygous genotype (*GT*) and combined (*GT+GG*) genotype reported a decreased risk of developing adenocarcinoma (**p=0.03**). For *MSH6* (*rs1800932*) polymorphism, our study reported that patients suffering from Small cell lung cancer (SCLC) and harboring heterozygous genotype (*AG*) have a significant increase in median survival time (MST) (**p=0.03**). Furthermore, for *MSH6 rs1042821* polymorphism, patients undergoing docetaxel chemotherapeutic regimen with *CT* genotype reported a significant increase in MST (**p= 0.03**). Four *MSH2* (*118T>C (*rs2303425*), 1032G>A (*rs4987188*), T>C/-6 (*rs2303428*),* and *Asn^{127}Ser (*rs17217772*)*) polymorphisms reported a strong propensity towards the risk of developing lung cancer in subjects having heterozygous and mutant genotype. *T>C/-6* polymorphism showed a 13-fold (**Pcorr=0.0006**)

increased risk of developing lung cancer in patients carrying the variant allele. Our study also found that in *118T>C* polymorphism, minor allele C was significantly associated with poor survival (MST=3.13, **p=0.03**) in lung cancer patients. In lung cancer patients undergoing paclitaxel therapy, heterozygous carriers for the $Asn^{127}Ser$ polymorphism were shown to significantly reduce survival (**p=0.02**). For *2063T>G* polymorphism undergoing irinotecan therapy, subjects with a mutant genotype showed poor survival (**p=0.0004**). Six out of eleven SNPs of four genes (*MLH1*, *MSH2*, *MSH3*, and *MSH6*) located in the coding region were considered for computational analysis. For MSH2 protein, *rs63749993* and *rs17217772* polymorphism was shown to be deleterious by SIFT, PROVEAN, PANTHER, Polyphen-2, PhD-SNP and SNP&GO whereas I-mutant and MuPRO predicted it to have decreased stability. *rs4987188* polymorphism was predicted to have damaging effect by PANTHER, Polyphen-2 only and I-mutant and MuPro predicted it to have decreased stability. *rs1800932 (MLH1)*, *rs1042821* (*MSH6*) and *rs26279* **(***MSH3***)** was shown to have damaging effect by PANTHER, Polyphen-2 whereas only *rs1800932*, *rs1042821* were predicted to have increased stability by I-mutant and MuPRO.

Conclusion: *MLH1* (*rs1800734*), *MSH6* (*rs3136228*), *MSH2* (*rs2303425, rs4987188*, *rs2303428* and *rs17217772*) and *MSH3* (*rs26279*) polymorphism were associated with modulating the risk of lung cancer in North Indian population. Computational analysis of nsSNPs selected for this study showed that they play a key role in altering the structure of proteins, thereby altering the stability of proteins associated with the development of various cancers.

Table of Contents

List of Abrreviations

Abbreviated form	Full Form
NCI	National Cancer Institute
WHO	World Health Organization
NSCLC	Non-Small Cell Lung Cancer
SCLC	Small-Cell Lung Cancer
NEC	Neuroendocrine Cells
SQCC	Squamous Cell Carcinoma
LCC	Large Cell Carcinoma
DNA	Deoxyribonucleic Acid
ROS	Reactive Oxygen Species
RNS	Reactive Nitrogen Species
MMR	Mismatch Repair
BER	Base Excision Repair
NER	Nucleotide Excision Repair
DBSR	Double-Strand Break Repair
MSH2	Mut S Homolog 2
MSH3	Mut S Homolog 3
MSH4	Mut S Homolog 4
MSH6	Mut S Homolog 6
MLH1	Mut L Homolog 1
MLH3	Mut L Homolog 3
PMS1	Post Meiotic Segregation 1
PMS2	Post Meiotic Segregation 2
SNP	Single Nucleotide Polymorphism
NCBI	National Center For Biotechnology Information
UTR	Untranslated Region
MDR	Multifactor Dimensionality Reduction
CART	Classification And Regression Tree Analysis
CGAP	Cancer Genome Anatomy Project
GAI	Genetic Annotation Initiative
HBA1	Hemoglobin A1C
LDLR	Low-Density Lipoprotein Receptor
PET	Positron Emission Tomography
CT	Computed Tomography
FNAC	Fine-Needle Aspiration Cytology
TNM	Tumour Node And Metastasis
ADCC	Adenocarcinoma
CEA	Carcinoembryonic Antigen
IHC	Immunohistochemistry
ADSC	Adipose Tissue-Derived Stem Cells
AIS	Adenocarcinoma In Situ
MIA	Minimally Invasive Adenocarcinoma
TTF-17	Thyroid Transcription Factor 1

APUD	Amine Precursor Uptake And Decarboxylation
PAH	Polycyclic Aromatic Hydrocarbons
NNK	4 4-(Methylnitrosamino)-1-(3-Pyridyl)-1-Butanone
IARC	International Agency For Research On Cancer
NCRP	National Cancer Registry Programme
ICMR	Indian Council Of Medical Research
PBCR	Population Based Cancer Registry
APC	Annual Percentage Change
AAR	Age-Adjusted Incidence Rates
GF	Growth Factors
EGFR	Epidermal Growth Factor Receptor
JAK	Janus Kinas
STAT	Signal Transducer And Activator Of Transcription
PIK3	Phosphatidylinositol-3-Kinase
AKT/PKB	Protein Kinase B
RAS	Rat Sarcoma
RAF	Rapidly Accelerated Fibrosarcoma
MAPK	Mitogen-Activated Protein Kinase
mTOR	Mammalian Target Of Rapamycin
ATP	Adenosine Triphosphate
ADP	Adenosine Diphosphate
TKI	Tyrosine Kinase Inhibitor
MAX	Myc-Associated Factor X
GTP	Guanosine Triphosphate
GDP	Guanosine Diphosphate
TSG	Tumor Suppressor Genes
TP53	Tumor Protein P53
MDM2	Mouse Double Minute 2 Homolog
E2F	E2 Transcription Factor
PTEN	Phosphate Tensin Homolog Deleted On Chromosome Ten
PIP3	Phosphatidylinositol (3,4,5)-Trisphosphate
Bcl-2	B-Cell Lymphoma-2
Mcl-1	Myeloid Cell Leukemia 1
CREB	Camp Response Element-Binding Protein
CML	Chronic Myeloid Leukemia
BCR-ABL	Breakpoint Cluster Region-Abelson Proto-Oncogene
CDDP	Cisplatin, Cisplatinum, Or Cis-Diamminedichloroplatinum(Ii)
ABCB1	Atp Binding Cassette Subfamily B Member 1
CYP3A4	Cytochrome P450 Family 3 Subfamily A Member 4
CYP3A5	Cytochrome P450 Family 3 Subfamily A Member 5
TS	Thymidylate Synthase
DHFR	Dihydrofolate Reductase
GARFT	Glycinamide Ribonucleotide Formyltransferase
CTLA-4	Cytotoxic T Lymphocyte-Associated Molecule-4
FDA	Food And Drug Administration

APC	Antigen Presenting Cell
PD-1	Programmed Cell Death Receptor-1
PD-L1	Programmed Cell Death Ligand-1
ICI	Immune Checkpoint Inhibitor
TIM3	T Cell Immunoglobulin And Mucin-Domain Containing-3
B7H3	B7 Homolog 3
VISTA	V-Domain Ig Suppressor Of T Cell Activation
LAG3	Lymphocyte-Activation Gene 3
TIGIT	T Cell Immunoreceptor With Ig And ITIM Domains
dNTPs	Deoxynucleotide Triphosphates
CPD	Cyclobutane Pyrimidine Dimers
6-4PPs	6-4 Photoproducts
EMS	Ethylmetahnosulfate
SSB	Single-Stranded DNA Binding Protein
DNA pol III	DNA Polymerase III
PCNA	Proliferating Cell Nuclear Antigen
RFC	Replication Factor C
RPA	Replication Protein A
HNPCC	Hereditary Non-Polyposis Colon Cancer
RASGRF2	Ras Protein Specific Guanine Nucleotide Releasing Factor 2
IDL	Insertion-Deletion Loops
BRCA-1	Breast Cancer Gene 1
PWWP	Pro-Trp-Tro-Pro
BPH	Benign Prostatic Hyperplasia
CRC	Colorectal Cancer
EC	Esophageal Cancer
OSCC	Oral Squamous Cell Carcinoma
LTLC	Long-Term Survivorship In Lung Cancer
FEV	Forced Expiratory Volume
dMMR	Deficient MMR
OS	Overall Survival
pMMR	Proficient MMR
FU	5-Fluorouracil
TTR	Time To Recurrence
SIFT	Sorting Intolerant From Tolreant
MAPP	Multivariate Analysis Of Protein Polymorphism
PANTHER	Protein Analysis Through Evolutionary Relationships
PROVEAN	Protein Variation Effect Analyzer
Polyphen-2	Polymorphism Phenotyping V2
PhD-SNP	Predictor Of Human Deleterious Single Nucleotide Polymorphisms
RMSD	Root Mean Square Deviation
TM	Template Modeling
HOXB13	Homeobox B13
MD	Molecular Dynamics

PPARγ	Peroxisome Proliferator-Activated Receptor Gamma
ORR	Objective Response
RECIST	Response Evaluation Criteria In Solid Tumor Group
PGIMER	Post Graduate Institute Of Medical Education And Research
ECOG	Eastern Cooperative Oncology Group
ALT	Alanine Aminotransferase
ULN	Upper Limits Of Normal
SD	Stable Disease
PD	Progression Disease
CR	Complete Remission
PR	Partial Remission
AE	Adverse Events
TRAE	Treatment-Related Adverse Events
GFR	Glomerular Filtration Rate
WBC	White Blood Cells
EDTA	Ethylene Diaminetetra Acetic Acid
SDS	Sodium Dodecyl Sulfate
TBE	Tris/Borate/Edta
TE	Tris-EDTA Buffer
RBC	Red Blood Cells
HR’	Adhusted Hazards Ratio
HR	Hazards Ratio
C.I	Confidence Interval
OD	Optical Density
dsDNA	Double Stranded DNA
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
BSA	Bovine Serum Albumin
NEB	New England Biolab
PAGE	Polyacrylamide Gel Electrophoresis
APS	Ammonium Persulfate
TEMED	N,N,N’.N’ Tetramethylethylene-1,2-Diamine
HWE	Hardy Weinberg Equilibrium
LD	Linkage Disequilibrium
FDR	Flase Discovery Rate
CVC	Cross Validation Consistency
MST	Median Survival Time
EtBr	Ethidium Bromide
OR	Odds Ratio
AOR	Adjusted Odds Ratio
AIC	Akaike Information Criterion
HMM	Hidden Markov Model
MSA	Multiple Sequence Alignment
subPSEC	Substitution Position-Specific Evolutionary Conservation Score
SVM	Support Vector Machine

GO	Gene Ontology
MAF	Minor Allele Frequency
W	Wild
H	Heterozgote
M	Mutant
Pro	Proline
Ser	Serine

List of Symbols

Symbol	Description
%	Percentage
<	Less than
≥	Equals to or greater than
˚	Degree
˚C	Degree Celsius
µg	Microgram
µm	Micrometer
cm	Centimeter
g	Gram
h	Hour
rpm	Revolutions per minute
mg/m2	Milligram per square meter
L	Litre
M	Molar
χ2	Chi-square
mg	Milligram
min	Minute
ml	Milliliter
mm	Millimeter
mM	Millimolar
Mg	Magnesium
nm	Nanometer
df	Degree of freedom
s	Second
U	Unit (activity)
v	Volume
Cl	Chloride
D	Coefficient of linkage disequilibrium
r2	Extent of linkage disequilibrium
µg/ml	Microgram per milliliter
µM	Micromolar
EtBr	Ethidium Bromide
HCl	Hydrochloride
ng/µl	Nanograms per microliter
bp	Base pair
α	Alpha
β	Beta
γ	Gamma
±	Plus minus
NaCl	Sodium chloride

A	Adenine
T	Thymine
G	Guanine
C	Cytosine

List of Figures

Chapter 2

Chapter 4

Chapter 5

List of tables

Chapter 2

Chapter 4

Chapter 5

Chapter 1

Introduction

1.1 Introduction

Cancer is a set of diseases involving abnormal cell growth with the potential to invade or spread to other parts of body. According to National Cancer Institute (NCI), Cancer can be defined as "a state in the human body where cells grow uncontrollably and spread to other parts of the human body." The fast, uncontrollably dividing cell may form tumors (lumps of tissue) which can be cancerous and non-cancerous, also known as benign. Sometimes cancer cells developed at one site can also migrate to other body organs *via* a process called "metastasis". Lung tumor is one of the commonest malignancies and reason for cancer related death throughout the world. In 2020, the number of new lung cancer patients was 19.3 million. It is responsible for an estimated 10 million deaths, globally (Sung *et al.,* 2021). In India, lung malignancy comprises 6.9 percent of all new growth cases and 9.3 percent of all tumor related deaths in both genders; it is the commonest disease and reason for tumor related mortality in men, with the most noteworthy occurrences in Mizoram. In India, lung cancer constitutes 6.9 per cent of all new cancer cases and 9.3 per cent of all cancer related deaths in both sexes, it is the commonest cancer and cause of cancer related mortality in men, with the highest reported incidences from Mizoram in both males and females (Malik *et al.,* 2015). Air pollution and tobacco use are the two leading risk factors for lung cancer in India (Hecht, 1999; Shankar *et al.,* 2019). A hospital-based study in northern India found that 90% of patients with lung cancer were diagnosed at an advanced stage of the disease and that diagnosis and care were delayed (Mathur et al., 2020). Single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide in the genome (or other shared sequence) differs between members of a biological species or paired chromosomes in an individual, When this single base change occurs in the coding region of gene resulting in an amino acid substitution (AAS) then it is termed as nonsynonymous SNP (nsSNP) (Mah et al., 2011). Genetic association studies with SNPs targeting cancer can be divided into two broad categories, investigation of susceptibility and outcomes. The latter seeks to determine prognostic information for survival, complications or response to pharmacological intervention (i.e., pharmacogenomics). To date, nearly all published studies have examined at most a few SNPs or genes, and on a rare occasion variants within genes of a pathway or related biological process, such as DNA repair enzymes (i.e., MLH1, MSH2, MSH3 and MSH6). The etiology of a specific cancer is probably associated with a set of genetic variants and it is notable that, the SNP can be protective in one context, whereas, in another, confer increased susceptibility for example presence of MLH1 in non-Asian population make them more vulnerable for cancer while same MLH1 SNP has no effect in Asian population (Zhu et al., 2016). Since India is a multiethnic society and each particular population residing might have different genotypic frequency of a particular SNP, therefore these populations might have a different risk for lung carcinogenesis and hence lung cancer susceptibility. Thus, the identification of new candidate lung cancer susceptibility genes would yield new insight into lung carcinogenesis/tumorigenesis process, and also help in screening particular populations who would be at a higher risk to develop lung cancer. Furthermore, SNPs can also be associated with outcomes and thus could be applied to clinical decision making. For instance, genetic variants could alter the risk for metastatic or aggressive tumor (Paris et al., 1999). Further, combining experimental and computational tools to analyze SNPs will enable us to give more sophisticated analysis from the ever-expanding data. We will be able to better address the complex issues of gene-gene interactions, when SNPs profiles can be used in clinical venue. This study will further enable us to derive the information about the significance of a specific SNP from a population based study and apply the knowledge to a particular individual (Erichsen et al., 2004).

Chapter 2

Literature Survey

2.1 Lung anatomy

The respiratory system is the network of organs and tissues that aid breathing. This system helps in bringing the oxygen from the outside air to the lungs, pass this oxygen to the bloodstream (oxygenated blood), which takes it to all tissues, and organs of our body for their proper functioning, while returning the blood carries CO_2 from our body organs (deoxygenated blood) and releases it to the air which we breathe out. The gaseous exchange of oxygen and carbon dioxide is based on the partial pressure of the two gaseous. The exchange of gases happens across an extensive epithelial surface area of approximately $70m^{2,}$ which is permeable to gases (Surate *et al.,* 2017).

The respiratory system is divided into two parts: Airways and lung parenchyma. Airways is composed of the trachea, which divides into a bronchus. The bronchus is another form of bronchioles that further divide into alveoli. The gas exchange, including alveoli, alveolar ducts, and bronchioles, forms the lung parenchyma. The texture of the lungs is spongy with a pinkish-grey tint (Chaudhry *et al.,* 2020). Anatomically speaking lung has an apex (blunt superior end of the lung), three borders (edges), and three surfaces (corresponds to the area of the thorax which they face). Each lung also has a smaller unit known as the lobe. These lobes are separated from each other by Fissures (Oblique fissures and horizontal fissures). There are three lobes in the right lung (Superior, middle, and inferior), whereas the left lung has two lobes (superior and inferior). A lobe has multiple broncho-pulmonary segments that divide lobes, and each segment has its tertiary bronchus and artery, which carries blood for supplying oxygen. There is also a covering of pleura (thin layer of mesothelial cells) on lung parenchyma (Chaudhry *et al.,* 2017). The pleura is a serous membrane surrounding the lungs.

Air comes inside the body *via* nose/mouth and moves towards the pharynx, larynx, and trachea successively during breathing. The pharynx can be divided into three parts: "Nasopharynx," “Oropharynx,” and “Laryngopharynx.” The nasopharynx is a pseudostratified columnar epithelium-lined cavity in most aero-digestive tracts. Oropharynx follows it; the middle part of the throat (Pharynx) is a hollow tube approximately 5 inches long, lined with stratified squamous epithelial tissue, and helps swallow the food (Mankowski *et al.,* 2020). The Laryngopharynx (hypopharynx), the last part of the pharynx, is also lined by stratified squamous epithelial tissue for swallowing. It is a point where the pharynx bifurcates anteriorly into the larynx and posteriorly into the esophagus.

2.2 Lung cancer

Lung cancer is a condition in which lung cells experience persistent pathological alterations, such as the formation of pre-neoplastic or premalignant lesions. A system of checks and balances on cells growth is present in a healthy body without cancer cells, allowing cells to divide to create new cells as and when required. Cancer is caused by an error or mutation within the DNA of a cell (Loeb and Loeb, 2000). A tumor mass might grow if these mutations impact cells. Thus, the regulation over cell death and development is disrupted. This tumor mass can be either "malignant" or "benign" in nature (Type of Tumors, 2016). Benign tumors are non-malignant and can be removed from the body. Tumors become malignant when they develop the capacity to enter normal tissues. Such tumors can travel through the bloodstream or lymphatic system, allowing them to spread to other parts of the body (Lung Cancer, 2017). This process of migrating the tumor to other parts of the body is termed metastasis. Lung cancer usually migrates to the following sites: adrenal glands, bones, liver, and brain (Tamura *et al.,* 2015). Lung cancer primarily arises from epithelial cells lining the larger and smaller air pathways; therefore, it can be called bronchogenic carcinoma (Press, 1991). When cancer arises from pleura, it is called mesotheliomas.

2.3 TNM staging

Eight editions of The Tumor-Node-Metastasis (TNM) classification was used for classifying tumor that the International Association gives for the study of Lung cancer (IASLC) (Lim *et al.*, 2018). In this classification scheme, "T" stands for staging, which involves the size and frequency of the primary tumor loco-regional invasions, "N" stands for the extent of lymph node involvement, and "M" stands for the presence/absence of distant metastasis. Since lung tumors do not manifest any symptoms early, most are diagnosed at a much-advanced stage. The tumor stage is assessed by considering both the clinical and pathological factors. The clinical stage is determined by analyzing the results of imaging tests, biopsies, and physical exams, whereas the pathological stage is determined after the surgery (Santos *et al.,* 2007). Generally, the stage is determined only after a confirmed diagnosis and is usually determined by using microscopic detection of tissue by immunohistochemistry, Positron emission tomography (PET) scan, fiber optic bronchoscopy, computed tomography (CT), and fine-needle aspiration cytology (FNAC). After analysis, applicants can be distinguished into eligible for surgery, those with a loco-regional disease, and those with metastatic disease. The lower the staging of the tumor, the better shall be the prognosis (Rami-porta *et al.*, 2009). **Table 2.1** summarizes the parameters commonly used by physicians to identify lung tumor stages. TNM classification and parameters aid in determining tumor level, and the specifics of the same are tabulated in **table 2.2**.

Table 2.1: TNM stage grouping (Eighth edition) (Source: Detterbeck *et al.*,2017)

Stage	Tx	N0	M0
Stage 0	Tis	N0	M0
Stage IA	T1a	N0	M0
	T1b	N0	M0
	T1c	N0	M0
Stage IB	T2a	N0	M0
Stage IIA	T2b	N0	M0
Stage IIB	T3	N0	M0
	T1a	N1	M0
	T1b	N1	M0
	T1c	N1	M0
	T2a	N1	M0
	T2b	N1	M0
Stage IIIA	T4	N0	M0
	T4	N1	M0
	T3	N1	M0
	T1a	N2	M0
	T1b	N2	M0
	T1c	N2	M0
	T2a	N2	M0
	T2b	N2	M0
Stage IIIB	T3	N2	M0
	T4	N2	M0
	T1a	N3	M0
	T1b	N3	M0
	T1c	N3	M0
	T2a	N3	M0
	T2b	N3	M0
Stage IIIC	T3	N3	M0
	T4	N3	M0
Stage IVA	T4	N0	M1a

	T4	N0	M1b
	T4	N1	M1a
	T4	N1	M1b
	T4	N2	M1a
	T4	N2	M1b
	T4	N3	M1a
	T4	N3	M1b
Stage IVB	T4	N0	M1c
	T4	N1	M1c
	T4	N2	M1c
	T4	N3	M1c

Table 2.2: Tumor, node, and metastasis (TNM) definition (Source: Detterbeck *et al.,* 2017)

Primary tumor (T)	
Tx	Tumor in sputum/ bronchial washings but not be assessed in imaging or bronchoscopy
T0	There is no primary tumor on the imaging
T1	A tumor is 3 cm or less in the most significant dimension, surrounded by lung/visceral pleura, not involving the main bronchus
T1a(mi)	Minimally invasive carcinoma
T1a	Tumor 1 cm or less in most significant dimension
T1b	Tumor more remarkable than 1 cm but 2 cm or less in most significant dimension
T1c	Tumor more significant than 2 cm but 3 cm or less in most significant dimension
T2	Tumor more remarkable than 3 cm but 5 cm or less in most significant dimension, *or* involvement of the main bronchus without carina, regardless of distance from carina or invasion visceral pleural *or* atelectasis or post obstructive pneumonitis extending to the hilum
T2a	Tumor more significant than 3 cm but 4cm or less in greatest dimension
T2b	Tumor more significant than 4 cm but fives m or more minor in most significant dimension
T3	Tumor more significant than 5 cm but 7 cm or less in most significant dimension *or* tumor of any size that involves chest wall, pericardium, phrenic nerve, *or* satellite nodules in the same lobe
T4	Tumor more remarkable than 7 cm in most significant dimension or any tumor with invasion of the mediastinum, diaphragm, heart, great vessels, recurrent laryngeal nerve, carina, trachea, esophagus, spine, *or* separate tumor in the different lobe of ipsilateral lung
Regional lymph nodes (N)	
N1	Ipsilateral peribronchial and hilar nodes and intrapulmonary nodes
N2	Ipsilateral mediastinal or subcarinal nodes
N3	Contralateral mediastinal or hilar; ipsilateral/contralateral scalene/supraclavicular

Distant metastasis (M)	
M1	Distant metastasis
M1a	Tumor in contralateral lung or pleural/pericardial nodule/malignant effusion
M1b	Single extrathoracic metastasis, including single non-regional lymph node
M1c	Multiple extrathoracic metastases in one or more organs

2.4 Global Epidemiology of Lung cancer

Lung cancer is the most frequently diagnosed cancer worldwide and is a leading cause of mortality, killing nearly 10 million people in 2020. According to WHO, nearly 2.21 million newly diagnosed lung cancer cases and 1.80 million deaths were reported in 2021. Lung cancer is the main cause of death associated with cancer (18%), followed by colon cancer (9.4%), liver cancer (8.3%), stomach cancer (7.7%), and breast cancer (6.9%). It is also the most prevalent cancer and the leading cause of death among men; in terms of incidence, prostate cancer is most commonly followed by colon cancer, whereas in terms of overall death, liver cancer leads followed by colon cancer. The most frequently diagnosed and leading cause of death in women is breast cancer, followed by colon and lung cancer concerning prevalence and death, respectively (Sung *et al.,* 2021).

According to the Globocan survey for prevalence in men, the first position in higher HDI (Human Development Index) countries is of lung cancer (39 per 100,000). Lung cancer is the leading form of cancer and death among men in 36 and 93 countries. The highest incidence rate is found in Eastern and Western Asia, Eastern and Southern Europe, , Polynesia, whereas the highest incidence rate among men worldwide is present in Turkey (Sung *et al.,* 2021). In 25 countries present in Oceania, North America, and Europe, Lung cancer is the most common cause of mortality among women. Micronesia/Polynesia, Australia, New Zealand, North America, and Northern & Western Europe are leaders in many incidences, with Hungary leading in the highest country-specific rates (Sung *et al.,* 2021). Geographically, the incidence of lung cancer varies between men and women, as depicted in **Figure 2.1**. This trend shows the historical pattern of tobacco smoking worldwide, as lung cancer prevalence is easily linked to evolving smoking patterns in developed and developing countries.

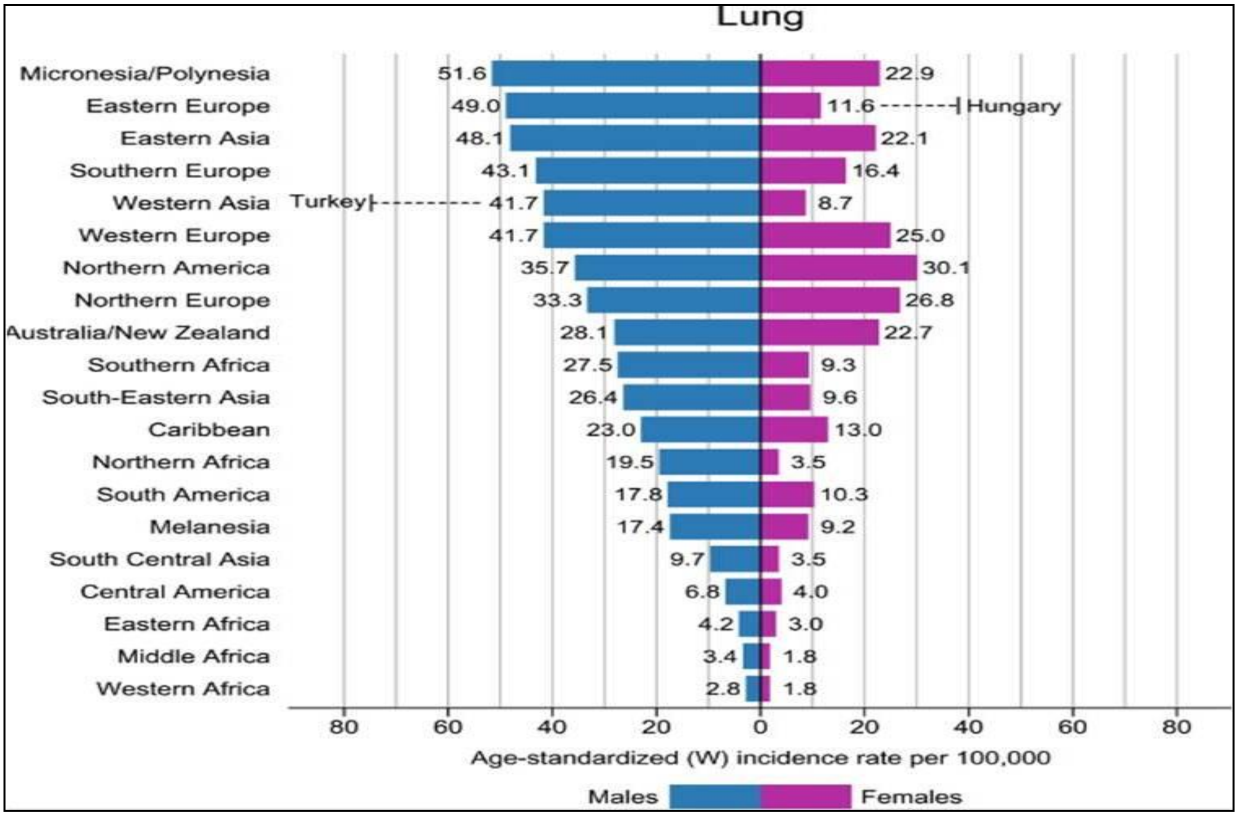

Figure 2.1: Region-Specific Incidence Age-Standardized Rates by Sex for Lung Cancer Among Men and Women in 2020. Rates are shown in descending order of the world (W) age-standardized rate in men, and the highest national rates among men and women are superimposed. (Source: Sung *et al.,* 2021)

2.5 Lung Cancer Epidemiology in India

Lung cancer and oral/mouth cancer are the two most common cancers among men in the Indian subcontinent. In India's southern region and metropolitan cities, the most commonly occurring cancer is lung cancer. Further, in India, smoking and air pollution are the two main risk factors for lung cancer (Mathur *et al.,* 2020).

In the early 19th century, Lung cancer was sporadic in India. In 1959, the number of lung cancer patients was 78.6 per million, representing an increase of more than 2.5 times compared to the early 1950s, with 27.4 cases per million (Behera and Balamugesh, 2004). In December 1981, the National Cancer Registry Programme (NCRP) was launched by The Indian Council of Medical Research (ICMR) with the main aim of gathering information from all the urban and suburban areas of the country (Gajalakshmi *et al.*, 2001). NCRP set up a country-wide network of cancer registries. On January 1, 1982, three hospital in Chandigarh, Thiruvananthapuram, Dibrugarh and population-based cancer registries started functioning in Bangalore, Mumbai, and Chennai

By 2020, the number of estimated lung cancer cases was 2.2 million globally. In India, of all the new cancer cases, 5.4% were diagnosed with lung cancer (https://gco.iarc.fr/today/online-analysis). It is the most common cancer in men and has become a foremost cause of cancer-related death. Mizoram had reported the highest number of incidences in both genders.

As per the National Cancer Registry Program report, 2020, 44% of men and 47.6% of women in Manipur and Mizoram reported having distant metastasis in most lung cancer patients (**Figure 2.2**). Females in 11PBCRs have reported a significant increase in lung cancer cases, Thiruvananthapuram taluk (7.1%; 2005-2016) reported a higher annual percentage change which was followed by Nagpur (6.8%; 2005-2016), and the lowest change reported in Mizoram (0.5%; 2004-2016) (**Figure 2.3**). This data further indicates that 1 in 68 men and 1 in 201 women shall develop lung cancer during their lifetime (Mathur *et al.,* 2020).

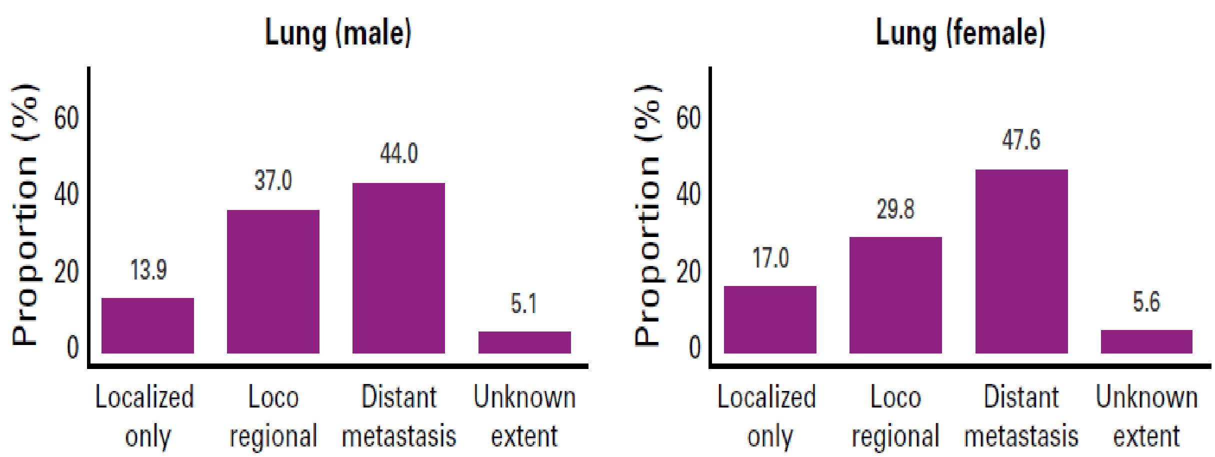

Figure 2.2: Relative proportion (%) of patients according to the clinical extent of disease, 2012-2016 (proportion [%] may not total 100% because of rounding) (Source: Mathur *et al.,* 2020)

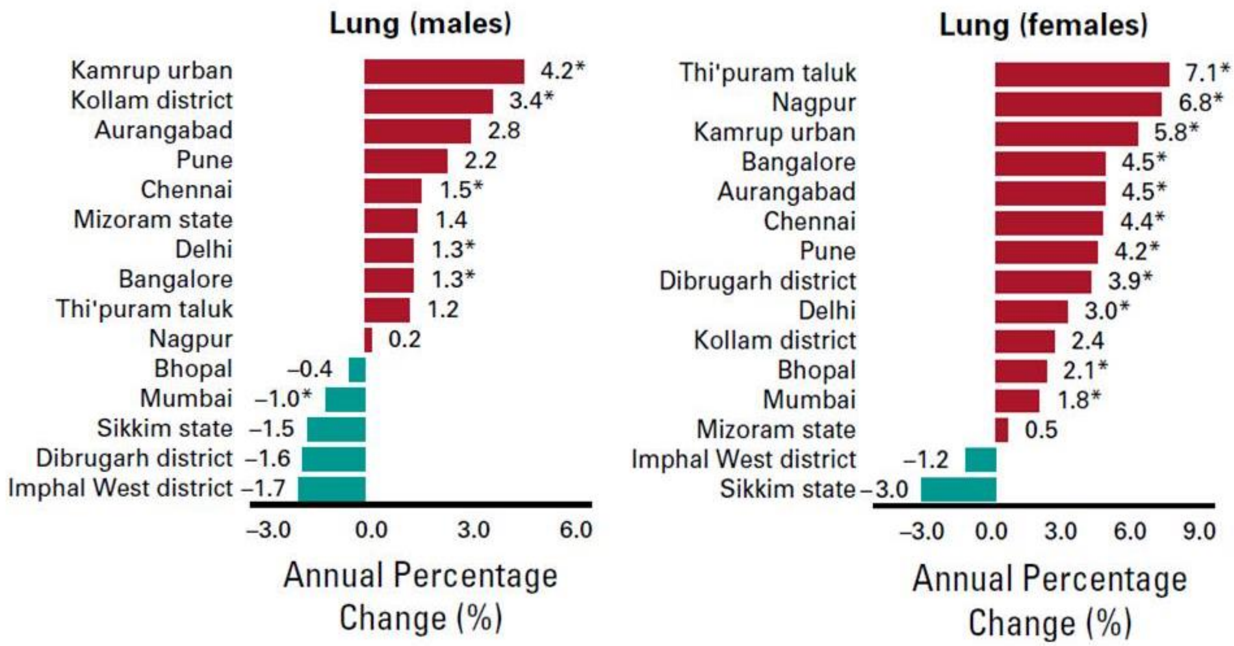

Figure 2.3: Annual percentage change (APC) in age-adjusted incidence rates (AAR) over the period. Calendar years of incidence data for each population-based cancer registry (PBCR) used in trend analysis (Source: Mathur *et al.,* 2020)

2.6 Treatment of lung cancer

The stage and type of lung cancer are the two most important factors determining which treatment strategy shall be employed. The metastatic status of the tumor should also be known before starting any treatment. The standard cancer treatment is surgery, chemotherapy, and radiotherapy (RadiologyInfo.org, 2017). When the tumor is not metastatic and confined to local regions, a local treatment like radiotherapy and surgery is done. When tumors show metastatic behavior (stage III B and IV), chemotherapy and targeted therapy are generally employed. Sometimes, chemotherapy can also be given after local treatment (Chang, 2011). Depending upon the location and stage of the tumor, chemotherapy can also be given even before surgery to shrink cancer cells or in combination with radiotherapy where distant metastasis is evident due to which essential organs are at extreme risk and performing surgery may put an individual at a higher risk. After surgery, it can remove any cancerous cells left behind or lead treatment. When chemotherapy is given as primary treatment, cisplatin/carboplatin in combination with docetaxel, gemcitabine, paclitaxel, and pemetrexed is given to patients diagnosed with NSCLC, whereas SCLC patients were given irinotecan (Behera, 2007).

2.6.1 Platinum-based chemotherapy

2.6.1.1 Docetaxel + cisplatin/ carboplatin

Docetaxel is a taxoid antineoplastic agent frequently used to treat various cancers such as locally advanced or metastatic breast cancer, metastatic prostate cancer, gastric adenocarcinoma, non-small cell lung carcinoma and head & neck cancer. It is a well-established anti-mitotic chemotherapy medication, and it binds reversibly to tubulin with a high affinity in a 1:1 stoichiometric ratio. The treatment success and objective response rate in docetaxel is around 23-33% and 21%, respectively, in NSCLC patients. Cisplatin, cisplatinum, or cis-diamminedichloroplatinum(II) (CDDP) is a platinum-based chemotherapy drug used to treat different types of cancer, including small cell lung carcinoma, lymphomas, ovarian cancer, and germ cell tumors (Aihara *et al.,* 2002). It was the first member of its class, including carboplatin and oxaliplatin. Cisplatin is an antineoplastic agent, belongs to the class of alkylating agents that can add alkyl groups to electronegative groups under normal conditions present in cells. Previous studies have reported that cisplatin-based chemotherapeutic regimens in NSCLC patients have a much greater response and survival rate than drugs without cisplatin (Donnadieu *et al.*, 1991).

2.6.1.2. Irinotecan + cisplatin/ carboplatin

Irinotecan is an antineoplastic enzyme inhibitor that is derived from camptothecin. Camptothecin interacts with topoisomerase-I, which helps DNA replication, recombination, and repair. Topoisomerase I induces reversible single strand breaks that facilitate single DNA strands to move across the break, which reduces torsional strain in DNA. After the torsional strain is removed by topoisomerase-I, the broken strand of DNA is reattached, and chemically unaltered topoisomerase is ready to begin transcription. An active metabolite of Irinotecan: SN-38 binds to topoisomerase-I and irinotecan itself to form a DNA stabilizing complex (Pillot *et al.,* 2006). In addition to its novel cytotoxic mechanism, Irinotecan differs from other chemotherapeutic agents in the multidrug-resistant cell lines that maintain irinotecan sensitivity. It is an authorized drug in the United States to treat metastatic colon cancer. It is either injected alone or combined with fluorouracil for colon cancer treatment. It is used in the treatment of SCLC in conjunction with cisplatin. When used as a single agent, the response rate in NSCLC can range between 15-34% (Pillot *et al.,* 2006).

2.6.1.3 Pemetrexed + cisplatin/carboplatin

Pemetrexed is a multitargeted antifolate that inhibits cancer growth by inhibiting three metabolic enzymes involved in the metabolism of folate, pyrimidine, and purine synthesis. These enzymes are thymidylate synthase (TS), glycinamide ribonucleotide formyltransferase (GARFT) and dihydrofolate reductase (DHFR) (Shih *et al.,* 1997). Pemetrexed is essentially used to treat NSCLC and pleural mesothelioma (Metro *et al.*, 2011). Its toxicity is minimized by vitamin B12, corticosteroids, and folic acid supplementation. The most common side effect includes mucositis, nausea, rash, vomiting, and fatigue. It is approved as a second line of therapy for NSCLC patients (Tomasini *et al.,* 2016). Pemetrexed is generally given in combination with platinum-based drug cisplatin or carboplatin.

2.7 DNA damage and repair

2.7.1 DNA damage repair pathways:

DNA damage response mechanism prevents further type DNA damages from permanently incorporating into the cell DNA. Four DNA damage responses work sequentially: DNA repair pathways, transcriptional response, checkpoints, and apoptosis (De Zio *et al.*, 2013). As soon as cell cycle checkpoints identify any mutation, either DNA repair pathways correct the induced mutation; otherwise, the process of apoptosis/programmed cell death is carried out (Aubrey *et al.,* 2018). The following DNA repair pathways carry out rectification of damaged DNA:

1. Mismatch repair pathways occur directly after DNA synthesis and utilize the parent strand to rectify faulty nucleotide incorporated in the new DNA strand.
2. Base excision repair acts on abnormal DNA bases and removes them *via* a specialized nuclease system followed by DNA synthesis.
3. Nucleotide excision repair acts on heavy base adducts and UV photo products.
4. Double-strand break repair acts on double strands break

2.7.1.1 Mismatch repair pathway (MMR):

2.7.1.1.1 Outline of Methyl directed MMR in E. coli:

The MMR pathway has been well studied and characterized genetically and biochemically in *Escherichia coli*. The following protein components are required for the proper functioning of the MMR pathway: MutS, MutL, MutH, DNA helicase (UvrD), exonucleases (RecJ, ExoI, ExoVII, and ExoX), single-stranded DNA binding protein (SSB), DNA Pol III holoenzyme, and DNA ligase (Lahue *et al.,* 1989, Obmolova *et al.,* 2000). MutS possesses intrinsic ATPase activity and is referred to as mismatch recognition protein (Modrich *et al.,* 1996). X-ray crystallographic studies have confirmed that MutS binds to mismatches in the form of homodimer (Obmolova *et al.,* 2000). MMR is ATP-dependent in *E. coli* and requires functional MutS ATPase. A MutL homodimer interacts with the MutS-DNA complex, and then a MutH restriction endonuclease is recruited and activated by MutL. Since MutL helps in the processivity of helicase II (UvrD) at MMR's initiation site, it plays a crucial part in assembling the functional MMR complex (Sancar *et al.,* 1993, Dao *et al.,* 1998 and Guarné *et al.,* 2004). MutL also has ATPase activity and works as a homodimer (Ban *et al.,* 1998). Some new studies suggest that MutL interacts with DNA pol III *via* its clamp loader subunit, proposing that MutL may be involved in enhancing the binding of DNA pol III to MMR intermediates (Li *et al.,* 2008, de Saro *et al.,* 2006). Since the mismatched base pair in the newly synthesized strand must be corrected, the MMR system should distinguish between error-containing strand and template strand (**Figure 2.4**) (Augusto-Pinto *et al.,* 2003). N-6 position of adenine in dGATC sequence is methylated in *E. coli;* the daughter strand synthesized in replication is hemimethylated at the dGATC site, which differentiate the new synthesized strand from parental strand. MutH, which belongs to the type II restriction endonucleases family, can recognize these hemimethylated sites (Ban *et al.,* 1998, Lee *et al.,* 2005). MutH nicks Hemi-methylated GATC site in un-methylated strand to make an entry point for the excision reaction. In the presence of MutL, helicase II binds on the nick and starts unwinding the DNA moving in the sirection of mismatch, thereby producing single strand DNA that is protected from nucleases by single-stranded DNA-binding protein (SSB) (Ramilo *et al.,* 2002). Helicases and exonuclease remove the error-containing strand, DNA pol III is used to synthesize a new strand, and DNA ligases seal all the gaps between newly synthesized DNA strands (Fukui *et al.,* 2010).

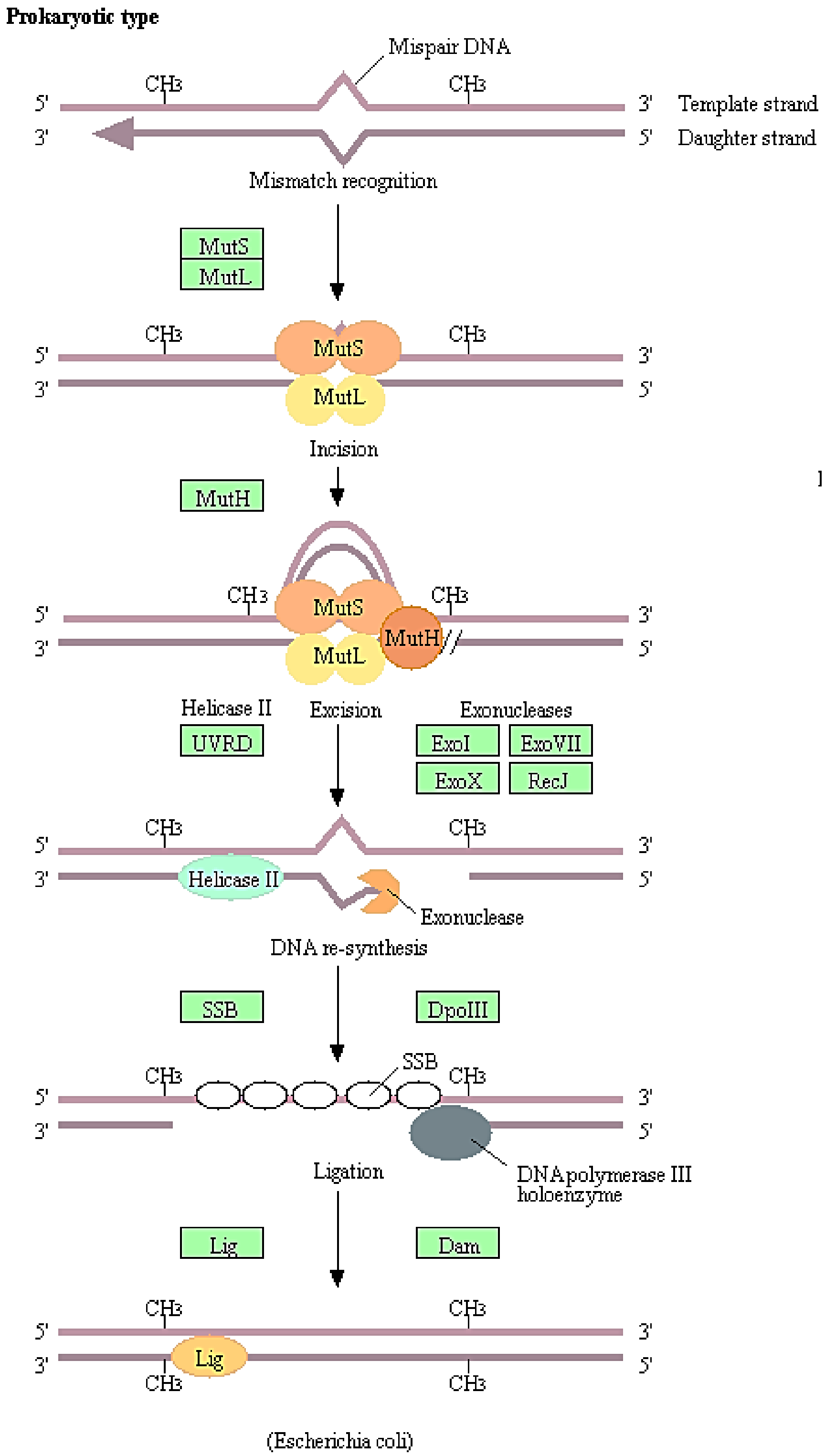

Figure 2.4: Overview of DNA Mismatch repair pathway (MMR) in Prokaryotes (E.*Coli)* (Source: https://www.genome.jp/pathway/ecd03430).

2.7.1.1.2 Overview of Eukaryotic MMR

Mismatch repair system consists of various proteins working simultaneously; they are broadly divided into two groups. One group comprises proteins involved in numerous aspects of DNA metabolism, including DNA recombination and replication like Exo1, RPA, Proliferating cell nuclear antigen (PCNA), replication factor C (RFC), and DNA Pol δ. The functional and structural information of these proteins is extensively studied. The proteins that are more related to the functioning of MMR are in group two, which includes namely MutS-related heterodimers MSH2-MSH3 and MSH2-MSH6 and the MutL-related heterodimers, MLH1-MLH3, and MLH1-MLH2 (MLH1-PMS1 in human), MLH1-PMS1 (MLH1-PMS2 in human) (Erie *et al.,* 2014, Lee *et al.,* 2014)

Gradia and co-workers have shown that MSH2-MSH6 and MSH2-MSH3 heterodimer complexes form a ring and a sliding clamp-like structure under some circumstances. Fishel's lab had done extensive studies on these mismatch repair complexes and further incorporated them into models for MMR mechanism (Gradia *et al.,* 1999) (Fishel *et al.,* 1998). In the absence of nucleotide or ADP, these mispair recognition factors tends to make an unstable ring on base-paired DNA and at the site of the mispair in mispaired DNA (Mendillo *et al.,* 2009) (**Figure 2.5**). In addition to ATP, the proteins bound to base-paired DNA disengage, whereas a stable clamp is formed by mispair-bound proteins that then slides along the DNA and, on linear DNAs, disengages from DNA (Gradia *et al.,* 1999, Wilson *et al.,* 1999, Mazur *et al.,* 2006). In an ATP hydrolysis-dependent process, MLH1-PMS2 can also be recruited by the MSH2-MSH6 complex. MLH1-PMS2 can move on the DNA strand in either direction to check for any gaps in the DNA strand. PCNA, commonly called DNA clip, binds to this complex and helps translocate along the DNA strand. The sliding of the complex activates an endonuclease enzyme, Exo1. This endonuclease has fragile activity on its own but can nick supercoiled DNA after activation by PCNA and RFC. Further it can also be initiated MSH2-MSH6 (or MSH2-MSH3), PCNA, and RFC to nick mimatched DNA having a pre-existing nick. Exo 1 starts from a nick in 5’ directions and travels towards the mismatch in 3’ direction, cleaving up to several hundred nucleotides. Upon deletion of the mispaired DNA, PCNA interacts with DNA polymerase; the complex then moves to mispaired DNA site and begins adding the suitable nucleotide (Kadyrov *et al.*, 2006, Pluciennik *et al.,* 2010) (**Figure 2.5**).

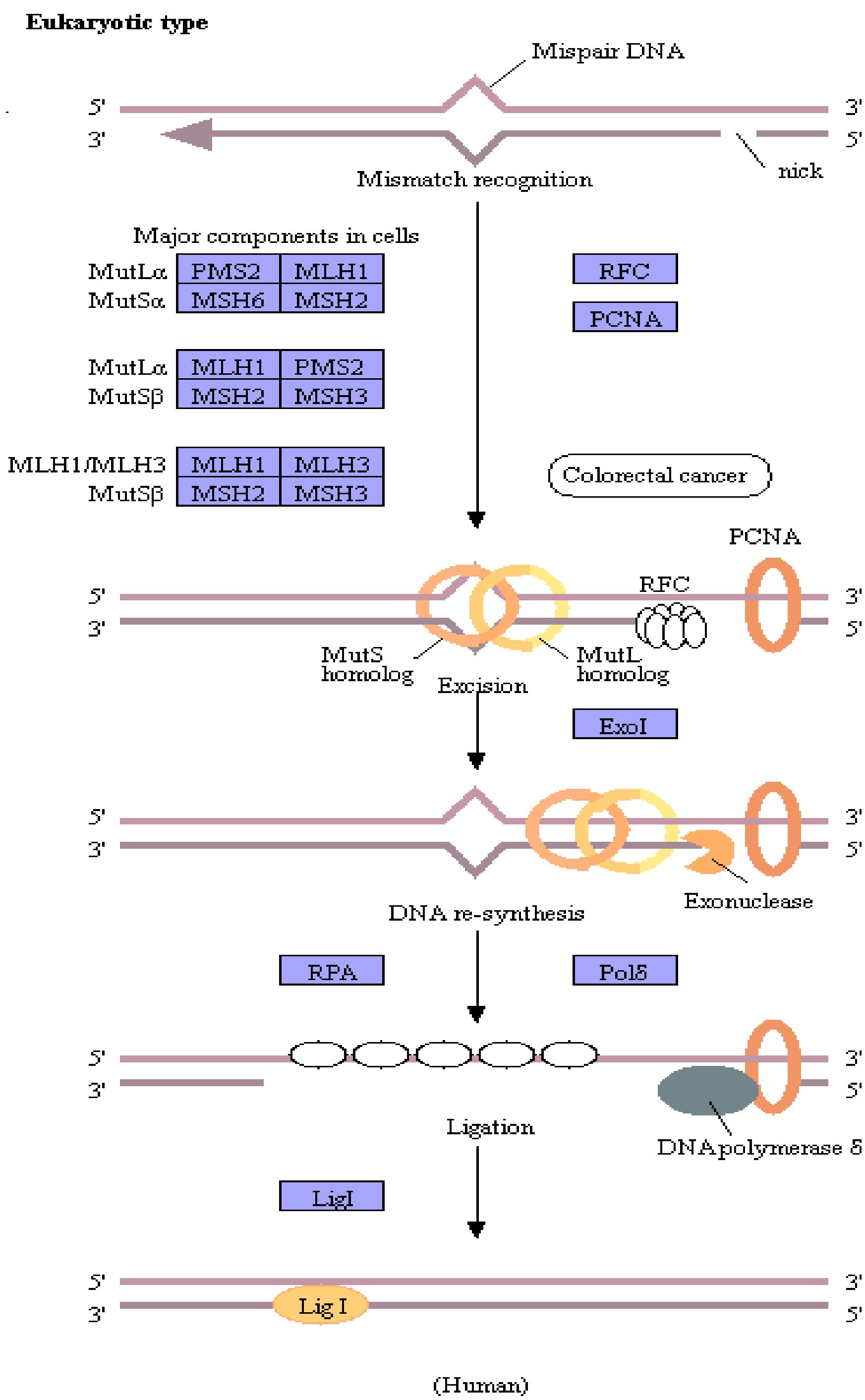

Figure 2.5: Overview of DNA mismatch repair pathway in Eukaryotes (Human) (Source: https://www.genome.jp/pathway/ecd03430)

2.7.2 DNA repair and carcinogenesis

DNA repair plays an essential role in preserving the genomic integrity of an individual. The genes involved in this process, “DNA repair genes,” are also known as “gatekeeper genes” because they act as a protector of the human body and influence the fate of tumor cells by releasing death or growth signals (Dominguez-Brauer *et al.,* 2010). The cell cycle governs the expected growth of cells inside a body. This cell cycle has various checkpoints to check the level of damage to DNA during cell division. As soon as it detects substantial damages, it halts the growth of cells and directs the cell towards a repair mechanism (DNA repair), whereas in some cases where the damage is beyond repair, it directs the cell towards cell death/apoptosis. Hence it prevents the passage of wrong information to the next generation. (Barnum *et al.,* 2014). Condition of excess DNA damage accumulation leads to shifting from DNA repair to apoptosis, as indicated in **Figure 2.6**. In some cases, cells fail to follow the signals leading to apoptosis, due to which the unrepaired damages can be passed on to the next generation, leading to genomic instability (Lengauer *et al.*, 1998). When DNA is synthesized during the cell cycle, and an error is detected at one of the checkpoints, the cell cycle stops until the damage is repaired *via* a DNA repair mechanism.

Several diseases are linked with the damaged DNA repair pathways due to faulty DNA repair genes. Hereditary non-polyposis colon cancer (HNPCC) is the most frequently occurring hereditary colon cancer due to a faulty DNA mismatch repair pathway. Additionally, people having HNPCC are also often reported to have other cancers such as endometrial cancer, small bowel cancer, renal and pelvis cancer. The international Collaborative Group for HNPCC patients manages a database for mutations linked with HNPCC (http://www.nfdht.nl). To date, 300 different predisposing mutations in more than 500 HNPCC families from all over the world are reported in this database. Majorly the mutations in *MSH2* and *MLH1* are reported to affect as the leading cause of faulty mismatch repair pathways (Peltomäki *et al.*, 2001).

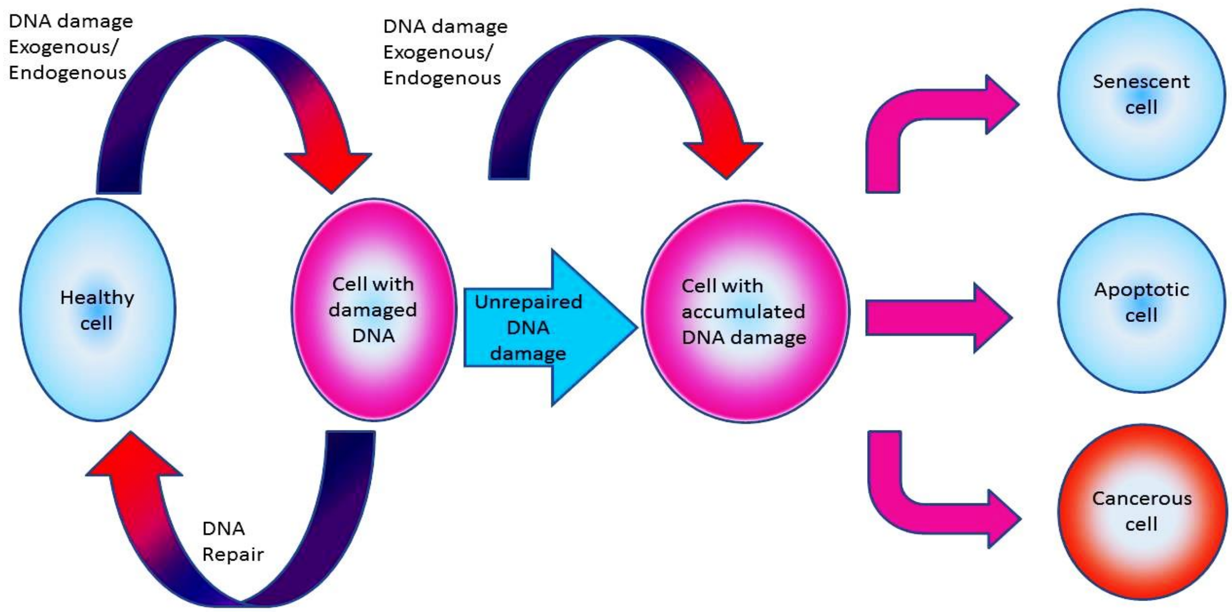

Figure 2.6: Diagramatic representation of the cellular response to DNA damage (Source: Houtgraaf *et al.,* 2006)

The genes that are associated with DNA repair system are as follows:

2.7.2.1 MLH1 (human mutL homolog 1):

MLH1 gene is present on chromosome 3p22.2, composed of 19 exons spanned in the region of 57360 bp, and provides instructions to make a protein that plays a crucical role in DNA repair (**Figure 2.7**). The molecular weight of MLH1 protein is 84.6 kDa and is 756 amino acids long. It is one part of a system of seven mismatch repair proteins that function sequentially to repair human DNA mismatches (Pal *et al.,* 2008). In approximately 13% of colorectal cancer, the mutation in MLH1 protein is more frequently observed than any other DNA mismatch repair protein (Truninger *et al.,* 2005). It is a human homolog of the *E.coli* DNA mismatch repair gene, mutL, and is found to be frequently mutated in hereditary nonpolyposis colon cancer. The primary functions of the *MLH1* gene are mediating protein-protein interaction during the process of identifying mismatches, strand discrimination, and removal (RefSeq, Aug 2017)

This protein helps remove errors when DNA is copied in preparation for cell division. It contains two interaction domains, and one ATPase domain (Domingo *et al.,* 2005). MLH1 is not known for any enzymatic activity. MLH1 forms a heterodimer with PMS2 known as MutLα, part of the DNA mismatch repair system; however, it can also bind to PMS1 or MLH3 to form MutLγ. The heterodimer complexes of MLH1 are accountable for the enlistment of the proteins required for the excision and repair synthesis (Kadyrov *et al.*, 2006). MutLα binds to the heteroduplex MutSα (composed of MSH2 and MSH6), and some accessory proteins or MutSβ (MSH2-MSH3), the PMS2 subunit of MutLα generates a single-strand break near the site of DNA mismatches, thereby generating an entry point for exonuclease EXO1 to remove the strand having the mismatch (RefSeq, Aug 2017).

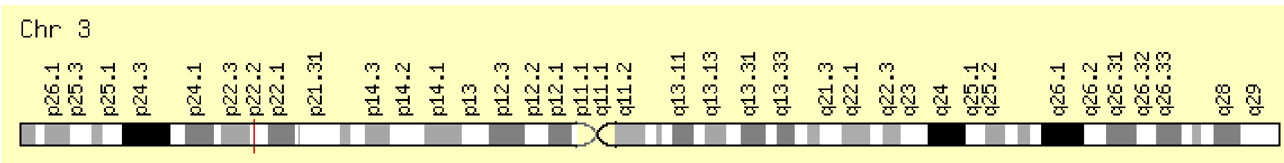

Figure 2.7: Chromosome 3 showing the position of the *MLH1* gene on it (Source: Domingo *et al.*, 2005).

2.7.2.2 MSH2 (human mutS homolog 2):

MSH2 gene, also known as MutS homolog 2, is present on chromosome 2p21, is composed of 16 exons spanned across a region of 80098 bp, and provides guidelines for making a protein that serves a crucial role in DNA repair (**Figure 2.8**). MSH2 protein molecule has a molecular weight of 104.7 kDa with 934 amino acids in its structure. It contains two interaction domains, one for MSH3 or MSH6 and MutL homologs and a binding domain.

MSH2 can bind to MSH6/MSH3 to form the MutSα or the MutSβ complexes, that binds to mispaired DNA, initiating DNA repair. When bound, heterodimers bend the DNA helix and shield approximately 20 base pairs (Domingo *et al.,* 2005). MutSα complex binds to single base mismatches and dinucleotide insertion-deletion loops (IDL). MutSβ only binds to insertion-deletion mismatches. Upon binding to the mismatch, the MutS complex interacts with the MutL complex (composed of MLH1 and PMS2), that is responsible for controlling the downstream MMR events, including strand discrimination, resynthesis and excision. ATP binding and hydrolysis play a essential part in mismatch repair functions. The ATPase activity linked with MutSα controls binding just like a molecular switch: mismatched DNA provokes ADP→ATP exchange, resulting in a discernible conformational transition that converts MutSα into a sliding clamp competent of hydrolysis-independent diffusion along the DNA backbone. This transition is essential for mismatch repair (Blackwell *et al.,* 1998).

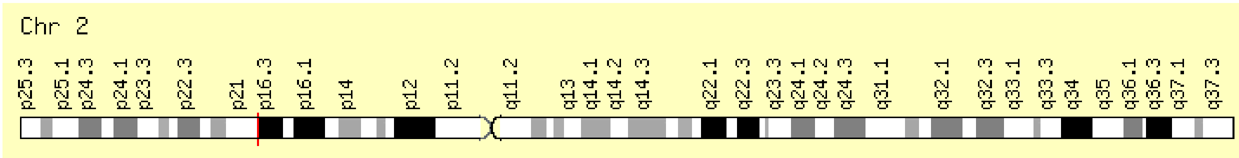

Figure 2.8: Chromosome 2 showing the position of the *MSH2* gene on it (Source: Domingo *et al.,* 2005).

2.7.2.3 MSH3 (human mutL homolog 3):

The *MSH3* gene is a human homolog of the mismatch protein found in bacteria "MutS," which participated in the DNA mismatch repair system. It is located on chromosome 5 (5q14.1) between dihydrofolate reductase (DHFR) and Ras Protein Specific Guanine Nucleotide Releasing Factor 2 (RASGRF2) genes (**Figure 2.9**). It is composed of 24 exons spanned across the region of 222 Kb. MSH3 protein has a molecular weight of 127 kDa and 1137 amino acids in its structure. It is a protein associated with mismatch repair process after DNA replication (Domingo *et al.,* 2006). It is a post-replicative DNA mismatch repair system (MMR) component. It heterodimerizes with MSH2 to form MutSβ, that binds to DNA mismatches, repairing DNA. When bound, the MutSβ heterodimer bends the DNA helix and shields approximately 20 base pairs. MutSβ recognizes large insertion-deletion loops (IDL) up to 13 nucleotides long. After mismatch binding, it forms a ternary complex with the MutLα heterodimer, that is thought to be responsible for directing the downstream MMR events, including strand discrimination, excision, and re-synthesis (Park *et al.*, 2013).

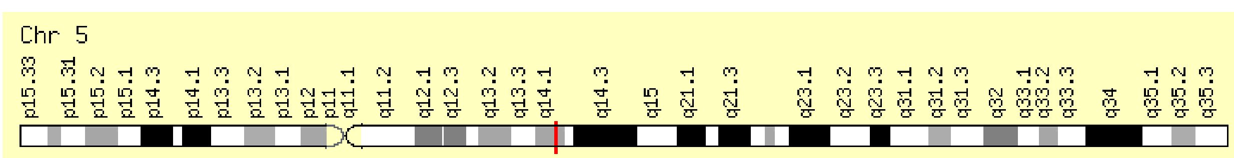

Figure 2.9: Chromosome 5 showing the position of the *MSH3* gene on it (Source: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MSH3).

2.7.2.4 MSH6 (human mutS homolog 6):

The *MSH6* gene is a member of mismatch repair (MMR) genes located on chromosome 2 (2p16.3) (**Figure 2.10**) (Banerjee *et al.*, 2007). Like other MutS homologs, *MSH6* also has intrinsic ATPase activity. In the late G1/early S phase, the *MSH6* gene is 2.5-fold transcriptionally upregulated, and the protein content remains constant throughout the cell cycle (Banerjee *et al.*, 2007).

The *MSH6* gene provides guidelines for making a protein that plays a crucical role in DNA repair. MSH6 protein is 1360 amino acids long and has a molecular weight of 152786Da. MSH6 and BASC or BRCA-1 associated genome surveillance complex is also a part of the substantial multi-subunit protein complex of DNA damage sensors, tumor suppressors, and signal transducers. It heterodimerizes with MSH2 to make MutSα, that binds to DNA mismatches, repairing DNA. When bound, MutSα bends the DNA helix, shields approximately 20 base pairs, and identifies single base mismatches & insertion-deletion loops (IDL)(Cederquist K *et al.,* 2004). After binding, it make a ternary complex with the MutLα heterodimer, which is thought to be responsible for directing the downstream MMR events, including strand discrimination, excision, and re-synthesis. MutSα is also associated with DNA homologous recombination repair.

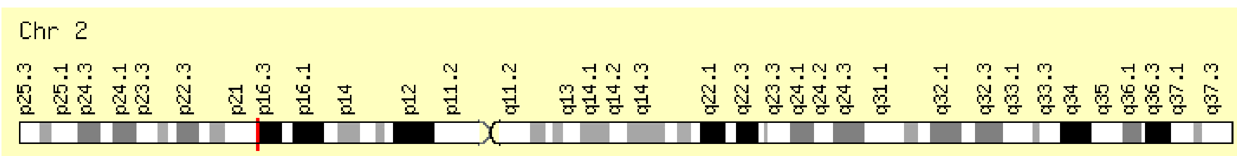

Figure 2.10: Chromosome 2 showing the position of the *MSH6* gene on it (Source: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MSH6).

2.8 Genetic variability in MMR genes and their correlation with cancer predisposition

2.8.1 Genetic variability in *MLH1* genes and their correlation with cancer predisposition

Ensuring the fidelity of DNA replication is the key to safeguarding genomic integrity, and for saving the fidelity of replication, DNA mismatch repair is of prime importance. An analysis of previous literature has revealed that 130 genes are involved in the DNA repair pathways. Single nucleotide polymorphism in a few of them plays an essential role in cancer development of cancer whereas some polymorphism either reduced the risk of developing cancer or showed no association with it. Polymorphism associated with MMR genes and their possible role in cancer development is further discussed. Tanaka *et al.,* 2009 have shown that polymorphisms of the *MLH1* gene can be a risk factor for benign prostatic hyperplasia (BPH) and prostate cancer in the Asian population. They have demonstrated that *MLH1* polymorphism at 219 and 384 codon is not responsible for developing a benign form of prostate cancer. In support of this Fredriksson *et al.,* 2006 have also demonstrated that codon 219 also have no risk for benign prostatic hyperplasia in Finnish population (Fredriksson *et al.,* 2006, Tanaka *et al.,* 2009).

2.8.2 Genetic variability in *MSH2* genes and their correlation with cancer predisposition

For *MSH2 2063T>G* polymorphism, Li and colleagues have reported a 4-fold increased risk towards the development of Esophageal cancer (EC) in individuals harboring mutant and heterozygous type genotype (Li *et al.,* 2012). Another study conducted by Mik and co-workers demonstrated that *1032G>A* polymorphism has no association with sporadic cancer development in the polish populace (Mik *et al.,* 2017). For *MSH2 T>C/-6* polymorphism, Beiner and co-workers have reported a strong association of individuals carrying mutant type genotype (*CC*) with the development of endometrial cancer (Beiner *et al.,* 2006). Moreover, *T>C/-6* polymorphism is located near the splice acceptor site at the exon-intron border and has also been linked with modulating the susceptibility towards sporadic colorectal cancer (Smolarz *et al.,* 2015).

2.8.3 Genetic variability in *MSH3* genes and their correlation with cancer predisposition

hMSH3 gene is located on chromosome 5q14.1 *rs26279* ($Ala^{1045}Thr$) is located on exon 23 and leads to G→A transition (G3133A), thus resulting in alanine (Ala) to threonine (Thr) amino acid change (Nogueira *et al.,* 2018). Smith and co-workers have reported no association between *rs26279* polymorphism and susceptibility towards breast cancer in the Caucasian population (Smith *et al.,* 2008). However, a few studies have found an association between *MSH3* $Ala^{1045}Thr$ polymorphism with an increased propensity towards colorectal cancer (Berndt *et al.,* 2007 and Terrazzino *et al.,* 2012). Another well-known risk factor for lung cancer is tobacco smoking, and few studies have reported an association of increased risk of lung cancer with smoking. Vogelsang and co-workers have reported that smoking and alcohol intake were associated with an increased risk for oesophageal cancer in the South African populace. In their study, authors have compared two populations (Black vs. mixed ancestry) and have found that Black and mixed ancestry population has approximately five and nineteen-times increased risk towards oesophageal cancer due to smoking (Vogelsang *et al.,* 2012).

2.8.4 Genetic variability in *MSH6* genes and their correlation with cancer predisposition

MSH6 gene is located on chromosome 2p16.3. It comprises 1360 amino acid residues and has a molecular mass of 152 kDa. *MSH6* heterodimerizes with *MSH2* protein to form the MutSα complex (Banerjee *et al.,* 2007). Its crucial role is to recognize base–base mismatches and dinucleotide insertion-deletion loops. Mutations in the *MSH6* gene are commonly linked with thyroid, colorectal, colon, and breast cancer (Santos *et al.,* 2018, Liu *et al.,* 2019, Curtin *et al.,* 2009, Smith *et al.,* 2009). The most frequently studied SNP of *MSH6* is *rs1042821*, which alone acts as a risk factor for breast cancer (Lee *et al.,* 2014). SNP *rs1042821* was also linked with reduced risk of primary hepatocellular cancer, and along with *MSH3 rs26279* polymorphism with reduced risk of oesophageal cancer; patients harboring heterozygous genotype of *rs1042821* polymorphism did not report an increased risk of CRC (Zelga *et al.,* 2017). A case-control investigation invloving 106 thyroid cancer subjects and 212 age and gender-matched controls reported that individuals having *rs1042821* mutant genotype are at a increased risk of this cancer (Santos *et al.,* 2018). A very noteworthy link with pancreatic cancer was also reported for *A/G* or *A/A* genotype (Dong *et al.,* 2012).

2.9 Review of Literature for Computational study

Bioinformatics tools have gained recent interest in detecting germline susceptibility, somatic single-nucleotide and small insertion/deletion mutations, copy number alterations, and structural variants in a wide variety of diseases. *In silico* tools and databases are being developed to focus on SNPs and understand their association with different diseases, especially cancer. From various SNPs categories, non-synonymous SNPs (nsSNPs) are of particular interest as they occur in the coding region, which influences protein functionality (Chandrasekaran *et al.,* 2017). The developed computational tools can predict and discriminate between disease-causing and neutral nsSNPs.

SNPs predictions tools are developed based on sequence and structure-based approaches (Tang *et al.,* 2016). Most of the predictions tools being developed use a sequence-based approach compared to structural-based. SIFT, MAPP, PANTHER, and PAREPRO are sequence-based approach prediction software (Mah *et al.,* 2011). The structural-based approach cannot know the information of three-dimensional (3D) structures. However, for understanding the mechanism of the effect of SNPs on the protein phenotypes, the structural approach is more relevant. Prediction tools such as SNPeffect, SNAP, and PMUT integrate both approaches to cover different aspects of SNPs analysis (Mah *et al.,* 2011).

Several studies have used computational tools to predict the nature of mutation and its role in developing diseases. Abdelraheem *et al.,* 2016 have done the computational analysis of MSH6 protein structure and function where they have used various tools like SIFT, Polyphen-2, MUpro, PhD-SNP, and SNP&GO. In this study, SIFT predicted 108 nsSNPs (288 mutations) to be deleterious while 21 nsSNPs (59 mutations) to be highly damaging by SIFT and Polyphen, 19 nsSNPs (53 mutations) and 14 SNPs (39mutations) were predicted by PHD-SNPs and SNP&GO, respectively to be disease-related (Abdelraheem *et al.*, 2016). In silico analysis on proliferating cell nuclear antigen have shown five nsSNPs viz *rs780735449* (Q38R), *rs1050525* (S39R), *rs781573975* (E104G), *rs772308650* (L182W) and *rs753494859* (K248N) were identified as deleterious (Chernapalli *et al.*, 2017). It was reported in this study that the I-Mutant tool led to a decrease in the protein stability for the five high-risk nsSNPs. Deviance in the energy minimization was observed for the variants concerning the native protein. The RMSD (root mean square deviation) and TM (template modeling) values predicted that the mutants were structurally similar to the wild-type protein (Chernapalli *et al.*, 2017). In another report on the human HOXB13 gene, seven nsSNP's (*rs761914407*, *rs8556*, rs13821319i7, *rs772962401*, *rs778843798*, *rs770620686*, and *rs587780165*) were found to be having a damaging and deleterious effect (Chandrasekaran *et al.*, 2017).

Chapter 3
Gaps in study and objectives

3.1 Gaps in study

Polymorphism of DNA repair genes is one of the critical reasons for cancer susceptibility worldwide. In India, lung cancer patients with various genetic polymorphisms are considerably high. Besides this, the local cigarette, commonly known as “bidi,” is more frequently consumed in India, containing exceptional amounts of carcinogens/mutagens. The probability that the North Indian population may be more susceptible to developing lung cancer is very high. On the other hand, studies trying to understand the role of different polymorphisms in developing lung cancer are limited in the Indian populace. DNA repair genes carry polymorphisms that can modulate gene function, altering DNA repair capacity and impairing the DNA repair mechanisms involved in removing DNA adducts. Therefore, identifying polymorphism of DNA repair genes and understanding their correlation to survival in lung cancer patients, susceptibility to lung cancer, and toxicity would yield insights into how the gene polymorphism adversely or favorably influences the chemotherapy outcomes in the North Indian population. The selection of a chemotherapeutic regimen could improve response and, hence, cancer patients' overall survival. *In Silico,* studies on SNPs are relatively new, and so far, no report in the direction of computational analysis of SNPs on the function of the Mismatch repair system is available. Furthermore, these SNPs could provide targets for the future development of personalized therapeutics based on genotype and determine the association between these polymorphisms concerning the clinical outcome, *i.e.*, objective response (ORRs) using RECIST criteria and overall survival (OS) of patients.

3.2 Objectives of the study

1) To study the single nucleotide polymorphic variants present in the mismatch repair (MMR) system
2) To test for associations and susceptibility of different MMR genes polymorphism towards lung cancer susceptibility and clinical outcomes
3) Computational analysis of SNPs to determine their effect on the function of Mismatch repair (MMR) system.

Chapter 4

Materials and methods

4.1 Design of the study

This study is a prospective non-interventional study done in newly diagnosed lung cancer patients recruited from the Lung Cancer Clinic (New OPD, PGIMER) and in the Department of Pulmonary Medicine patients, PGIMER. This prospective cohort study has enrolled all diagnosed (either cytological or histological) lung cancer patients on or before starting irinotecan, docetaxel, and Pemetrexed-based palliative chemotherapy from the places mentioned above. The subjects were analysed after subdividing them in two small groups based on histology (Small cell lung carcinoma, Adenocarcinoma and Squamous cell carcinoma), sex, and smoking status (Smokers and non-smokers, Heavy and light smokers). The study aims to find an association between risk of lung cancer and polymorphism in DNA mismatch repair pathway genes. However since in truly prospective cohorts, the study subjects are recruited prior to any cancer diagnosis hence the study design of the current study entails a significant risk of bias in reporting the lifestyle questionnaire data.

Patients have been evaluated at baseline for complete blood count, including absolute neutrophil count with other biochemical parameters like renal function tests, liver function tests, chest X-ray, contrast-enhanced computed tomography (CT) scan of thorax (including upper abdomen). Demographic characteristics (including age, gender, PS), histological type, stage of disease, and details of smoking status have been noted for all patients.

Before starting platinum-based doublet and second-line chemotherapeutic regimens like irinotecan, docetaxel, and pemetrexed, 10ml of the blood sample was collected for the study. A complete blood count was done on day 10 of the first chemotherapy cycle and subsequent cycles if indicated. Complete blood count and renal and liver function tests were done before each chemotherapy cycle and 3-4 weeks after the last cycle of chemotherapy.

This study recruited 500 lung cancer patients from the Department of Pulmonary Medicine, PGIMER, and 500 healthy controls. The researchers involved in the project explained the objectives and importance of this study to patients and controls recruited. All patients shall sign an informed consent form for blood sample collection. The Ethics Committee of Post

Graduate Institute of Medical Education and Research (PGIMER) Chandigarh shall approve the proposed study. The ethical clearance letter is attached in Appendix-I. Informed written consent was obtained from all participants or their representatives, as attached in Appendix II.

Patients meeting all the following requirements were enrolled for this study:

- Diagnosis of lung cancer (NSCLC or SCLC) is confirmed either by histology or cytology.
- Stage I,II, III or IV disease.
- No age, gender, smoking, histology, or staging restrictions were applied.
- Untreated and intent to treat with definitive chemotherapy (Treated with platinum-based drugs cisplatin/ carboplatin, either as the first or second line).
- An Eastern Cooperative Oncology Group (ECOG) performance status (PS) of 0–2.
- At least one bi-dimensionally measurable lesion, according to the RECIST criteria
- Adequate organ function, defined as absolute neutrophil count>1500/μL, platelet count >100,000/μl, and levels of creatinine, liver enzymes, and alanine aminotransferase (ALT) less than two times the upper limits of normal (ULN).
- Written informed consent was obtained.

The exclusion criteria included the following:

- Previous history of any cancer.
- Active infection or immunosuppression (HIV).
- Patients were taking any systemic steroids.
- Patients are suffering from a chronic form of diarrhea.
- Patients who are undergoing chemotherapy.

Along with 500 lung cancer patients, 500 unrelated subjects having no history of any type of cancer who visited PGIMER for regular health check-ups were also recruited as a controls group. The interviewer filled out a detailed questionnaire asking for information about demography, age, gender, smoking, and drinking habits for both cases and the controls population. Every precaution was taken to match the control population with the test population regarding age (±ten years), sex, and smoking habits. Smokers in the Indian population are used to smoking local cigarettes called "bidi" (a native cigarette-like stick of coarse tobacco hand-rolled in a dry temburni leaf). Pack-years was calculated as an indicator of cumulative smoking exposure using the following formula:

$$\frac{\text{number of cigarettes/bidis smoked per day}}{20} * number\ of\ years\ smoked$$

The essential medical information of cases, such as histology, TNM classification, clinical staging, performance status (KPS and ECOG), chemotherapeutic regimen, and objective response following the Response Evaluation Criteria in Solid Tumor Group (RECIST) criteria, was obtained from the hospital's medical records.

4.1.1 Chemotherapeutic regimen

The patients in this study were given platinum-based chemotherapy as these cases were inoperable. The chemotherapy drugs given included carboplatin and cisplatin and 3rd generation agents such as pemetrexed, irinotecan, paclitaxel, or docetaxel. These drugs were administered intravenously at a fixed dosage based on the body's surface area. The combinatorial therapy involved 500 mg/m^2 pemetrexed, 100 mg/m^2 irinotecan, or 75 mg/m^2 docetaxel (administered as a 1-h infusion, followed by cisplatin 65 mg/m^2 administered over 3-hrs as an intravenous infusion). Vitamin B12 and routine folate supplementation were given to all the patients. Certain drugs such as dexamethasone, granisetron, and ranitidine were administered as a part of the premedication routine. The treatment procedure was repeated every 3-4 weeks, and the chemotherapeutic drugs were administered intravenously. Four cycles of chemotherapy were given as per standard protocol followed by PGIMER, which assessed tumor response. If there was a manifestation of critical toxicity or clinic-radiological proof of disease continuation before completing four cycles, the tumor response was evaluated, and, if required, chemotherapy was stopped. Patients showing an objective response to chemotherapy were given two extra cycles limited to a maximum of six cycles.

4.1.2 Follow-up and response assessment

After every two months, all recruited subjects were followed up *via* telephonic conversation until the end of the study or the patient's death. The survival time was calculated from when the patient was diagnosed with lung cancer to death. The primary endpoint of this investigation was overall survival; the second endpoint was the association between various genetic polymorphic variants and treatment responses. Response Evaluation Criteria in Solid Tumors (RECIST) criteria evaluated tumor response. According to RECIST criteria, subjects were stratified into "responders," those who show complete remission (CR) or partial remission (PR), and "non-responders," those who show stable disease (SD) or progression disease (PD).

- CR: Complete disappearance of all target lesions.
- PR: At least a 30% decrease in the sum of diameters of target lesions, taking the baseline sum diameter as reference.
- PD: At least a 20% increase in the sum of diameters of target lesions, taking the smallest sum on study (this includes the baseline sum if that is the smallest on the study). In addition to the relative increase of 20%, the sum must also demonstrate an absolute increase of at least 5 mm. (Note: the appearance of one or more new lesions is also considered progression).
- SD: Neither sufficient shrinkage to qualify for PR nor sufficient increase to qualify for PD, taking as reference the minor sum diameters while on a study (Eisenhauer *et al.*, 2009).

  Adverse events (AEs) were noted and classified as standard toxicity criteria (CTC) version 3.0. Treatment-related adverse events (TRAE) were assessed for subjects receiving three or more chemotherapy cycles and those who have received at least one chemotherapy cycle. As far as hematological toxicity is concerned, leukopenia, neutropenia, thrombocytopenia, and anemia were evaluated. For determining Nephrological toxicity, Glomerular filtration rate (GFR) was determined. Gastrointestinal toxicity was also evaluated. The standard PGIMER protocol was followed to determine TRAEs. All the chemotherapy patients for at least two cycles were investigated for treatment-related adverse events (TRAEs).

4.2 DNA Extraction

4.2.1 Genomic DNA Extraction

Blood samples of cases and controls were used to isolate genomic DNA as per the protocol of Bartlett and White (2003) with slight modifications.

- 5ml of whole blood was mixed with an equal volume of washing buffer (320mM sucrose, 1.5% Triton X-100, 20mM Tris-HCl (pH 8.0), and 5mM $MgCl_2$) and centrifuged at 3,500 rpm for 10 min at 4˚C to separate white blood cells (WBCs) from whole blood. This step was repeated three times to increase the amount and quality of white blood cells.

- The RBC-free pellet (white-colored) was incubated for lysis at 45˚C (overnight) by mixing it with an equal volume of WBC lysis buffer: 400 mM Tris-HCl (pH 8.0), 60 mM EDTA, 150 mM NaCl, 100 µg/ml proteinase K, and 1% SDS.
- The next day, proteins are removed from the lysed mixture with the addition of phenol/chloroform/isoamyl alcohol solution (25:24:1). This mixture was given a gentle shake and centrifuged at 8000 rpm for 10 min at 4˚C, and two transparent layers were formed. The upper aqueous layer (containing soluble DNA) was removed and put in a new vial; the step was repeated until the exhaustion of the upper layer.
- In the new vial, an equal volume of chloroform/isoamyl alcohol (24:1) solution was added to the previously separated aqueous layer, and after giving a gentle shake, it was centrifuged at 6500 rpm for 5 min at 4˚C to remove the aqueous layer into a separate vial.
- To this separated aqueous layer equal volume of chilled isopropanol or 2.5 times of chilled absolute ethanol was added to precipitate the soluble DNA. The thread-like structure of DNA was stored at 20˚C for 1-2 hours, and then after centrifugation at 12000 rpm for 10 min, the pellet of DNA was formed.
- The pellet obtained was given ice-cold ethanol (70%) washing by adding 1ml and centrifugation at 12000 rpm for 5min. The purified pellet was then air-dried to remove any content of ethanol.
- DNA pellet was then suspended in 100µl of sterile Tris-EDTA buffer.

4.2.2 Qualitative and quantitative estimation of genomic DNA

Qualitative estimation of isolated DNA was done using agarose gel electrophoresis. 0.8% agarose gel with 0.5 µg/ml ethidium bromide was used to visualize the isolated DNA. The sample was mixed with 6X loading dye and diluted with sterile water before loading into the gel wells. TBE gel of 0.5X concentration was used for running the electrophoresis unit. After the run was complete, the gel was visualized under a UV-transilluminator to see the quality of DNA. A simple straight line and a smear indicate isolated DNA's excellent and impaired quality, respectively (Johansson, 1972).

Spectrophotometric analysis was used to complete the quantitative estimation of DNA using Nanodrop (NanoDrop ND-1000 spectrophotometer). A single unit of absorbance (1 OD) for dsDNA is comparable to 50ng/l of DNA at 260nm wavelength. Quantity of DNA in samples was calculated by determining absorbance at 260nm and multiplying it with 50ng/l and dilution factor if any (as shown below):

Concentration (g/ml) = Absorbance at 260nm X 50 g/ml X Dilution factor.

The purity of the sample was also estimated by calculating the ratio of absorbance at 260 and 280nm. A ratio of 1.8 indicates that the DNA sample is pure; a ratio less than 1.8 indicates contamination of protein, whereas a ratio of more than 1.8 indicates RNA or phenol impurities (Desjardins and Conklin, 2010). After complete analysis, the DNA was diluted to 100ng/l and stored at 20˚C for further use

4.3 Genotyping of polymorphic sites of Mismatch repair gene (*MLH1*, *MSH2*, *MSH3*, and *MSH6*)

4.3.1 Genotyping of *MLH1* polymorphism by PCR-RFLP

Genotype analysis was complete by PCR-RFLP technique utilizing the primers, restriction enzymes, and annealing temperature as previously reported (Mik *et al.*, 2017), tabulated in **Table 4.1.** For amplification of polymorphic sites, a standard PCR was used. The PCR reaction was carried out in a 25 µl mixture composed of 1X PCR buffer, 0.5 µM of forward and reverse primer, 100 g/ml bovine serum albumin (BSA), 1.5 mM $MgCl_2$, 200 µM dNTPs, and 1U Taq polymerase (DNAzyme, Thermo Scientific) and 200 ng DNA. The PCR was performed at the following temperature settings: melting step/initial denaturation at 95°C for 5 min; denaturation at 94°C for 30sec (29 cycles); annealing step at 60°C for 45sec; primer extension at 72°C for 30sec and final extension at 72°C for 5 min.

Restriction fragment length polymorphism

After completion of DNA amplification, the PCR product (387bp) was incubated with five units of *PvuII* restriction enzymes procured from Takara Bio USA Inc. 10 µl of each PCR amplicon was treated overnight with five units of restriction enzyme, restriction buffer, and water to make the final volume to 20 µl. This reaction mixture was incubated overnight for 10-12 hours at 37˚C temperature. The restriction product was visualized using the 10% polyacrylamide gel electrophoresis technique. Two different individuals analyzed the results to remove any biasedness. Further, randomly selected 20% of the samples were analysed again, and the results were 100% reproducible.

Table 4.1: List of primers and restriction enzymes used for *MLH1* genetic variant

SNP	**SNP ID #**	**Primers**	**PCR product size (bp)**	**Restriction enzyme**	**Fragments identifying genotypes (bp)**
-93G>A	*rs1800734*	5'- CCG AGC TCC TAA AAA CGA A -3' – forward primer 5'- CTG GCC GCT GGA TAA CT -3' – reverse primer	387	*PvuII*	GG 180, 207 GA 180, 207, 387 AA 387

Polyacrylamide gel electrophoresis (PAGE)

- The separation of nucleic acids and protein can be achieved by polyacrylamide gels that work on the principle that when an external electric field is applied, charged molecules moves towards an electrode with the opposite sign in a buffer. PAGE gel is a cross-linked polymer of bis-acrylamide and acrylamide. The total concentration of monomeric units of bis-acrylamide and acrylamide determines the pore size of the gel. Ammonium persulfate (APS) is used to initiate the polymerization, and this reaction is catalyzed by TEMED (N, N, N, N-Tetramethylethylenediamine). The resolving power of polyacrylamide gels is very high, but the range of separation is relatively minimal. When these gels are used to separate DNA molecules, they can be used only to separate fragments lower than 500bp. After ethidium bromide staining, the gels were further analyzed by viewing them under a UV transilluminator. This protocol used was given by Sambrook and Russel 2006 with slight modifications.
- A free radical generating system starts polymerization.
- The rate of generation of free radicals from persulfate was hastened by TEMED, which catalyzes polymerization.

- The persulfate free radicals convert acrylamide monomers to free radicals which react with inactivated monomers to begin the polymerization chain reaction
- The polymer chains are randomly cross-linked by bis-acrylamide forming a gel.
- The chemical required for this technique are 5X TBE (Tris borate EDTA), 6X loading buffer, 30% Acrylamide solution [(29:1) acrylamide: bis-acrylamide], 10% APS.

Ethidium bromide staining

Ethidium bromide (EtBr) solution was used for staining native-PAGE gels. It is prepared by adding 10 µl of EtBr having a concentration of 0.5µg/ml in 50ml of sterile water. The gel was dipped in EtBr solution for approximately 30-40 min in the dark. After staining, the gel was visualized in a GelDoc system employing UV-transillumination to produce results. The band size after restriction enzyme digestion was determined by comparing the bands with the DNA ladder bands in the same gel.

4.3.2 Genotyping of *MSH2* polymorphism by PCR-RFLP

Genotype analysis was complete by PCR-RFLP technique utilizing the primers, restriction enzymes, and annealing temperature as previously reported with slight modifications (Sun *et al.*, 2014, Jung *et al.*, 2006, Mik *et al.*, 2017, Worrillow *et al.*, 2003, Smolarz *et al.*, 2015). The list of primers used is listed in **Table 4.2**. For amplification of polymorphic sites, a standard PCR was used. The PCR reaction was carried out in a 25 µl mixture composed of 1X PCR buffer, 0.5 µM of forward and reverse primer, 100 g/ml bovine serum albumin (BSA), 1.5 mM $MgCl_2$, 200 µM dNTPs, and 1U Taq polymerase (DNAzyme, Thermo Scientific) and 200 ng DNA. The PCR was performed at the following settings: 95 ˚C for 5 min and 94 ˚C for the 30s (denaturation); 48 ˚C (*2063T>G*), 59 ˚C (*118T>C*), 58 ˚C (*IVS1+9G>C*), 50 ˚C (*1032G>A*), 64 ˚C (T>C/-6), 64 ˚C (*Asn^{127}Ser*) for 45s (annealing); followed by 72 ˚C for 30s (extension) for 29 cycles and the final extension for 5 min at 72 ˚C. The PCR product of *2063T>G*, *118T>C*, *IVS1+9G>C*, *1032G>A*, *T>C/-6*, and 380A>G were 183, 190, 159, 252, 150, and 260 bp was visualized on 1.5% agarose gel and then digested with five units of *Hsp92 II*, *BseN I*, *Bgl I*, *Hinf I*, *Pst I* restriction enzymes (Takara Bio USA Inc.) respectively at 37 ˚C. The digested products were visualized on 10% Native-PAGE after staining with EtBr (procedure as described in 4.3.1) for *2063T>G* /, *118T>C*, *IVS1+9G>C*, and *T>C/-6* variants, whereas for *1032G>A* and *Asp^{127}Ser* variants the digested samples were run on 2.5% agarose gel that was dissolved in 0.5X TBE with 0.5μg/ml ethidium bromide. The electrophoresis was run in 0.5X TBE at 50 volts/cm. The digested samples are mixed with 2μl of 6X loading buffer containing sucrose, bromophenol blue, and xylene cyanol. The resolved gel was visualized under the UV trans-illuminator, and a restriction pattern was observed. To remove the biasedness, two different individuals analyzed the results, and a further 20% of the samples were used to reproduce the results that resulted in 100% reproducibility.

Table 4.2: List of primers and restriction enzymes used for *MSH2* genetic variant

Codon for protein	**SNP**	**SNP ID #**	**Primers**	**PCR product size (bp)**	**Restriction enzyme**	**Fragments identifying genotypes (bp)**
Met^{688}Arg	c.*2063T>G*	*rs63749993*	FP-5'-CGC GAT TAA TCA TCA GTG-3' RP-5'-ATG GCA CAA AAC ACC CAA-3'	183	*Hsp92 II*	TT (WT)- 25,158 GG (MT)- 183 TG (HT)-25,158,183
	118T>C	*rs2303425*	FP-5'-AGG CAT GCG CAG TAG CTA AA-3' RP-5'-CCC ACA CCC ACT AAG CTG TT-3'	190	*BseN I*	WT-190 HT-190,118,72, MT-118,72
	IVS1+9G>C	*rs2303426*	FP-5'-GAC CGG GGC GAC TTC TAT AC-3' RP-5'-AAA GGA GCC GCG CCA CAA GG-3'	159	*Bgl I*	WT 159 MT 113,47 HT 159,113,47
Gly^{322}Asp	*1032G>A*	*rs4987188*	FP-5'-GTT TTC ACT AAT GAG CTT GC-3' RP-5'-AGT GGT ATA ATC ATG TGG GT-3'	252	*Hinf I*	WT-252, HT-252,182,70 MT-182,70
	T>C/-6	*rs2303428*	FP-5'-TGT GGG CAG GCT GTG GTT C-3' RP-5'-CTC CCA TAT TGG GGC CTG CA-3'	150	*Pst I*	WT- 150 MT- 130,20 HT- 150,130,20
Asn^{127}Ser	*c.380A > G*	*rs17217772*	ASN variant: FP-5'-TTA GGC TTC TCC TGG CAA -3' Ser variant: FP-5'-TTA GGC TTC TCC TGG CAG-3' Anti-sense primer: FP-5'-AGG AGA GCC TCA AGA TTG-3' Control primer: FP-5'-AAA ATT TTA AAG TAT GTT CAA G-3'	264	*No enzyme*	Control- 264bp ASN/SER variant-210bp

4.3.3 Genotyping of *MSH3* polymorphism by PCR-RFLP

Genotype analysis was complete by PCR-RFLP technique utilizing the primers, restriction enzymes, and annealing temperature as previously reported (Yadav *et al.*, 2018), tabulated in **Table 4.3**. For amplification of polymorphic sites, a standard PCR was used. The PCR reaction was carried out in a 25 µl mixture composed of 1X PCR buffer, 0.5 µM of forward and reverse primer, 100 g/ml bovine serum albumin (BSA), 1.5 mM $MgCl_2$, 200 µM dNTPs, and 1U Taq polymerase (DNAzyme, Thermo Scientific) and 200 ng DNA. The PCR was performed at the following temperature settings: melting step at 95°C for 5 min; denaturation at 94°C for 30sec (29 cycles); annealing step at 60°C for 45sec; primer extension at 72°C for 30sec and final extension at 72°C for 5 min. After amplification, the PCR product of 225bp was incubated with five units of restriction enzyme *Hha I* (Takara Bio USA Inc.) for 10-12 hours at 37°C. The digested PCR products were run on 2.5% agarose gel that was dissolved in 0.5X TBE with 0.5µg/ml ethidium bromide. The electrophoresis was run in 0.5X TBE at 50 volts/cm. The digested samples are mixed with 2µl of 6X loading buffer containing sucrose, bromophenol blue, and xylene cyanol. Two different persons visualized the results to remove any bias, and 20% of randomly selected samples were used to check reproducibility.

Table 4.3: List of primers and restriction enzymes used for *MSH3* genetic variant

SNP	SNP ID #	Primers	PCR product size (bp)	Restriction enzyme	Fragments identifying genotypes (bp)
G>A	*rs26279*	5'- TCT AACAGG CAA GTA GGA AC -3' – forward primer 5'- TAG CCA CAT TTA ATC CAT AAC -3' – reverse primer	225	*HhaI*	AA 225 AG 225, 138, 87 GG 138, 87

4.3.4 Genotyping of *MSH6* polymorphism by PCR-RFLP

Genotype analysis was complete by PCR-RFLP technique utilizing the primers, restriction enzymes, and annealing temperature as tabulated in **Table 4.4**. The PCR mixture of 25 µl consists of 1X PCR buffer, 1.5 mM $MgCl_2$, 0.5µM of reverse and forward primer, 200 µM dNTPs, 1U Taq polymerase (DNAzyme, Thermo Scientific), 100 µg/ml BSA and 200 ng DNA. PCR setting used for the mixture were: 95 ˚C for 5 min and 94 ˚C for 30s (denaturation), 65˚C (*557G>T*), 58˚C (276A>G) and 60˚C (C>T) for 45s (annealing) followed by 72˚C for 30s (extension) for 29 cycles and the final extension was oerformed for 5 min at 72˚C. The PCR product of 355 bp for *557G>T* variant, 434 bp for *276A>G* variant, and 286 bp for *C>T* variant after amplification were visualized on 1.5% agarose gel and then restriction digestion was completed by adding 5U of *MspI* and *NciI* restriction enzyme (Takara Bio Inc.) respectively at 37 ˚C. The digested PCR samples were run on a 2.5% agarose gel dissolved in 0.5X TBE with 0.5µg/ml ethidium bromide. The electrophoresis was run in 0.5X TBE at 50 volts/cm. The digested samples are mixed with 2µl of 6X loading buffer containing sucrose, bromophenol blue, and xylene cyanol. Two different people examined the results to eliminate biasedness, and further 20% of the randomly selected samples were analyzed again, and the results' were 100% reproducible.

Table 4.4: List of primers and restriction enzymes used for *MSH6* genetic variant

Codon for protein	SNP	SNP ID #	Primers	PCR product size (bp)	Restriction enzyme	Fragments identifying genotypes (bp)
	557G>T	rs3136228	5'- GGC TCA GAT AAC GGA CTG TGG -3' – forward primer 5'- ACC CGA AAG GCC TCG GAA AG -3' – reverse primer	355	MspI	GG 264, 90 GT 355, 264, 90 TT 355
Pro92Pro	276A>G	rs1800932	5'- CCT GCC ATC AGC ATT ATA CCA -3' – forward primer 5'- CTG TAC ATG AAC ACG GAC TGA -3' – reverse primer	434	NciI	AA 313, 121 AG 434, 313, 121 GG 434
Gly39Glu	C>T	rs1042821	5'- TTA GGA GCT CCG TCC GAC AGA AC -3' – forward primer 5'- CCT CCG TTG AGG TTC TTC GCC TT -3' – reverse primer	286	NciI	CC 177, 109 CT 286, 177, 109 TT 286

4.4 Statistical analysis

4.4.1 Demographic analysis

Demographic characteristics like gender, age, smoking pattern, and pack-years were stratified into continuous and categorical variables. Chi-square test was employed to compare controls and cases for the categorical variables such as smoking pattern and gender, whereas t-test was employed to compare cases and controls for continuous variables like smoking and age.

4.4.2 Genotypic and allelic Frequency analysis

To calculate the genotype frequencies of all the SNPs, Hardy-Weinberg equilibrium theory (p^2+2pq+q^2=1; where p and q are the frequency of the wild and variant type allele respectively) was employed using the χ^2 test for both cases and control population. Minor allele frequency for all SNPs was also calculated. To check if there is any statistically noteworthy difference in allelic and genotypic frequencies between control and cases population, Pearson's χ^2 test was employed.

4.4.3 Association analysis

To investigate the risk of developing lung cancer and the polymorphism associated with it, odds ratio (ORs) and 95% confidence interval (CIs) were calculated. A logistic regression method was employed to determine the adjusted odds ratio (AORs), which considers all the confounding factors such as gender, age, smoking status. Apart from age which is a continuous variable, all the variables are dichotomous. In the logistic regression analysis, the advent of lung cancer was considered the dependent variable, whereas genotype was considered an independent variable that can predict susceptibility towards lung cancer. Apart from these comprehensive association studies, a subgroup study investigating the risk of developing lung cancer based on smoking and gender was also completed. Further, the association study related to susceptibility towards specific subgroups of lung cancer (based on histology) was also performed. All statistical analysis was performed with Medcalc version14.8.1 (Medcalc Software, Ostend, Belgium).

4.4.4 Haplotype analysis

To determine the haplotype frequency and linkage disequilibrium between groups SHEsis software was employed (http://202.120.31.177/myAnalysis.php) (Yong and Lin, 2005). The polymorphic sites that belong to the same gene and share a common chromosomal locus were investigated for LD to determine D and r^2 values. Haplotypes blocks that showed a frequency of more than 0.03 for both cases and controls were evaluated, and remaining haplotype blocks were not evaluated for their relation with lung cancer predisposition. A global Chi-square test was performed to determine the difference in frequencies, whereas the odds ratio for each haplotype was utilized to determine the association.

4.4.5 The false discovery rate analysis and Akaike information criterion

An online tool based on Benjamini and Hochberg methodology used the false discovery rate method to do multiple comparison corrections for various genotypic combinations (SNP-SNP) (http://sdmproject.com/ utilities/?show=FDR). The adjusted p values are calculated, and FDR adjusted $p<0.1$ was considered statistically significant (Benjamini and Hochberg, 1995). Further, the Akaike information criterion (AIC) was employed to evaluate the best fit model for each SNP. (AIC) is a mathematical technique for determining how well a model fits the data. AIC compares different possible models in statistics and evaluates which one is the best fit.

4.4.6 CART analysis

CART analysis was applied using the 8.0 Salford system, San Diego, California, to identify high order SNP-SNP interaction. It is a binary recursive-partitioning algorithm that generates a decision tree to recognize the high-risk subjects. The starting node in a decision tree is the most critical and is the most substantial risk factor. It then uses the statistical hypothesis testing method to find the subsequent split from the parent node, with each multiplicity adjusted p values for tree growth control ($P<0.05$). This process get repeated until the terminal nodes have no statistically significant splits or the terminal nodes reaches a pre-stated minimum size. The risk of all genotypes sets was calculated by declaring the node with a low case rate as the reference node to estimate the ORs and 95% CIs (Gupta *et al.*, 2016).

4.4.7 Overall survival and Hazard rate analysis

To determine if polymorphic genetic variants can predict the survival of lung cancer patients, median survival time (MST) and its corresponding curve were calculated using the Kaplan Meier method (Univariate regression analysis). To check the effect of confounding factors like Sex, Age, Cancer stage, histology, ECOG, and smoking status Cox method (multivariate regression analysis) was used. For Univariate analysis, log-rank p-value, and for multivariate analysis, the p-value has been considered. Cox proportional hazard models were used to determine hazard ratio (HRs) and 95% confidence intervals (CIs). The hazard ratio is the probability of a patient's death. A p-value of less than 0.05 was considered as a benchmark to decide if the analysis was significant or not. Medcalc statistical software package was used to perform all the analyses (MedCalc Software, Ostend, Belgium).

4.5 Computational analysis:

4.5.1 Retrieval of SNPs

All the SNPs related to the DNA mismatch repair genes and corresponding protein sequence in FASTA format were retrieved from NCBI's SNP and protein database. The complete flow chart is provided below (**Figure 4.1**):

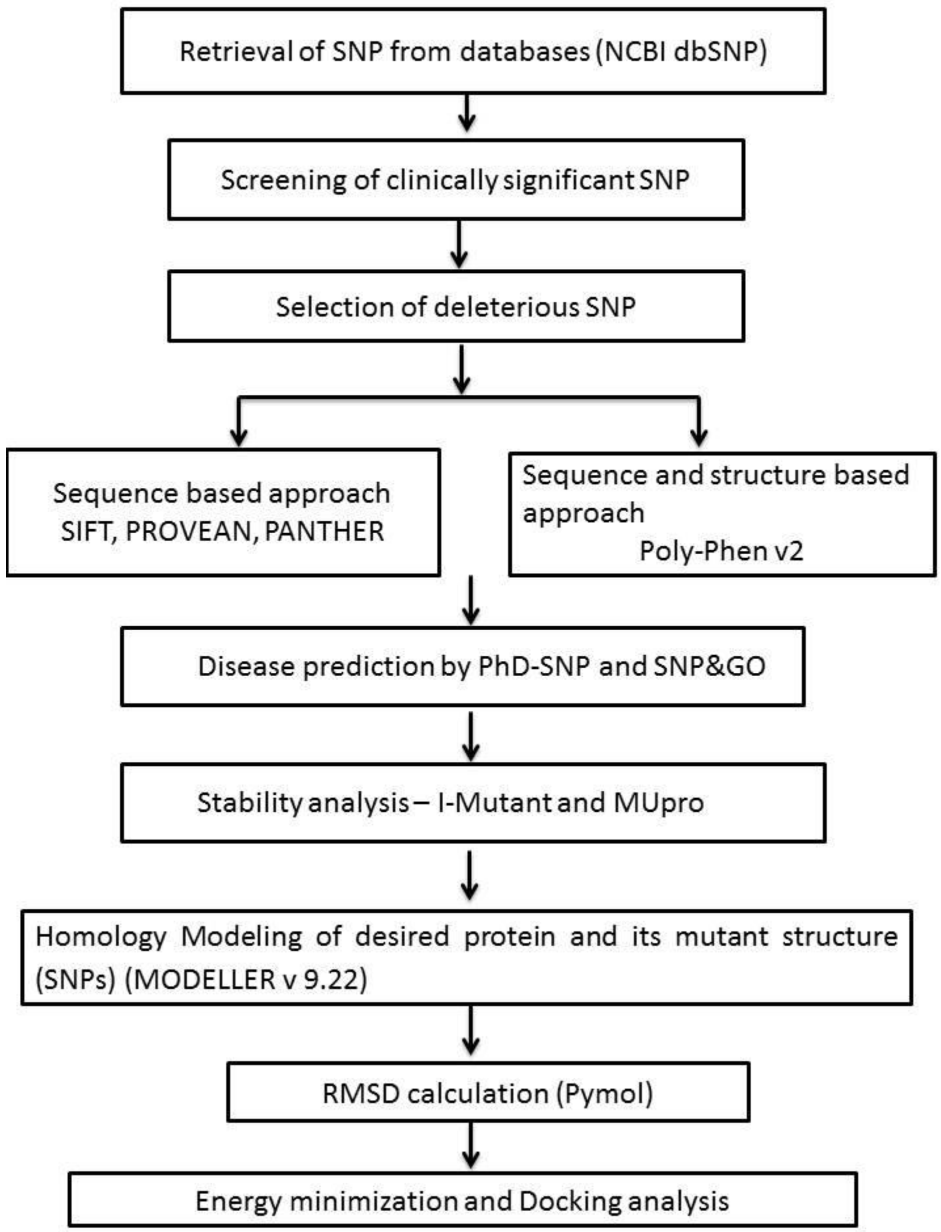

Figure 4.1: Flowchart for computational analysis of SNPs.

4.5.2 Prediction tools for identification of deleterious nsSNP's

SIFT (Ng and Henikoff, 2003), PROVEAN (Choi and Chan, 2015), and PANTHER (Mi *et al.*, 2021), which utilize sequence-based algorithms, was used to screen the deleterious nsSNP's. SIFT (Sorting Intolerant From Tolerant) tool uses sequence homology, physiochemical properties of amino acids & degree of evolutionary conservation to sort intolerant and tolerant amino acid substitutions and uses it to predict whether a particular nsSNP is deleterious or not (Mah *et al.*, 2011). In SIFT, SNPs having score > 0.05 (0.00–0.05) and < 0.05 were classified as intolerant and tolerant respectively. The sensitivity of SIFT tool is 73% (Frousios *et al.*, 2013). PROVEAN (Protein Variation Effect Analyzer) analyzes how the functionality of a protein will change due to a particular amino acid substitution by taking into account the common characteristics of amino acid and protein structure. The sensitivity of the PROVEAN tool is 79.4% (Choi *et al.*, 2012). It can also predict the non-synonymous mutations that can affect protein structure and, hence, alter its functionality. PANTHER (Protein Analysis Through Evolutionary Relationships) tool measures the time of a provided amino acid which has been evolutionarily preserved among different species and tells if there is any functional and structural effect on the protein due to that particular amino acid substitution. It uses statistical modeling based on HMM (Hidden Markov Model) and multiple sequence alignment (MSA) to calculate the evolutionary relationship of coding SNPs. It calculates the likelihood of a particular nsSNP altering the protein-based function on the substitution position-specific evolutionary conservation score (subPSEC) (Thomas *et al.*, 2003). The sensitivity of the PANTHER tool is 69% (Frousios *et al.*, 2013)

Polyphen (Polymorphism Phenotyping v2) considers Sequence and Structure-based approaches in predicting whether a particular amino acid substitution alters the functionality and structure of a human protein. It predicts the results based on various features, including sequence, phylogenetic and structural alterations where various algorithms such as THMM (Transformed Hidden Markov Model) and signalP are applied. Polyphen2 performs sequence-based characterization of the substitution site, maps the SNPs to known 3D protein structures, extracts protein structural & sequence attributes, then predicts whether a missense mutation is probably damaging, damaging, and benign (Mah *et al.*, 2011). The sensitivity of Polyphen 2 is 84% (Frousios *et al.*, 2013).

4.5.3 Prediction tools for identification of disease-associated nsSNP's

PhD-SNP (Capriotti and Fariselli, 2017) and SNP&GO (Calabrese *et al.,* 2009) tools were used to identify the disease-associated SNPs. PhD-SNP (Predictor of human Deleterious Single Nucleotide Polymorphisms) is a support vector machine (SVM) and sequence's evolutionary information-based classifier (Capriotti *et al.,* 2006); which uses three different algorithms viz 'sequence & profile based,' 'sequence-based,' and 'hybrid method.' The sequence-based method utilizes the local sequence environment of the mutation given to predict the outcome. The hybrid method utilizes an SVM sequence or SVM profile to predict the sequence profile is available for the sequence of interest. In our study, we have used a hybrid method. SNP&GO server utilizes gene ontology (GO) database to derive information based on which it predicts whether a particular mutation can be associated with a disease or not. The main characteristic feature of this tool over the other computational/bioinformatics techniques is available GO terms. It also incorporates the PANTHER classification system and its data, making this tool one of the most comprehensive tools for SNP analysis (Calabrese *et al.,* 2009). The sensitivity of PhD-SNP and SNP&GO servers is 62% and 71%, respectively (Frousios *et al.,* 2013).

4.5.4 Prediction tool for stability analysis of nsSNP's

I-Mutant (Capriotti *et al.,* 2005) and MUpro (Cheng *et al.*, 2006) were used to analyze nsSNP's effect on protein stability. I-Mutant is an SVM-based computational tool that uses either protein structure or sequence to predict the changes in protein stability upon single amino acid substitution. ProTherm, a comprehensive database for practical information regarding protein mutations, has trained this tool (Gromiha *et al.,* 1999). Approximately 80% of the query where the 3D structure and 77% in the query where only protein sequence is known can estimate if a mutation will stabilize or destabilize the protein. MUpro is a machine learning program based on SVM and Neural networks to predict changes in protein stability by single amino acid substitution (Cheng *et al.,* 2006). I-Mutant and MUpro are very reliable tools with 70.2% and 77.5% accuracy (Frousios *et al.,* 2013).

4.5.5 Modeling the native and mutant MSH2 protein

MODELLER v9.22 modeled the complete structure of DNA mismatch repair proteins. It is software that utilizes comparative homology modeling techniques for building the structure of proteins. Protein structure generation *via* comparative homology modeling includes the following steps: selection of template having maximum coverage and percentage identity using BLAST, aligning the selected template structure with the query sequence, building the model. The Ramachandran plot further validated generated structures. The selected model was saved as WILD. The mutant model was generated by MODELLER v9.22 using energy minimized WILD as a template structure. To mirror the protein's in vivo folding conditions and parameters, energy minimization of wild and all the mutant structures was performed with the help of the Chimera tool that utilizes the steepest gradient and conjugate gradient energy minimization algorithms. Further, RMSD values of mutant and native protein structures were calculated by superimposing native and mutant protein structures in PyMol software and utilizing its "align" functionality. RMSD value indicates structural deviation from native protein; the higher the RMSD value, greater is the deviation from its native form.

4.5.6 Molecular docking analysis

Docking was performed for the mutant, wild model of desired protein and the protein molecule with which the wild molecule has close/critical association. HADDOCK server was utilized for protein-protein docking keeping all the parameters to their default values. HADDOCK server provided HADDOCK score, RMSD, Desolvation energy, Electrostatic energy, Restraints violation energy, Van der Waals energy, Z-Score and Buried Surface Area.

Chapter 5

Results

5.1 Genomic DNA isolation

Both qualitative and quantitative estimation of genomic DNA was carried out. Qualitative estimation was done using agarose gel electrophoresis (agarose concentration 0.8%), whereas quantitative estimation was completed using “nanodrop” to determine its concentration. The bands of the isolated DNA were closely checked for any shearing loss, as represented in **Figure 5.1.** The exact process was repeated for control subjects.

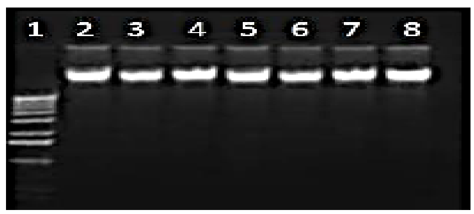

Figure 5.1: 0.8% agarose gel representing genomic DNA isolated from blood; Lane 1: 1kb ladder, Lane 2-4: genomic DNA isolated from lung cancer patient’s blood, Lane 5-8: genomic DNA isolated from healthy controls.

5.2 PCR amplification

For all the SNPs, PCR amplification was confirmed by running the PCR product on 2.0% agarose gel along with a DNA marker ladder of either 50 bp or 100 bp size to verify the correct product formation. A typical example of agarose gel is depicted in

Figure 5.2 demonstrating the amplification of *MSH2 118T>C* (*rs2303425*). The amplified product of 190 bp is obtained, confirmed by a 100 bp DNA marker in the first well.

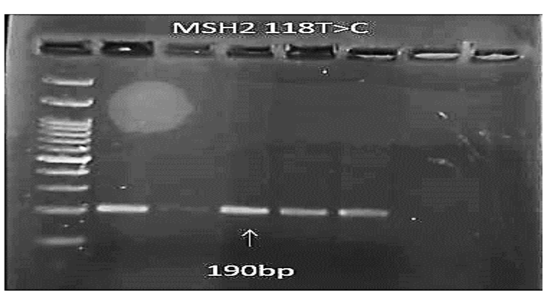

Figure 5.2: Representative example of agarose gel electrophoresis showing PCR amplification for *MSH2 118T>C*.

5.3 Genotyping with RFLP

5.3.1 Genotyping of *MLH1* genetic variants

The Native-PAGE gel images showing the restriction pattern for *MLH1* polymorphism are represented in **Figure 5.3**. The *rs1800734 (-93G>A)* polymorphism is a 5' UTR variant and is located in the core promoter of *MLH1*, 93 bases upstream of the transcription start site in a region that is required for maximal transcriptional activity (Ito *et al.*, 1999, Arita *et al.*, 2003). The size of the amplified PCR product was 387 bp. For genotyping of *MLH1* gene, the restriction enzyme used was *Pvu II*, a single band of 387 bp corresponds to mutant type genotype (*AA*), two bands of size 207 and 180bp belong to wild type genotype (*GG*), and three bands of length 387, 207 & 180 bp belongs to heterozygous type genotype (*GA*) as shown in **Figure 5.3**

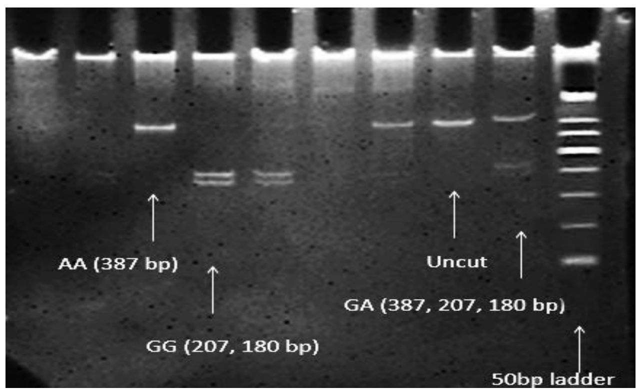

Figure 5.3: Representative example of 10% Native -PAGE showing PCR-RFLP pattern for *MLH1* (*rs1800734*/ *-93G>A*) polymoprhism. Band size: *AA*: 387 bp, *GG*: 207, 187 bp and *GA*: 387, 207, 187 bp. Lane 1: 50bp ladder; Lane 2,4: GA; lane 3: Uncut; lane 6,7: GG and lane 8: AA.

5.3.2 Genotyping of *MSH3* genetic variants

The Native-PAGE gel images showing the restriction pattern for *MSH3* polymorphism are represented in **Figure 5.4:** Representative example of 10% Native -PAGE showing PCR-RFLP pattern for *MSH3* (*rs26279* G>A) polymorphism. Band size: *GG*: 138, 87 bp, *AA*: 225 bp, and *AG*: 225, 138, and 87 bp.. The size of the amplified PCR product was 225 bp. The polymorphic variants for the *MSH3* gene were analyzed after digesting the PCR product with *HhaI* (Takara) at 37˚C overnight and running the digested product in 10% Native-PAGE. The wild allele produced a single band of 225bp, mutant allele produced two bands of 138 & 87 bp, and heterozygous allele produced three bands of 225, 138 & 87 bp, as shown in **Figure 5.4**

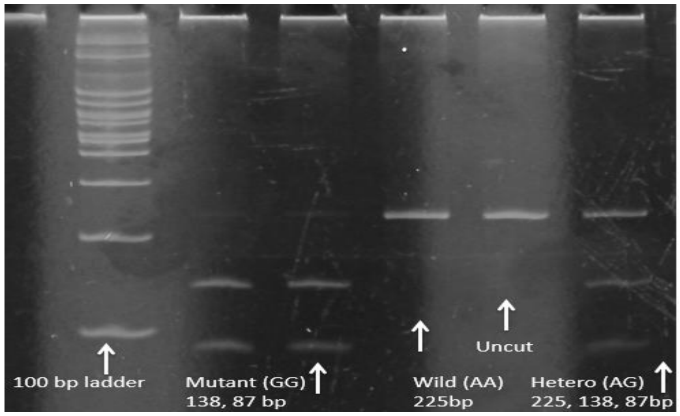

Figure 5.4: Representative example of 10% Native -PAGE showing PCR-RFLP pattern for *MSH3* (*rs26279* G>A) polymorphism. Band size: *GG*: 138, 87 bp, *AA*: 225 bp, and *AG*: 225, 138, and 87 bp.

5.3.3 Genotyping of *MSH6* genetic variants

In this study, we have analyzed three variants alleles of *MSH6* (*viz.* (*557G>T, rs3136228*), (*A>G, rs1800932*) and (*C>T, rs1042821*) polymorphism. These mutations were distinguished by native –PAGE, banding pattern generated after digesting the PCR product of 355bp, 434bp, and 286 bp respectively by *MspI* for *rs3136228* and *NciI* for *rs1800932* and *rs1042821* polymorphisms at 37 ˚C. For *rs3136228* polymorphism, the wild type produced two bands of 264 and 90bp, the mutant type produced a single band of 355bp, and the heterozygous type produced three bands of 355, 264, and 90bp, as shown in . For *rs1800932* polymorphism, the wild type produced two bands of 313 and 121bp, and the heterozygous type produced three bands *viz.* 434, 313, and 121bp, whereas the mutant type produced a single band of 434bp as depicted in **Figure 5.5**.

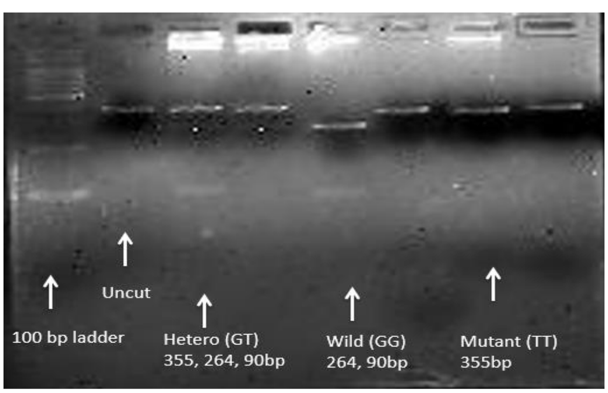

Figure 5.6: Representative example of 10% Native -PAGE showing PCR-RFLP pattern for *MSH6* (557G>T, rs3136228) polymoprhism. Band size: *GG*: 264, 90bp; *GT*: 355, 264, 90bp; *TT*:355bp. Lane 1: 100bp ladder, lane 2: Uncut; lane 3: GT; lane 4,6,7,8: TT and lane 5: GG

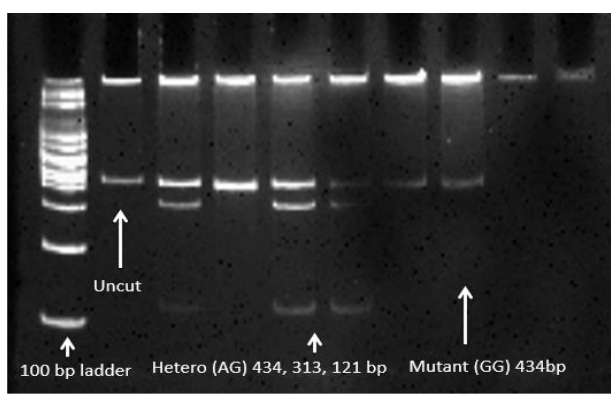

Figure 5.5: Representative example of 10% Native -PAGE showing PCR-RFLP pattern for MSH6 (A>G, rs1800932) polymorphism, band size: GG: 434bp; AG: 434, 313 and 121bp. Lane 1: 100bp ladder; labe 2: Uncut; lane 3,5,6: AG; lane 4,7,8: GG

For *rs1042821* polymorphism, the mutant type produced a single band of 286bp, and the heterozygous type produced three bands of 286, 177, and 109 bp, whereas the wild type produced two bands of 177 and 109bp, as shown in **Figure 5.7**

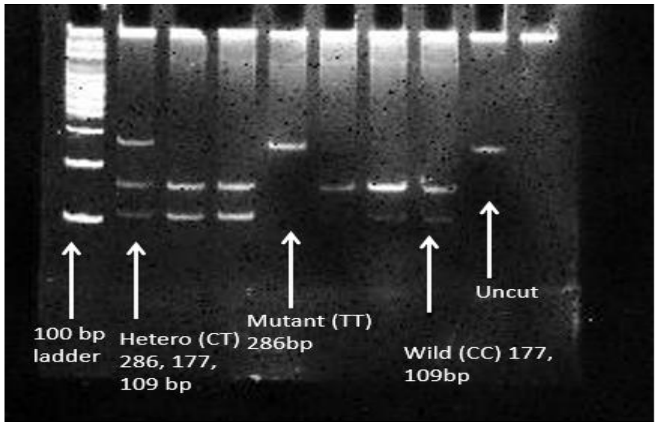

Figure 5.7: Representative example of 10% Native -PAGE showing PCR-RFLP pattern for MSH6 (C>T, rs1042821) polymorphism, band size: CC: 177, 109bp, CT: 286, 177 and 109bp, TT: 286bp. Lane 1: 100 bp ladder; lane 2: CT; lane 3,4,6,7,8: CC; lane 5: TT and lane 9: Uncut.

5.3.4 Genotyping of *MSH2* genetic variants

In this study we have investigated six polymorphism of *MSH2* gene (*2063T>G, rs63749993*;*118T>C, rs2303425; IVS1+9G>C, rs2303426; 1032G>A , rs4987188; T>C/-6, rs2303428; Asn^{127}Ser, rs17217772*). The PCR product of *2063T>G*, *118T>C*, *IVS1+9G>C*, *1032G>A*, and *T>C/-6*, and *380A>G* were 183, 190, 159, 252, 150, and 264 bp was checked on 1.5% agarose gel and then digested with five units of *Hsp92 II*, *BseN I*, *Bgl I*, *Hinf I*, *Pst I* restriction enzymes (Takara Bio USA Inc.) respectively at 37 ˚C. The digested products were visualized on 10% Native-PAGE for *2063T>G*, *118T>C*, *IVS1+9G>C*, and *T>C/-6* variants, whereas for *1032G>A* variants, the digested samples were run on 2.5% agarose gel. MSH2 *Asn^{127}ser* polymorphism was carried out using allele-specific PCR; hence the "Asn" and "Ser" variant band size was visualized on 2.5% agarose gel. *For* MSH2 (*2063T>G*, rs63749993), the wild type produced two bands of 158, 25bp, the mutant type produced a single band of 183bp, and the heterozygous type produced three bands of 183, 158, and 25bp (**Figure 5.8a**). For *118T>C, rs2303425* polymorphism, the wild type produced a single band of 190bp, the heterozygous type produced three bands of 190, 118, and 72bp, whereas the mutant type produced two bands of 118, 72bp (**Figure 5.8b**).

a) MSH2 (*2063T>G*, rs63749993) *b)* MSH2 (118T>C, rs2303425)

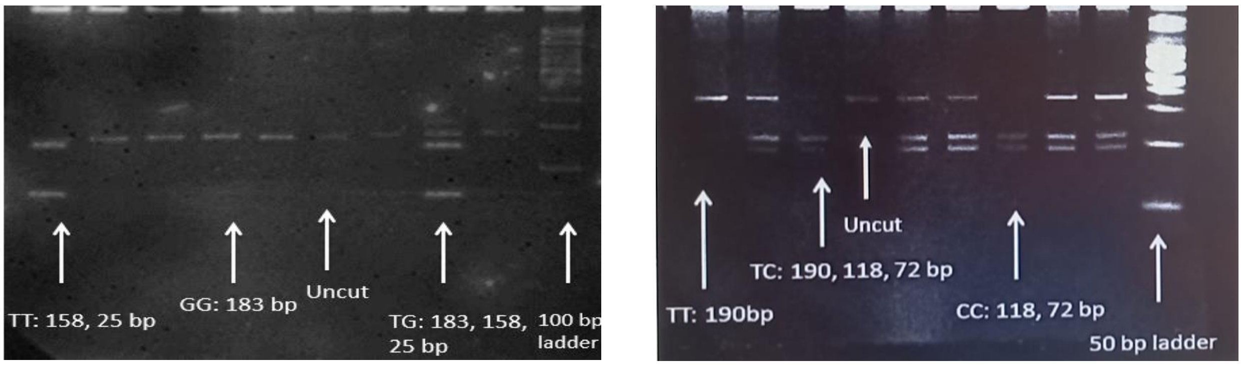

Figure 5.8: a) Representative example of 10% Native -PAGE showing PCR-RFLP pattern for *MSH2* (*2063T>G, rs63749993*) polymoprhism. Band size: *TT*: 158, 25bp, *GG*: 183bp and *TG*: 183, 158 and 25bp. b) MSH2 (*118T>C*, *rs2303425*) polymorphism, band size: *TT*:190bp, *TC*: 190, 118 and 72bp, *CC*: 118, 72bp.

For *IVS 1+9G>C*, *rs2303426* polymorphism, the wild type produced a single band of 159 bp, the mutant type produced two bands of 113 and 47bp, the heterozygous type produced three bands of 159, 113, and 47bp (**Figure 5.9a**). For *1032G>A*, *rs4987188* polymorphism, the mutant type produced two bands of 182 and 72bp, the heterozygous type produced three bands of 252, 182, and 70bp, the wild type produced a single band of 252bp (**Figure 5.9b**).

a) MSH2 (*IVS1+9G>C, rs2303426*) b) MSH2 (*1032G>A , rs4987188*)

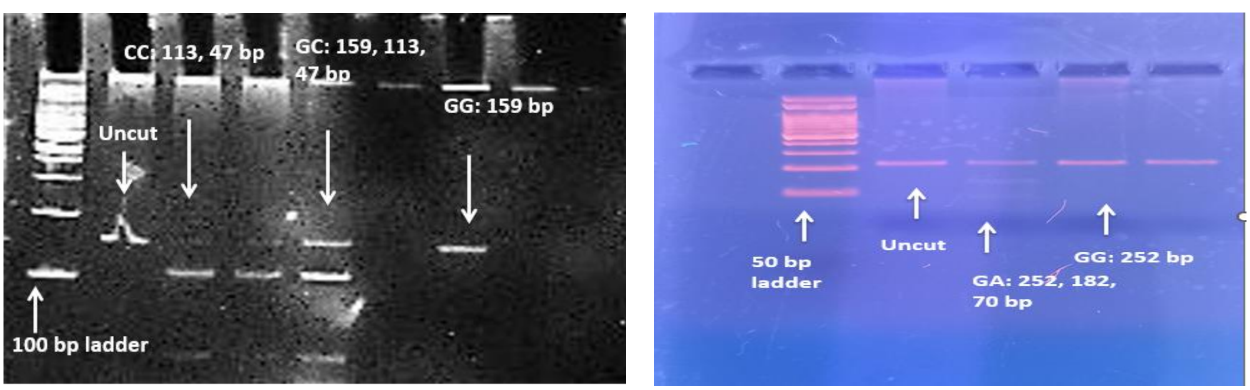

Figure 5.9: a) Representative example of 10% Native -PAGE showing PCR-RFLP pattern for MSH2 (*IVS1+9G>C, rs2303426*) polymorphism, band size: *GG*: 159, *CC*: 113, 47bp, *GC*: 159, 113 and 47bp. b) MSH2 (*1032G>A , rs4987188*) polymorphism, band size:GG:252bp, *GA*: 252, 182, 70bp, *AA*: 182, 72bp (no mutant was observed).

For *T>C/-6, rs2303428* polymorphism, the wild type produced a single band of 150bpm heterozygous type produced three bands of 150, 130, and 20bp, the mutant type produced two bands of 130 and 20bp (**Figure 5.10a**). For *Asn*^{127}*Ser*, rs17217772 polymorphism, a control band was observed of 264bp, whereas for the “Asn” and “Ser” variant, a band size of 210bp was observed (**Figure 5.10b**).

a) MSH2 (*T>C/-6*, *rs2303428*) b) MSH2 (*Asn*^{127}*Ser*, *rs17217772*)

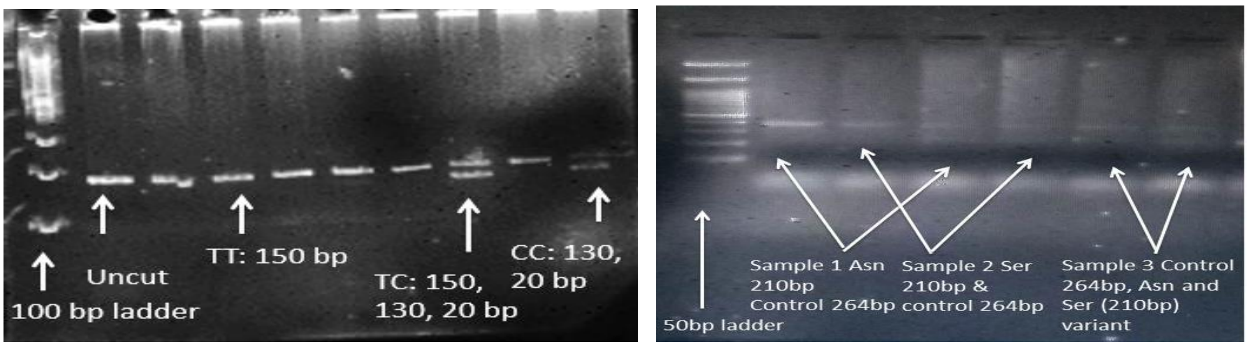

Figure 5.10: a) Representative example of 10% Native -PAGE showing PCR-RFLP pattern for MSH2 (*T>C/-6, rs2303428*) polymorphism, band size: *TT*: 150bp, *TC*: 150, 130, 20bp and *CC*: 130, 20bp. b) MSH2 (*Asn*^{127}*Ser*, *rs17217772*) polymorphism, band size: Control: 264bp, Asn and Ser variant: 210bp.

5.4 Epidemiological statistics of North Indian population under study

5.4.1 Regional Distribution of Lung cancer patients

Blood samples of lung cancer patients registered at the Department of Pulmonary Medicine, PGIMER Chandigarh was collected. PGIMER is a strategically located specialized government hospital to cater to the needs of nearby states in Northern India. Patients requiring specialized treatment are referred to this hospital by local hospitals of the surrounding states. This hospital caters to nearby states, including Punjab, Haryana, Himachal Pradesh, Uttar Pradesh, Uttarakhand, Jammu & Kashmir, and Chandigarh. For this study, we have recruited 500 lung cancer patients, out of which the maximum patients are from Punjab: (142, 28%), Himachal Pradesh: (154, 31%), and Haryana: (92, 18%), as represented in **Figure 5.11**. The remaining patients belonged to Uttar Pradesh: (55,11%), Jammu and Kashmir: (19, 4%), and Uttarakhand:(11, 2%).

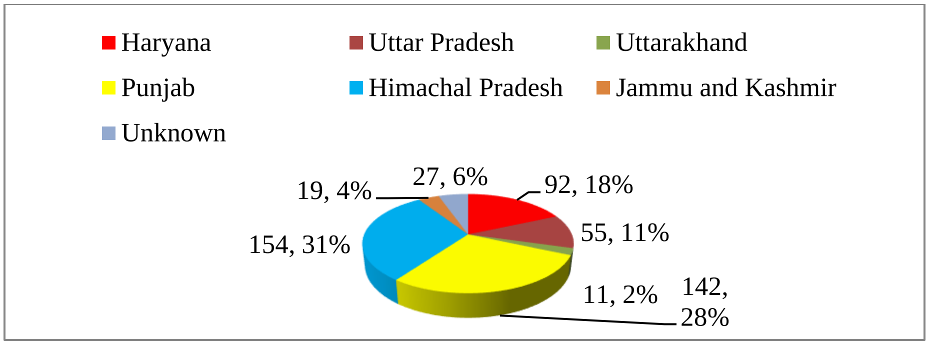

Figure 5.11: Pie chart illustrating the regional distribution of North Indian lung cancer patients.

The recruited patients for this study were analyzed for their regional distribution based on gender. Our study comprises 401 (80.2%) males and 99 (19.8%) females in cases as compared to 455 (91.0%) males and 45 (9.0%) females in controls. The maximum number of males and females belonged to Himachal Pradesh (128, 32%) and Punjab (26, 27%) (**Figure 5.12**). The lowest number of females were in Uttarakhand and Jammu & Kashmir (3, 3%) (**Figure 5.13**).

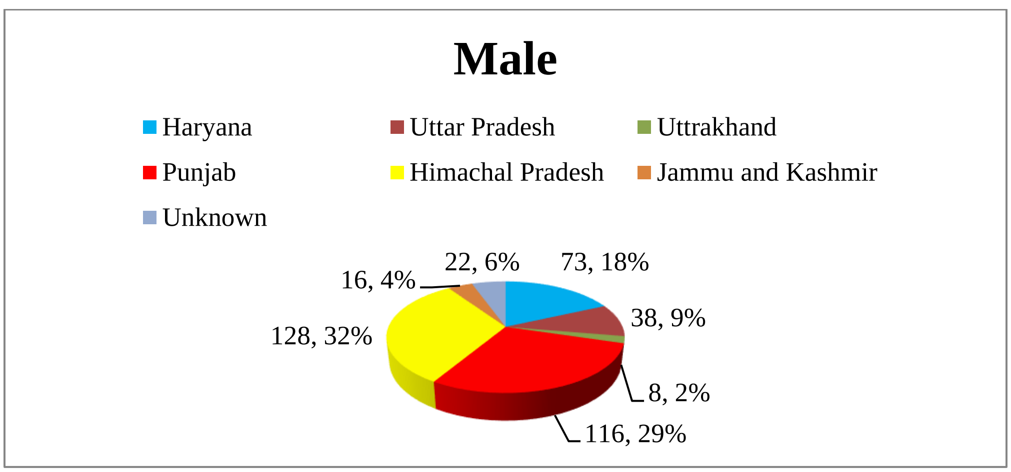

Figure 5.12: Pie chart illustrating the regional distribution of males' North Indian Lung cancer patients.

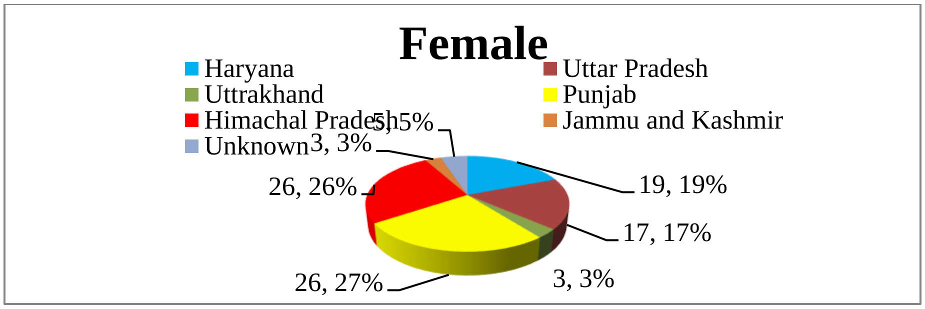

Figure 5.13: Pie chart illustrating the regional distribution of females' North Indian Lung cancer patients.

5.4.2 Distribution of lung cancer patients based on smoking

Smoking is considered one of the root causes of lung cancer development, so smoking habits were also collected in this study. The population was stratified into non-smokers and smokers based on the smoking pattern. Since smoking is a well-documented reason for cancer, smokers were also significantly more than non-smokers in cases. To avoid sample bias, we have tried to mimic the smoking pattern in the control population. The number of smokers in cases was 398, 79.6% compared to 449, 89.8% in controls, as depicted in **Figure 5.14**. Pack years for each subject were calculated to estimate smoking quantitatively. The mean pack-year of cases was 20.82 ± 21.03, higher than 13.84 ± 11.27 for controls. This indicates that the prevalence of smokers in cases was higher than the control population.

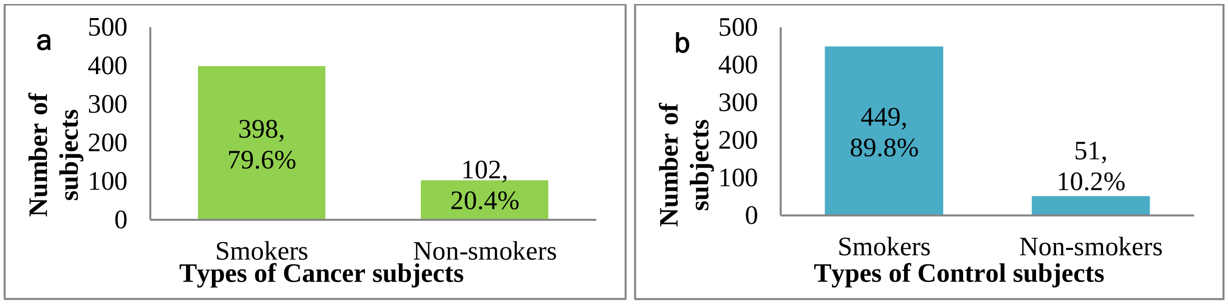

Figure 5.14: Graph showing the distribution of study subjects based on smoking status. a: Among cases, b: Among controls

Further, data about the type of smoking was also recorded. This data shed light on the various types of smoking carried out by the population under study, including cigarette, beedi, combined consumption of cigarette and tobacco. In our study, beedi was the most commonly used source of smoking in both cases and controls. The smoking habits of lung cancer and control subjects were also evaluated based on gender. The number of smokers in male lung cancer subjects (366, 91.3%) is lower than the male smokers in the control population (441, 96.9%), as shown in **Figure 5.15**.

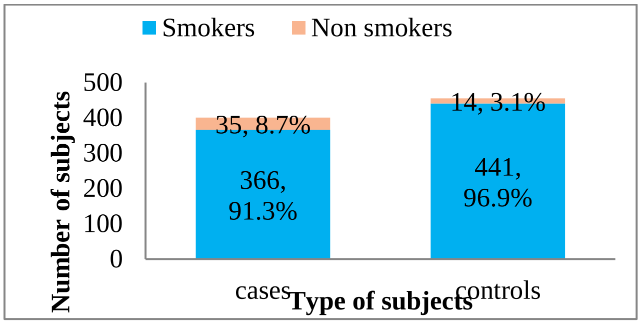

Figure 5.15: Graph showing the distribution of male study subjects based on smoking status.

5.4.3 Distribution of lung cancer patients based on age

Age is another fundamental demographic characteristic for any case-control study; henceforth, it is associated with the onset of many cancers. The age for all the recruited cases and control population was recorded. The mean age of lung cancer cases in our study was 60.5 ± 9.86 years, with a range of 30-86 years. We have also estimated the median age for cases as 62 years. The mean age for the control population in our study was 61.6 ± 11.4 years, with a range of 31-83 years. The median age for the control population was 61 years. The diagrammatic representation of the distribution of lung cancer subjects based on gender and age is depicted in **Figure 5.16**. Male subjects recruited for this study having age ≥ 62 years are slightly more in number than the males having age < 62 years. On the contrary, females aged ≥62 years are slightly less than females aged < 62 years.

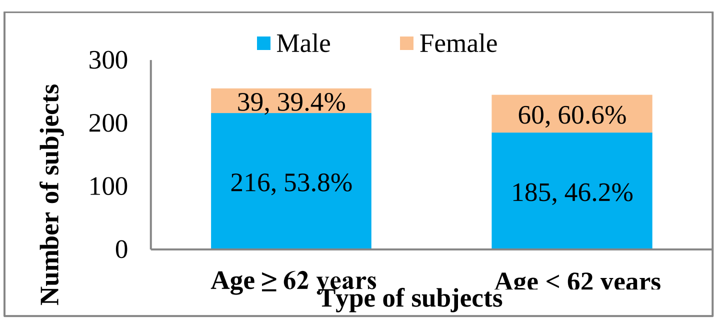

Figure 5.16: Graph showing the age-wise distribution of lung cancer subjects based on gender

5.4.4 Distribution of lung cancer patients based on the stage of tumor and TNM

All the clinical data were analyzed after reviewing the medical records of all the recruited patients from the Department of Pulmonary Medicine, PGIMER Chandigarh. The study population was stratified based on clinical stages I, II, III, and IV. The patients with stage I and II tumors were significantly less than patients with stage III and IV tumors (5% vs. 90%). Hence most of the lung cancer patients reported advanced stage of the tumor. The number of stages I patients was 6, 1.2%, stage II was 19, 3.8%, stage III was 193, 38.6%, and stage IV was 257, 51.4%, as shown in **Figure 5.17**. Further, we have also stratified the study population based on other clinic-pathological features such as direct tumor extension (T), lymph node involvement (N), and metastasis (M). The detailed stratification of lung cancer patients based on tumor size is as follows: TX+T1 (45, 9.0%), T2 (41, 8.2%), T3 (85, 17.0%), and T4 (302, 60.4%). Based on lymph node involvement, the subjects under study were distributed as NX+N0 (68, 13.6%), N1 (34, 6.8%), N2 (210, 42.0%), and N3 (163, 32.6%). Similarly, the patients were stratified based on metastatic status into M0+MX (211, 42.2%) and M1 (263, 52.6%). TNM characteristics for the remaining 5.2% cases were not available.

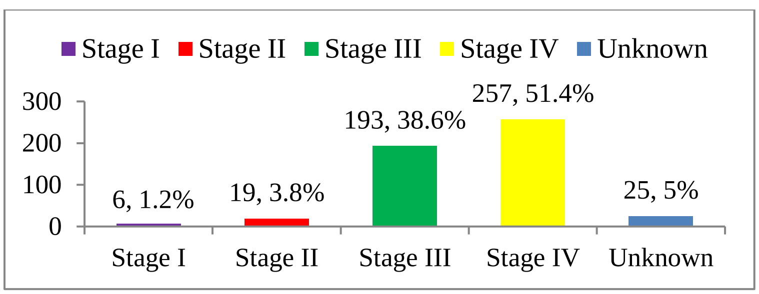

Figure 5.17: Graph illustrating the distribution of North Indian Lung cancer patients based on clinical-stage

5.4.5 Distribution of lung cancer patients based on histology of the tumor

Based on histology, lung cancer can be classified into ADCC, SQCC, and SCLC. All the patients recruited for this study were grouped into these three categories. In our study, most of the patients that were enrolled were suffering from SQCC (205, 41.0%) followed by ADCC (203, 40.6%) and SCLC (84, 16.8%) remaining 8, 1.6% of subjects involved in this study belonged to either large cell carcinoma, malignant mesothelioma, Oat cell carcinoma or are histo-pathologically undifferentiated as shown in the **Figure 5.18**. Hence, SQCC is the most commonly occurring lung cancer in the North Indian population.

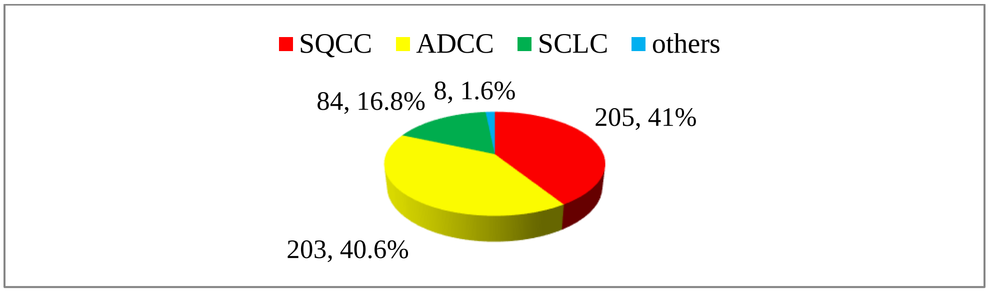

Figure 5.18: Pie chart illustrating the distribution of North Indian Lung cancer patients based on histological subtypes.

5.4.6 Distribution of lung cancer patients based on performance status

Lung cancer patients' two leading performance indicators are KPS-PS (Karnofsky performance status) and ECOG-PS (Eastern Cooperative Oncology Group). With the help of these parameters, the functional condition of the patients can be evaluated. In our study, the numbers of patients having a KPS score between 70 and 80 were 259, 54.4%, between 90 and 100 were 131, 27.5%, and below 60 were 86, 18.1%, indicating abysmal performance as shown in **Figure 5.19**. Similarly, the patients were also evaluated based on ECOG score, and it was observed that 214, 44.9% of patients had ECOG scores between 0 and 1, 188, 39.49% of patients had ECOG score of 2, and 74, 15.54% of patients had ECOG scores between 3 and 4 as represented in the **Figure 5.20**.

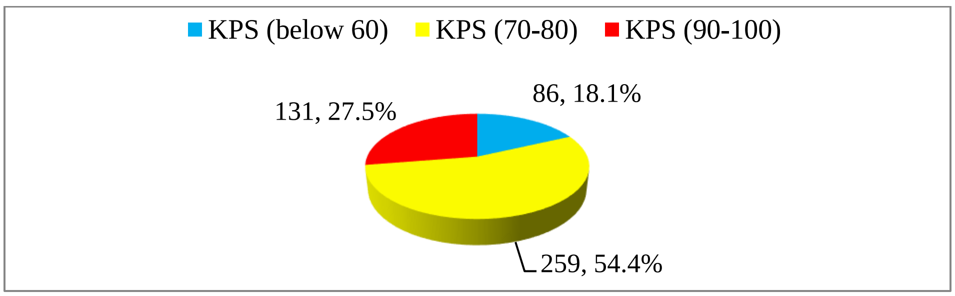

Figure 5.19: Pie chart illustrating the distribution of North Indian Lung cancer patients based on karnofsky performance status (KPS)

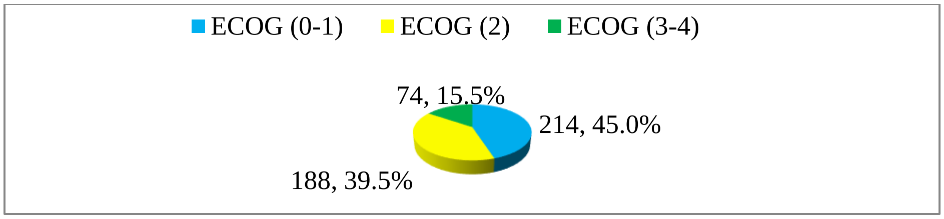

Figure 5.20: Pie chart illustrating the distribution of North Indian Lung cancer patients based on Eastern Cooperative Oncology Group -performance status (ECOG-PS).

5.4.7 Distribution of lung cancer patients based on platinum-based doublet chemotherapeutic regimen

The most common treatment for lung cancer patients is platinum-based doublet chemotherapy. Platinum drugs such as carbo/cisplatin in combination with other chemotherapeutic drugs are administered to lung cancer subjects in a cyclic pattern *via* intravenous mode. Commonly given drugs along with carbo/cisplatin are paclitaxel, gemcitabine, docetaxel, irinotecan, and pemetrexed. The distribution of lung cancer patients based on the chemotherapeutic regimen provided to them is as follows: Pemetrexed cis/carboplatin drug combination was given to 112, 31.63%, Paclitaxel cis/carboplatin combination was given to 82, 23.16%, Irinotecan cis/carboplatin drug combination was administered to 45, 12.71%, Docetaxel cis/carboplatin drug combination was given to 37, 10.45%, and Gemcitabine cis/carboplatin drug combination was administered to 29, 8.19% patients and Ceritinib was given to least number of patients 1, 0.28% as shown in the **Figure 5.21**. Due to the unavailability of data about chemotherapeutic regimens, 48, 13.55% of patients were put in the others category.

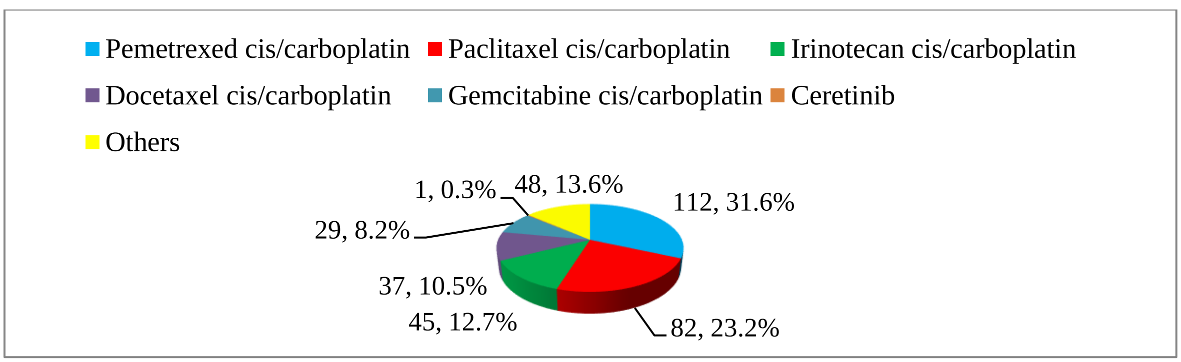

Figure 5.21: Pie chart illustrating the distribution of North Indian Lung cancer patients based on platinum-based doublet chemotherapeutic regimen

5.5 Statistical Analysis for polymorphic sites of *MLH1* gene

5.5.1 Demographic characteristics

The study group comprised 500 lung cancer subjects and 500 healthy controls. The main characteristics of the study group like age, sex, smoking habit, pack-years, histology, TNM staging, and other clinical data are represented in **Table 5.1.** The mean age of control and cases are 61.6 ± 11.4, 60.5 ± 9.86, respectively. The study includes 401 (80.2%) male subjects in cases and 455 (91.0%) males subjects in controls, whereas 99 (19.8%) and 45 (9.0%) were female subjects both in case and control groups, respectively.. In this study, a comparison between smokers and non-smokers was also made. The lung cancer cases population comprises 398 (79.6%) smokers compared to 449 (89.8%) smokers in the control population. The number of non-smokers in cases and control population was 102 (20.4%) and 51 (10.2%), respectively; a significant difference was observed in the population of smokers and non-smokers ($p<0.0001$). We have attempted to match the number of smokers and non-smokers sufficiently, as smoking is well documented to be noteworthly associated with the development of lung cancer, the same was recorded in this study population. Smoking can be further quantified in the form of pack-years [(cigarettes or beedis per day/20) X years smoked]. Hence we compared the means of pack years in controls and cases with a significant difference ($p<0.0001$). Out of 500 lung cancer subjects, 205 (41%) had squamous cell carcinoma (SQCC), 203 (40.6%), and 84 (16.8%) were diagnosed with adenocarcinoma (ADCC) and small cell lung carcinoma (SCLC), respectively. Among the cases, the TNM stage data is as follows: stage I: 6 (1.2%), stage II: 19 (3.8%), stage III: 193 (38.6%), and stage IV: 257 (51.4%). The cases with a relative tumor size of T3 and T4 were significantly more than T1 and T2 (387 vs. 67). In cases, the patient having lymph node category of N0 was 65 (13%) whereas patients having N1 were 34 (6.8%), N2 was 210 (42%), N3 163 (32.6%), and N4 category was 0%. 210 (42%) had no sign of metastasis (M0) compared to 263 (52.8%) cases with distant metastasis (M1). Further, KPS and ECOG as performance indicators were used to evaluate patients; in our study, 86 (18.1%) of cases have KPS below 60, 259 (54.4 %), and 131 (27.5%) cases reported a KPS in the range of 70-80 and 90-100, respectively. Based on ECOG, 214 (44.9%) cases have reported an ECOG value between 0-1, 188(39.49%) and 74 (15.54%) cases have ECOG values of 2 and between 3 & 4. As depicted in Table 5.1 out of 500 cases, 354 patients have undergone platinum-based doublet chemotherapy, out of which 112 (31.63%) have received pemetrexed + cisplatin/carboplatin, 45 (12.71%) cases were given irinotecan+cisplatin/carboplatin, 37 (10.45%) cases were administered docetaxel+cisplatin/carboplatin, 82 (23.16%) received paclitaxel+cisplatin/carboplatin and 29 (8.19%) cases were given gemcitabine in combination with cisplatin/carboplatin.

Table 5.1: Demographic characteristics among cases and controls

Variable	**Total (N)**	**Cases, n (%)**	**Controls, n (%)**	**[a]*p*-value**
Age (years)	500			
Mean ± SD		60.5 ± 9.86	61.6 ± 11.4	0.103
Range		30-86	31-83	
Gender	500			
Male		401 (80.2)	455 (91.0)	
Female		99 (19.8)	45 (9.0)	<0.0001
Smoking status	500			
Smokers		398 (79.6)	449 (89.8)	
Non-smokers		102 (20.4)	51 (10.2)	<0.0001
Pack years	500			
Mean ± S.D.		20.82 ± 21.03	13.84 ± 11.27	<0.0001
Histological types	500			
SQCC		205 (41.0)		
ADCC		203 (40.6)		
SCLC		84 (16.8)		
Others		6 (1.2)		
Unknown		2 (0.4)		
Overall survival	475			
Dead		356 (74.9)		
Alive		119 (25.0)		
Staging	500			
I		6 (1.2)		
II		19 (3.8)		
III		193 (38.6)		
IV		257 (51.4)		
Unclassified		25 (5.0)		
Tumor Size	500			
Tx		19 (3.8)		
T1		26 (5.2)		
T2		41 (8.2)		
T3		85 (17.0)		

T4		302 (60.4)	
Unknown		27 (5.4)	
Lymph Node	500		
Nx		3 (0.6)	
N0		65 (13.0)	
N1		34 (6.8)	
N2		210 (42.0)	
N3		163 (32.6)	
N4		0 (0)	
Unknown		25 (5.0)	
Metastasis	500		
Mx		1 (0.2)	
M0		210 (42.0)	
M1		263 (52.6)	
Unknown		26 (5.2)	
Performance status	476		
KPS (below 60)		86(18.1)	
KPS (70-80)		259(54.4)	<0.0001
KPS (90-100)		131(27.5)	<0.0001
ECOG (0-1)	476	214(44.9)	
ECOG (2)		188(39.49)	<0.0001
ECOG (3-4)		74(15.54)	<0.0001
Chemotherapy regimen	354		
Pemetrexed cis/carboplatin		112(31.63)	
Irinotecan cis/carboplatin		45(12.71)	
Docetaxel cis/carboplatin		37(10.45)	
Paclitaxel cis/carboplatin		82(23.16)	
Gemcitabine cis/carboplatin		29(8.19)	
Ceritinib		1(0.28)	
Others		48(13.55)	

Abbreviations: S.D. = Standard Deviation, n = total number of lung cancer cases or control subjects, a) *p*-values were derived from Pearson Chi-square test except for age and pack-years; Student t-test was used for age and pack-years. All *p*-values are two-sided. $p < 0.05$ was considered statistically significant

5.5.2 Genotypic and allelic distribution of the *MLH1* gene and risk of lung cancer based on histological subtypes

To separate the population into high-risk subgroups, the patients were stratified into subgroups based on their histological subtypes. Genotypic frequency of *MLH1* variants was compared to determine the prevalence of the high-risk subgroups from all the categories, as shown in **Table**

5.2. We reported a reduction in risk for heterozygous type genotype (*GA*) as compared to wild type genotype (*GG*) (AOR=0.48; 95%CI= 0.32-0.73, *P*=**0.0007**) in the co-dominant model for the patients diagnosed with ADCC. The dominant model, also exhibited reduced risk towards lung cancer for combined genotype (GA+AA) (AOR= 0.8; 95%CI= 0.67-1.15; *P*=**0.008**). When the recessive model was analyzed for ADCC histology, a 1.8-fold increased risk of developing lung cancer was observed (OR=1.81; 95%CI=1.06-3.05; p=**0.02**) in the mutant type genotype (*AA*). Similarly, when the odds ratio was calculated for patients suffering from SCLC, we reported a 1.78-fold increased risk of developing heterozygous type genotype (*GA*) in the co-dominant model (AOR= 1.78; 95%CI= 1.04-3.02; *P* =**0.03).** We did not find any association in dominant and recessive models for SCLC patients. (**Table 5.2**)

5.5.3 Effect of smoking status on the association of *MLH1* gene polymorphisms and lung cancer risk

Gene-environment interaction was analyzed after subdividing patients into smokers and non-smokers. These were further analyzed by stratifying based on pack-years (number of cigarettes/bidis smoked per day / 20) x number of years smoked). When the co-dominant model was evaluated, it was observed that heavy smokers have a 2-fold increased risk of lung cancer development for heterozygous type genotype (AOR=2.18, 95%CI= 1.33-3.55, *P* =**0.001**) (**Table 5.3**). In the dominant model for the same group, combined genotype (*GA+AA*) also showed a significant increase in risk for development of lung cancer (AOR=1.83, 95% CI= 1.18-2.85, *P* =**0.007**). Further, we did not find any association of occurrence of lung cancer with light smokers, as shown in **Table 5.3**.

Table 5.2: Genotypic and allelic distribution of the *MLH1* genetic variant and its association with risk of lung cancer according to tumor histology

ADCC				
rs1800734 **-93G>A (*MLH1*)**	**Controls (500) N(%)**	**Cases (203) N(%)**	**AOR (95% CI) [a]**	***p*** **[b]**
Co-dominant model				
GG	225 (45)	102 (50.2)	1.00 (Reference)	
GA	223 (44.6)	74 (36.4)	0.48 (0.32-0.73)	**0.0007**
AA	52 (10.4)	27 (13.4)	1.51 (0.79-2.86)	0.20
Dominant model				
GG	225 (45)	102 (50.2)	1.00 (Reference)	
GA+AA	275 (59.82)	101 (49.8)	0.88 (0.67-.1.15)	**0.008**
Recessive model				

GG+GA	448 (89.6)	176 (86.6)	1.00 (Reference)	
AA	52 (10.4)	27 (13.43)	1.81 (1.06-3.05)	**0.02**
		SCLC		
rs1800734 ***-93G>A*** **(*MLH1*)**	**Controls (500) N(%)**	**Cases (84) N(%)**	**AOR (95% CI)** [a]	***p***[b]
Co-dominant model				
GG	225 (45.0)	28 (39.37)	1.00 (Reference)	
GA	223 (44.6)	51 (46.46)	1.78 (1.04-3.02)	**0.03**
AA	52 (10.4)	5 (14.17)	0.57 (0.18-1.83	0.35

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided χ2 test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

Table 5.3: Relationship of different *MLH1* genotypes with the heavy and light smokers for cases and controls

		HEAVY SMOKERS		
Genotype ***rs1800734*** ***-93G>A*** **(*MLH1*)**	**Controls (126) N(%)**	**Cases (243) N(%)**	**AOR (95% CI)** [a]	***p***[b]
Co-dominant model				
GG	78 (61.9)	112 (46.09)	1.00 (Reference)	
GA	33 (26.19)	106 (43.62)	2.18 (1.33-3.55)	**0.001**
AA	15 (11.9)	25 (4.93)	1.09 (0.53-2.22)	0.81
Dominant model				
GG	78 (61.9)	112 (46.09)	1.00 (Reference)	
GA+AA	48 (38.09)	131 (53.90)	1.83 (1.18-2.85)	**0.0071**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided χ2 test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.5.4 Relation between *MLH1* Genotypic Distribution based on gender

The present study was evaluated further for gender and association with the polymorphic variants *MLH1*. **Table 5.4** show the genotype distribution of *MLH1* polymorphism according to gender and their relation with lung cancer. We employed the logistic regression method to analyze any association of gender, with *rs1800734* polymorphism, and susceptibility towards lung cancer. The genotypic frequencies were calculated for the *MLH1* polymorphic site for both sexes, *viz.* male and female, as shown in **Table 5.4**. Our results showed that in the co-dominant model of female category there was a significant decrease in risk of developing lung cancer in subjects harboring a variant allele (AOR= 0.016; 95%CI= 0.002-0.12; *P* **=0.0001**) and subjects having combined genotype (GA+AA) in dominant model (AOR= 0.02; 95%CI= 0.002-0.16; *P* =**0.0002**) (**Table 5.4**).

Table 5.4: Genotypic and allelic distribution of the *MLH1* genetic variant and its association with risk of lung cancer according to gender.

Genotype *rs1800734* *-93G>A* (*MLH1*)	**Controls (45) N(%)**	**FEMALE Cases (99) N(%)**	**AOR (95% CI)** [a]	***p*** [b]
Co-dominant model				
GG	1 (2.22)	51 51.51)	1.00 (Reference)	
GA	44 (97.77)	36 (36.36)	0.016 (0.002-.12)	**0.0001**
AA	0 (0)	12 (12.12)	3.5E(0.0-0)	0.99
Dominant model				
GG	1 (2.22)	51 (51.51)	1.00 (Reference)	
GA+AA	44 (97.77)	48 (48.48)	0.02 (0.002-0.16)	**0.0002**
Recessive model				
GG+GA	45 (100)	87 (87.87)	1.00 (Reference)	
AA	0 (0)	12 (12.12)	7.86E+07	0.99

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table

5.6 Statistical Analysis for polymorphic sites of *MSH3* gene

5.6.1 Genotypic and allelic frequencies of *MSH3* polymorphic variants among cases and controls and their role in modifying the susceptibility towards lung cancer

As depicted in **Table 5.5** regarding the polymorphic variant of *MSH3* $Ala^{1045}Thr$, in control subjects, 355 (71%) were homozygous for the wild genotype type (*AA*), 18 (3.6%) were homozygous variant carriers (*GG*), and 127 (25.4%) were carrying the heterozygous (*GA*) genotype. In contrast, in the lung cancer cases, 369 (73.8%) were homozygous for the wild genotype type (*AA*), 117 (23.4%) had the heterozygous (*GA*) genotype, and 14 (2.8%) were carrying both the mutant alleles (*GG*). For the *MSH3* polymorphism significant variation between genotypic frequencies was found between cases and controls ($\chi 2$=1.81; df=2; **p=0.03**). Further for *MSH3* polymorphic variants there was no deviation from Hardy-Weinberg equilibrium (HWE), *MSH3* {Cases: χ^2=1.58; df =1; p=0.20; Controls: χ^2=2.38; df=1; p=0.12}, Minor allele frequency (MAF) of *MSH3* is as follows*:* Cases=0.145, Controls=0.163.

To evaluate whether the polymorphic variants of *MSH3* genes had any relationship with lung cancer, we applied three genetic models to assess this relationship. Logistic regression analysis was applied, and the adjusted odds ratio and 95% CI were calculated. In the co-dominant model for *MSH3* $Ala^{1045}Thr$ polymorphism, no noteworthy relationship was observed with the polymorphic variants of *MSH3* (AOR=0.90; 95%CI= 0.67-1.21; p=0.52) and lung cancer susceptibility (**Table 5.5**). Similarly, the dominant model did not find any significant association between the *MSH3* $Ala^{1045}Thr$ polymorphic variant and lung cancer.

Table 5.5: Genotypic and allelic distribution of the *MSH3* genetic variant and its association with risk of lung cancer

	OVERALL			
rs26279 **G>A (*MSH3*)**	**Controls (500) N(%)**	**Cases (500) N(%)**	**AOR (95% CI)** [a]	p^b
Co-dominant model				
AA	355 (71.0)	369 (73.8)	1.00 (Reference)	
GA	127 (25.4)	117 (23.4)	0.90 (0.67-1.21)	0.52
GG	18 (3.6)	14 (2.8)	0.75 (0.36-1.56)	0.45
G(Allele)	837	855		
A(Allele)	163	145		
MAF	0.163	0.145		
	$\chi2$=1.81; df=2			**0.03**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.6.2 Relation between *MSH3* Genotypic Distribution based on gender

The present study was evaluated for gender and its association with the polymorphic variants of *MSH3*. **Table 5.6** show the genotype distribution according to gender and their relation to lung cancer. We employed logistic regression to analyze any association of gender with *MSH3* polymorphism and susceptibility towards lung cancer. The genotypic frequencies were calculated for the *MSH3* polymorphic site for both males and females, as depicted in **Table 5.6**. Univariate analysis was also employed to estimate the association of gender with MSH3 $Ala^{1045}Thr$ polymorphism and lung cancer occurrence. When the co-dominant model was applied, it was visible from the data that female lung cancer patients who were heterozygous carriers (*GA*) reported a 2.35-fold (95%CI= 0.85-6.52; **p=0.04**) risk towards the development of lung cancer; such a trend was also clearly visible in the dominant model, where female subjects exhibited a 2.4-fold (OR=2.39, 95%CI=0.90-6.23; **p=0.03**) increased risk towards the development of lung cancer. Although we reported a decreased risk of developing lung cancer in male patients in both co-dominant and dominant models, the results were not statistically significant.

Table 5.6: Relationship of different *MSH3* and polymorphisms based on gender

		FEMALE		
Genotype *rs26279* **G>A (*MSH3*)**	**Controls (45) N(%)**	**Cases (99) N(%)**	**AOR (95% CI)** [a]	***p*** [b]
Co-dominant model				
AA	37 (82.3)	71 (71.7)	1.00 (Reference)	
GA	7 (15.5)	24 (24.2)	2.35 (0.85-6.52)	**0.04**
GG	1 (2.2)	4 (4.1)	2.11 (0.19-22.0)	0.53
Dominant model				
AA	37 (82.3)	71 (71.7)	1.00 (Reference)	
GA+GG	8 (17.7)	28 (28.3)	2.39 (0.90-6.23)	**0.03**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided χ2 test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table

5.7 Statistical Analysis for polymorphic sites of *MSH6 (rs3136228)* gene

5.7.1 Genotypic and allelic frequencies of *MSH6 (rs3136228)* polymorphic variants among cases and controls and their role in modifying the susceptibility towards lung cancer

As depicted in **Table 5.7** regarding the polymorphic variant of *MSH6 557G>T*, in control subjects, 407 (81.4%) were homozygous for the mutant type genotype (*TT*), which is considered reference, 8 (1.6%) were homozygous wild carriers (*GG*), and 85 (17%) were carrying the heterozygous (*GT*) genotype. In contrast, in the lung cancer cases, 432 (86.4%) were homozygous for the mutant genotype (*TT*), 68 (13.6%) had the heterozygous (*GT*) genotype. Our study did not find any case subject harboring wild-type genotype (GG). It was observed that for *MSH6* polymorphism, there is no noteworthy distinction between the distribution of genotypic frequencies between cases and controls ($\chi2$=10.63; df =2; p=0.10). We have also found no deviation for Hardy-Weinberg equilibrium (HWE) for *MSH6* polymorphic variant {Cases: χ^2=2.66; df =1; p=0.10; Controls: χ^2=2.03; df=1; p=0.15}, Minor allele frequency (MAF) of *MSH6* is as follows*:* Cases=0.068, Controls=0.101.

We applied three genetic models to assess this relationship to evaluate whether the polymorphic variants of *MSH6 rs3136228* had any association with lung cancer. Logistic regression analysis was applied, and the adjusted odds ratio and 95% CI were calculated. In the co-dominant model for *MSH6 557G>T* polymorphism, no significant association was observed with the polymorphic variants of *MSH6* (AOR=0.75; 95%CI= 0.53-1.07; p=0.12) and lung cancer susceptibility (**Table 5.7**). Further, the dominant model predicted a decreased risk of developing lung cancer in the combined genotype (AOR=0.69; 95%CI=0.49-0.98; **p=0.03**); in contrast to the dominant model, the recessive model predicted a 1.5-fold higher risk of developing lung cancer in individuals carrying mutant type genotype (AOR=1.43; 95%CI=1.01-2.03; **p=0.03**).

Table 5.7: Genotypic and allelic distribution of the *MSH6* genetic variant and its association with risk of lung cancer

		OVERALL		
Genotype rs3136228 557 G>T	**Controls (500) N(%)**	**Cases (500) N(%)**	**AOR (95% CI)** [a]	p^b
Co-dominant model				
TT	407 (81.4)	432 (86.4)	1.00 (Reference)	
GT	85 (17.0)	68 (13.6)	0.75 (0.53-1.07)	0.12
GG	8 (1.6)	0 (0)	0	0.99
T (Allele)	899	932		
G (Allele)	101	68		
MAF	0.10	0.068		
	$\chi2$=10.63	p=0.10		
Dominant model				
TT	407 (81.4)	432 (86.4)	1.00 (Reference)	
GT+GG	93 (18.6)	68 (13.6)	0.69 (0.49-0.98)	**0.03**
Recessive model				
GT+GG	93 (18.6)	68 (13.6)	1.00 (Reference)	
TT	407 (81.4)	432 (86.4)	1.43 (1.01-2.03)	**0.03**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.7.2 Genotypic and allelic distribution of the *MSH6 (rs3136228)* gene and risk of lung cancer based on histological subtypes

In order to separate the recruited patients in to high-risk subgroups, the subjects were bifurcated into categories according to their histological subtypes. The genotypic frequency of *MSH6* polymorphism is depicted in **Table 5.8**. For *MSH6 557G>T* polymorphism, when we segregated the lung cancer subjects based on histological subtypes; the co-dominant and dominant model for the *MSH6 557G>T* polymorphism predicted a decreased risk of developing adenocarcinoma in subjects carrying the heterozygous genotype (*GT*) (AOR=0.6; 95%CI= 0.36-1.02; p=0.06) and combined (*GT+GG*) genotype (AOR=0.56; 95%CI= 0.34-0.95; p=**0.03**). In the recessive model, which predicted 1.75 fold risk for adenocarcinoma subjects carrying the *TT* genotype (AOR=1.75; 95% CI= 1.05-2.93; **p=0.03**).

Table 5.8: Genotypic and allelic distribution of the *MSH6* genetic variant and its association with risk of lung cancer according to tumor histology

		ADCC		
Genotype rs3136228 557 G>T (MSH6)	**Controls (500) N(%)**	**Cases (203) N(%)**	**AOR (95% CI)** [a]	***p*** [b]
Dominant model				
TT	407 (81.4)	177 (87.2)	1.00 (Reference)	
GT+GG	93 (18.6)	26 (12.8)	0.56 (0.34-0.95)	**0.03**
Recessive model				
GT+GG	93 (18.6)	26 (12.8)	1.00 (Reference)	
TT	407 (81.4)	177 (87.2)	1.75 (1.05-2.93)	**0.03**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.7.3 Effect of smoking status on the association of *MSH6 (rs3136228)* gene polymorphisms and lung cancer risk

In order to evaluate gene-environment interaction, subjects were bifurcated into smokers and non-smokers according to their smoking status. These were further analyzed by stratifying based on pack-years (number of cigarettes/bidis smoked per day / 20) x number of years smoked). We reported no smokers carrying wild-type genotype for *MSH6* polymorphism (**Table 5.9**). When we applied the co-dominant model between smoking and genotype, for non-smokers carrying the heterogeneous genotype (*GT*), a decreased propensity towards lung cancer was visible (AOR=0.31, 95%CI=0.12-0.78, p=**0.01**). Further, in the recessive model, the subjects carrying the mutant-type genotype (*TT*) recorded a 3-fold increase risk towards the development of lung cancer (AOR=3.22; 95%CI= 1.26-8.18; **p=0.01**) (**Table 5.9**).

Table 5.9: Relationship of different *MSH6* genotypes with the smoking status of cases and controls

		NON SMOKERS		
Genotype rs3136228 557 G>T (MSH6)	**Controls (51) N(%)**	**Cases (102) N (%)**	**AOR (95% CI)** [a]	***p*** [b]
Co-dominant model				
TT	36 (70.6)	91 (89.2)	1.00 (Reference)	
GT	15 (29.4)	11 (10.8)	0.31 (0.12-0.78)	**0.01**
GG	0 (0)	0 (0)	0	0.99
Dominant model				
TT	36 (70.6)	91 (89.2)	1.00 (Reference)	
GT+GG	15 (29.4)	11 (10.8)	0.31 (0.12-0.78)	**0.01**

Recessive model				
GT+GG	15 (29.4)	11 (10.8)	1.00 (Reference)	
TT	36 (70.6)	91 (89.2)	3.22 (1.26-8.18)	**0.01**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided χ2 test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.8 Statistical Analysis for polymorphic sites of *MSH2* (*rs63749993, rs2303425, rs2303426, rs4987188, rs2303428, rs17217772*) gene

5.8.1 Genotypic and allelic frequencies of *MSH2* polymorphic variants among cases and controls and their role in modifying the susceptibility towards lung cancer

As depicted in **Table 5.10** regarding the polymorphic variant of *MSH2 2063T>G* (*rs63749993*) polymorphism in control subjects, 211 (42.2%) were homozygous for the wild genotype type (*TT*), 63 (12.6%) were homozygous variant carriers (*GG*), and 226 (45.2%) were carrying the heterozygous (*TG*) genotype. In contrast, in the lung cancer cases, 277 (55.4%) were homozygous for the wild genotype type (*TT*), 196 (39.2%) had the heterozygous (*TG*) genotype, and 27 (5.4%) had homozygous variant genotype (*GG*). It is also reported that there is a noteworthy dissimilarity between the genotypic frequencies distribution of cases and controls (χ2=25.45; df =2; p<0.0001). Further, no deflection from Hardy-Weinberg equilibrium (HWE) was found for *MSH2* polymorphic variant {Cases: χ2=1.02, p= 0.310; controls: χ2=0.04, p=0.83}, Minor allele frequency (MAF) is as follows: cases=0.25, controls= 0.35 (**Table 5.10**). For *MSH2 IVS1+9G>C* (*rs2303426*), in the control population, 70 (14%) were carrying wild-type genotype (*GG*), 252 (50.4%) were carrying heterozygous genotype (*GC*), and 178 (35.6%) were carrying mutant genotype (*CC*). In contrast, in the lung cancer subject population, 105 (21.0%) were carrying wild genotype (*GG*), 230 (46%) were carrying heterozygous genotype (*GC*), and 165 (33.0%) were carrying mutant type genotype (*CC*). It is also reported that there is a noteworthy deviation between the genotypic frequencies distribution of cases and controls (χ2=8.49; df =2; **p=0.014**). Further, no deflection from Hardy-Weinberg equilibrium (HWE) was found for *MSH2* polymorphic variant {Cases: χ2=2.21, p=0.136; controls: χ2=1.64, p=0.199}, Minor allele frequency (MAF) is as follows: cases=0.560, controls= 0.60 (**Table 5.11**). For *MSH2 T>C/-6* (*rs2303428*), in the control population, 409 (81.8%) were carrying wild-type genotype (*TT*), 88 (17.6%) were carrying heterozygous genotype (*TC*), and 3 (0.6%) were carrying mutant genotype (*CC*). In contrast, in the lung cancer subject population, 314 (62.8%)

were carrying wild genotype (*TT*), 156 (31.2%) were carrying heterozygous genotype (*TC*), and 30 (6%) were carrying mutant type genotype (*CC*). It is also reported that there is a noteable distinction between the genotypic frequencies distribution of cases and controls ($\chi 2$=53.5; df =2; **p<0.0001**). Further, no deflection from Hardy-Weinberg equilibrium (HWE) was reported for *MSH2* polymorphic variant {Cases: $\chi 2$=3.10, p=0.07; controls: $\chi 2=0.554$, p=0.456}, Minor allele frequency (MAF) is as follows: cases=0.21, controls= 0.09 (**Table 5.12**). For *MSH2 Asn^{127}Ser* (*rs17217772*), in the control population, 474 (94.8%) were carrying wild-type genotype (*AA*), 25 (5.0%) were carrying heterozygous genotype (*AG*), and 1 (0.2%) were carrying mutant genotype (*GG*). In contrast, in the lung cancer subject population, 425 (85.0%) were carrying wild genotype (*AA*), 73 (14.6%) were carrying heterozygous genotype (*AG*), and 2 (0.4%) were carrying mutant type genotype (*GG*). It is also reported that there is a noticeable variation between the genotypic frequencies distribution of cases and controls ($\chi 2$=26.5; df =2; **p<0.0001**). Further, no delection from Hardy-Weinberg equilibrium (HWE) was reported for *MSH2* polymorphic variant {Cases: $\chi 2$=0.368, p=0.54; controls: $\chi 2$=1.17, p=0.27}, Minor allele frequency (MAF) is as follows: cases=0.077, controls= 0.027 (**Table 5.12**). We have used a logistic regression model for determining the association between *MSH2* polymorphism and the risk of the development of lung cancer (**Table 5.10**, **Table 5.11**, **Table 5.12**). Our results demonstrated that the *2063T>G* polymorphism was found to be associated with a low risk of developing lung cancer in the individuals harboring the heterozygous genotype (AOR=0.68; 95%CI=0.52-0.89; **p=0.006**), however after applying Bonferroni correction, this P value was not significant (Pcorr=0.036). Moreover, subjects carrying the mutant genotype (*GG*) manifested a protective effect toward acquiring the risk of lung cancer (AOR=0.33; 95%CI=0.20-0.55; **p<0.0001**). After applying Bonferroni correction, this P value was still found to be statistically significant (**Pcorr=0.0006**). Both the dominant and recessive models for this particular polymorphism also depicted low risk for developing lung cancer for combined genotype (*TG+GG vis. TT*) (AOR=0.62; 95% CI=0.48-0.80; **p=0.003**) and mutant type genotype (*GG vis. TT+TG*) (AOR=0.41; 95%CI=0.25-0.66; **p=0.0003**) respectively. However, after applying Bonferroni correction, this P-value for the dominant model was no longer statistically significant (Pcorr=0.018), in contrast to Pcorr=0.0018 for a recessive model that was still found be significant (**Table 5.10**). Our results also showed no significant association for the two polymorphisms *viz. 118T>C* and *1032G>A* towards susceptibility for lung cancer (**Table 5.11**,

Table 5.12). Based on the values of the Akaike information criterion (AIC), a dominant model with the lowest AIC value was found to be the best genotypic model indicating about 0.62-fold (95%CI=0.48-0.80; **p=0.03**) reduction in combined genotype (*TG+GG*) for *2063 T>G* polymorphism (**Table 5.10**).

Table 5.10: Genotypic and allelic distribution of the *MSH2* (*rs63749993*) genetic variant and its association with risk of lung cancer

Genotype *rs63749993*/ c.*2063T>G*	**Controls (500) N(%)**	**Cases (500) N(%)**	**AOR (95% CI)** [a]	***p***[b]	**Pcorr= *p***[b] ***6**	**AIC**
Co-dominant model						
TT	211 (42.2)	277 (55.4)	Reference			
TG	226 (45.2)	196 (39.2)	0.68 (0.52-0.89)	**0.006**	0.036	-2512
GG	63 (12.6)	27 (5.4)	0.33 (0.20-0.55)	**<0.0001**	0.0006	-1519
T(allele)	648	750				
G(allele)	352	250				
MAF	0.35	0.25				
		$\chi2$= 25.45, df=2		<0.0001		
HWE	0.04	1.02				
Dominant model						
TT	211 (42.2)	277 (55.4)	Reference			
TG+GG	289 (57.8)	223 (44.6)	0.62 (0.48-0.80)	**0.003**	0.018	-2764.5
Recessive model						
TT+TG	437 (87.4)	473 (94.6)	Reference			
GG	63 (12.6)	27 (5.4)	0.41 (0.25-0.66)	**0.0003**	0.0018	-2764.5

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. P corr indicates P-value after Bonferroni correction. AIC= Akaike's Information Criterion

Further, for *IVS1+9G>C* polymorphism, the codominant model also showed an inclination towards a decrease in risk of developing lung cancer for subjects who having a single variant allele (AOR=0.56; 95%CI=0.39-81; **p=0.001**) and mutant type genotype (AOR=0.57; 95%CI=0.39-0.83; **p=0.003**). After Bonferroni correction, the association remained statistically non-significant (Pcorr=0.006 and 0.018). The dominant model for the same polymorphism also concluded a decrease in the risk of developing lung cancer for the combined heterozygous and mutant type genotype (AOR=0.57; 95%CI=0.41-0.80; **p=0.001**); however, these values were not to be found statistically significant after applying Bonferroni correction (Pcorr=0.006) (**Table 5.11**). Based on the values of the Akaike information criterion (AIC), a dominant model with the lowest AIC value was found to be the best genotypic model indicating about 0.57-fold reduced risk (95%CI=0.41-0.80; p=0.006) for *IVS1+9G>C* polymorphism.

Table 5.11: Genotypic and allelic distribution of the *MSH2* (*rs2303426*) genetic variant and its association with risk of lung cancer

Genotype *rs2303426*/ *IVS1+9G>C*	**Controls (500) N(%)**	**Cases (500) N(%)**	**AOR (95% CI)[a]**	***p*[b]**	**Pcorr= *p*[b] *6**	**AIC**
Co-dominant model						
GG	70 (14.0)	105 (21.0)	Reference			
GC	252 (50.4)	230 (46.0)	0.56 (0.39-0.81)	**0.001**	0.006	-1813.0
CC	178 (35.6)	165 (33.0)	0.57 (0.39-0.83)	**0.003**	0.018	-1426.3
G(allele)	392	440				
C(allele)	608	560				
MAF	0.60	0.56				
HWE	1.64	2.21				
	$\chi 2$=8.49, df=2			0.014		
Dominant model						
GG	70 (14.0)	105 (21.0)	Reference			
GC+CC	430 (86.0)	395 (79.0)	0.57 (0.41-0.80)	**0.001**	0.006	-2764.5

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. P corr indicates P-value after Bonferroni correction. AIC= Akaike's Information Criterion

For the *MSH2 rs2303428 T>C/-6* polymorphic variant, the subjects who were harboring both the variant alleles (mutant genotype *CC*) expressed a 13-fold risk for developing lung cancer (AOR=13.31; 95%CI=3.99-44.3**; p<0.0001**). Both the dominant (*TC+CC vis. TT*) and recessive (*CC vis. TT+TC*) models for the same polymorphism also reported a 2.65 (AOR=2.65; 95%CI=1.97-3.55; **p<0.0001**) and 10-fold (AOR=10.4; 95%CI=3.16-34.7; **p=0.0001**) increased risk for developing lung cancer. Even after applying Bonferroni correction, the results for all the applied models were statistically significant (Pcorr=0.0006) (**Table 5.12**). For *MSH2 Asn^{127}Ser* polymorphism, our result concluded a 3-fold increase in risk for developing lung cancer in individuals harboring heterozygous type genotype (AOR=3.27; 95%CI=2.02-5.27; **p<0.0001**) in the codominant model. The dominant model for *Asn^{127}Ser* polymorphism also reported a 3-fold increased risk for lung cancer development in individuals carrying combined genotype (AOR=3.23; 95%CI=2.02-5.17; **p<0.0001**). Even after Bonferroni correction, the association remained statistically significant (Pcorr=0.0006) for the SNPs as mentioned above for the *MSH2* gene (**Table 5.12**). Based on the values of the Akaike information criterion (AIC), a dominant model with the lowest AIC value was found to be the best genotypic model indicating about 10-

fold increased risk (95%CI= 3.16-34.7; **Pcorr=0.0006**) towards lung cancer for *T>C/-6* polymorphism and 3.23-fold increased risk (95%CI= 2.02-5.17; **Pcorr=0.0006**) as to the best genotypic model for $Asn^{127}Ser$ polymorphism (**Table 5.12**).

Table 5.12: Genotypic and allelic distribution of the *MSH2* (*rs2303428* and *rs17217772*) genetic variant and its association with risk of lung cancer

Genotype *rs2303428*/ T>C/-6	**Controls (500) N(%)**	**Cases (500) N(%)**	**AOR (95% CI)[a]**	p^b	**Pcorr= p^b *6**	**AIC**
Co-dominant model						
TT	409 (81.8)	314 (62.8)	Reference			
TC	88 (17.6)	156 (31.2)	2.3 (1.69-3.11)	**<0.0001**	<0.0006	-2671.5
CC	3 (0.6)	30 (6.0)	13.3 (3.99-44.3)	**<0.0001**	<0.0006	-2075.8
T(allele)	906	784				
C(allele)	94	216				
MAF	0.09	0.21				
HWE	0.554	3.10				
	$\chi2$=53.5, df=2			P<0.0001		
Dominant model						
TT	409 (81.8)	314 (62.8)	Reference			
TC+CC	91 (18.2)	186 (37.2)	2.65 (1.97-3.55)	**<0.0001**	<0.0006	-2764.5
Recessive model						
TT+TC	497 (99.4)	470 (94.0)	Reference			
CC	3 (0.6)	30 (6.0)	10.4 (3.16-34.7)	**0.0001**	<0.0006	-2764.5

Genotype *rs17217772* *Asn[127]Ser*	**Controls (500) N (%)**	**Cases (500) N (%)**	**AOR (95% CI)[a]**	p^b	**Pcorr= p^b *6**	**AIC**
Co-dominant model						
AA	474 (94.8)	425 (85.0)	Reference			
AG	25 (5.0)	73 (14.6)	3.27 (2.02-5.27)	**<0.0001**	<0.0006	-2756.2
GG	1 (0.2)	2 (0.4)	2.06 (0.18-23.6)	0.55	3.3	-2487.7
A(allele)	973	923				
G(allele)	27	77				
MAF	0.027	0.077				
HWE	1.17	0.368				
	$\chi2$=26.5, df=2			P<0.0001		
Dominant model						
AA	474 (94.8)	425 (85.0)	Reference			
AG+GG	26 (5.2)	75 (15.0)	3.23 (2.02-5.17)	**<0.0001**	<0.0006	-2764.5
Recessive model						
AA+AG	499 (99.8)	498 (99.6)	Reference			
GG	1 (0.2)	2 (0.4)	-	-	-	-

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. P corr indicates P-value after Bonferroni correction. AIC= Akaike's Information Criterion

5.8.2 Genotypic and allelic distribution of the *MSH2* gene and risk of lung cancer based on histological subtypes

To assess the association of *MSH2* polymorphism and likelihood of lung cancer susceptibility, we further stratified our data based on histological subtypes as represented in **Table 5.13**, **Table 5.14**, **Table 5.15**. For *MSH2 2063T>G* polymorphism, subjects with the G allele showed a statistically decreased risk of lung adenocarcinoma than those with the T allele (OR = 0.45, 95% CI = 0.23-0.88; **p=0.02**). We could not find any association between *118T>C* and *1032G>A* polymorphism with the occurrence of ADCC. As shown in **Table 5.13** for the *MSH2 IVS1+9G>C* polymorphism, adenocarcinoma patients harboring mutant type genotype (*CC*) in the recessive model had a decreased risk of manifesting ADCC (AOR=0.58, 95%CI=0.39-0.86; **p=0.007**). Following Bonferroni correction, the p-value was found to be non-significant Pcorr=0.042). For *MSH2 rs2303428* (*T>C/-6*) polymorphism, using the *TT* genotype as a reference, ADCC susceptibility was found to be significantly higher in subjects with *TC* genotype (AOR=2.16; 95%CI= 1.43-3.26; **p=0.0002,** Pcorr=0.0012), *CC* genotype (AOR=12.0; 95%CI=3.11-46.9; **p=0.0003,** Pcorr=0.0018) and the dominant model (*TC* + *CC*; AOR=2.48; 95%CI=1.67-3.69; **p<0.0001;** Pcorr=0.0006) (**Table 5.13**). In the recessive model for the same polymorphism, individuals carrying heterozygous and mutant type genotype, respectively, have reported a 9-fold increased risk (AOR=9.48; 95%CI=2.5-35.9; **p=0.0009**) of developing ADCC. However, the results were statistically significant after applying Bonferroni correction (Pcorr=0.0054).

Our results further concluded that both the dominant and codominant model for *Asn*^{127}*Ser* polymorphism reported a 3-fold increase in the risk of developing ADCC (AOR=3.57; 95%CI=1.95-6.52; **p<0.0001**, Pcorr=0.0006). Based on ADCC histology, AIC predicted that the dominant model having the least AIC score is the best genotypic model, indicating approximately 2.48-fold (95%CI=1.67-3.69; **p=0.0006**) increased risk for combined genotype for *T>C/-6* polymorphism towards lung cancer. Similarly, for *Asn*^{127}*Ser* polymorphism, AIC predicted the dominant model with the least AIC score for the combined genotype as the best model, showing 3.57-fold (95% CI=1.97-6.46; **p=0.0006**) increased risk towards lung cancer (**Table 5.13**).

Table 5.13: Genotypic and allelic distribution of the *MSH2* genetic variant and its association with risk of lung cancer according to tumor histology (ADCC)

ADCC													
Genotype *rs63749993/ c.2063T>G /*	**Controls (500) N(%)**	**Cases (203) N(%)**	**AOR (95% CI)**[a]	p^b	**Pcorr= p^b *6**	**AIC**	**Genotype *rs17217772 Asn^{127}Ser***	**Controls (500) N(%)**	**Cases (203) N(%)**	**AOR (95% CI)**[a]	p^b	**Pcorr= p^b *6**	**AIC**
Co-dominant model							**Co-dominant model**						
TT	211 (42.2)	117 (57.7)	Reference				AA	474 (94.8)	171 (84.2)	Reference			
TG	226 (45.2)	73 (35.9)	0.74 (0.51-1.08)	0.12	0.72	-1530.4	AG	25 (5.0)	31 (15.4)	3.57 (1.95-6.52)	<**0.0001**	<0.0006	-1675.7
GG	63 (12..6)	13 (6.4)	0.45 (0.23-0.88)	**0.02**	0.12	-1007.1	GG	1 (0.2)	1 (0.4)	3.29 (0.19-55.2)	0.40	2.4	-1490.6
Dominant model							**Dominant model**						
TT	211 (42.2)	117 (57.7)	Reference				AA	474 (94.8)	171 (84.2)	Reference			
TG+GG	289 (57.8)	86 (42.3)	0.70 (0.49-1.01)	0.05	0.3	-1682.1	AG+GG	26 (5.2)	32 (15.8)	3.57 (1.97-6.46)	**<0.0001**	<0.0006	-1682.1
Recessive model							**Recessive model**						
TT+TG	437 (87.4)	190 (93.6)	Reference				AA+AG	499 (99.8)	202 (99.6)	Reference			
GG	63 (12.6)	13 (6.4)	0.56 (0.29-1.08)	0.08	0.48	-1682.1	GG	1 (0.2)	1 (0.4)	-	-	-	-
Genotype *rs2303426/ IVS1+9G>C*	**Controls (500) N(%)**	**Cases (203) N(%)**	**AOR (95% CI)**[a]	p^b	**Pcorr= p^b *6**	**AIC**	**Genotype *rs2303428/* T>C/-6**	**Controls (500) N(%)**	**Cases (203) N(%)**	**AOR (95% CI)**[a]	p^b	**Pcorr= p^b *6**	**AIC**
Co-dominant model							**Co-dominant model**						
GG	70 (14.0)	33 (16.3)	Reference				TT	409 (81.8)	129 (63.5)	Reference			
GC	252 (50.4)	110 (54.2)	0.81 (0.47-1.37)	0.43	2.58	-1139.8	TC	88 (17.6)	62 (30.5)	2.16 (1.43-3.26)	**0.0002**	0.0012	-1617.5
CC	178 (35.6)	60 (29.5)	0.49 (0.28-0.87)	**0.015**	0.09	-791.24	CC	3 (0.6)	12 (5.9)	12.0 (3.11-46.9)	**0.0003**	0.0018	-1247.8
Dominant model							**Dominant model**						
GG	70 (14.0)	33 (16.3)	Reference				TT	409 (81.8)	129 (63.5)	Reference			
GC+CC	430 (86.0)	170 (83.7)	0.67 (0.41-1.1)	0.11	0.66	-1682.1	TC+CC	91 (18.2)	74 (36.5)	2.48 (1.67-3.69)	**<0.0001**	<0.0006	-1682.1
Recessive model							**Recessive model**						
GG+GC	322 (64.4)	143 (70.4)	Reference				TT+TC	497 (99.4)	191 (94.1)	Reference			
CC	178 (35.6)	60 (29.6)	0.58 (0.39-0.86)	0.007	0.042	-1682.1	CC	3 (0.6)	12 (5.9)	9.48 (2.5-35.9)	**0.0009**	0.0054	-1682.1

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. P corr indicates P-value after Bonferroni correction. AIC= Akaike's Information Criterion

For the *MSH2 T>C/-6* polymorphism, subjects carrying heterozygous *TC* or homozygous *CC* genotype reported a 2-fold (AOR=1.93; 95%CI=1.30-2.86; **p=0.001**) and 9-fold increased risk for SQCC development (AOR=9.60; 95%CI=2.57-35.4; **p=0.0008**) in individuals harboring single mutant allele and individuals carrying both mutant alleles. When data from dominant (*TC+CC*) and recessive models (*CC*) were analyzed, we concluded that individuals having one mutant allele and both mutant alleles also reported a 2-fold and 8-fold increased risk of developing SQCC (AOR=2.18; 95%CI=1.5-3.19; **p<0.0001**) and (AOR=8.14; 95%CI=2.19-30.19; **p=0.0017**). However, after applying Bonferroni correction, the results were not statistically significant for the codominant model (Pcorr= 0.006, 0.0048) and recessive model (Pcorr= 0.0102) as compared to the dominant model (Pcorr= 0.0006), which was statistically significant. For *MSH2 Asn*^{127}*Ser*, the codominant and dominant models were found to be associated with a 3-fold increased risk for SQCC development (AOR=3.63; 95%CI=2.08-6.32; p<0.0001) and (AOR=3.58; 95%CI=2.07-6.20; p<0.0001). Moreover, these results were also statistically significant after applying Bonferroni correction (Pcorr=0.0006). Furthermore, we could not find any association of *118T>C* and *1032G>A* polymorphism with the risk of developing SQCC. Based on SQCC histology, AIC predicted dominant model having the lowest AIC score as the best model indicating 0.47-fold (95%CI= 0.33-0.66; **p=0.0006**) reduced risk of developing lung cancer for *2063T>G* polymorphism (**Table 5.14**).

All *rs2303428 T>C/-6* polymorphism models were related to lung cancer manifestation. With the *TT* genotype as a reference, SCLC susceptibility was found to be significantly higher in subjects with *TC* genotype in the codominant model (AOR=4.05; 95%CI=2.40-6.83; **p<0.0001**) and the dominant model (*TC + CC*; (AOR=4.37; 95%CI=2.82-6.7; **p<0.0001**). In the recessive model, the *CC* genotype of *rs2303428 T>C/-6* showed a significant increase (16.5-fold) in the risk of SCLC compared with the reference (*TT+TC* AOR=16.5; 95%CI=4.18-65.6; **p=0.0001**). These results were statistically significant even after applying Bonferroni correction (Pcorr=0.0006). Furthermore, as depicted in **Table 5.15**, when the association of *MSH2 2063T>G* polymorphism was evaluated for SCLC, subjects who had both copies of the mutant allele *(GG)* in the recessive model) showed a decrease in risk towards SCLC development (AOR=0.17; 95%CI=0.04-0.72**; p=0.01**). For *MSH2 118T>C* polymorphism, in the codominant model, subjects with the heterozygous genotype (*TC)* reported a 2-fold increased risk towards developing SCLC (AOR=2.18; 95%CI=1.32-3.60; **p=0.002**). When data from dominant

(*TC+CC*) and recessive model (*CC*) were analyzed, we concluded that individuals having either one copy (*TC*) or both copies of the mutant alleles (*CC*) reported a 2-fold increased risk for SCLC (AOR=2.11; 95%CI=1.29-3.45; **p=0.002**). Further, when the Bonferroni correction technique was applied, we did not find our result statistically significant (**Pcorr=0.012**). In this study, our results could not find any association of *Asn*^{127}*Ser* polymorphism with the risk of developing SCLC. Based on the Akaike information criterion (AIC), the recessive model with the least AIC score was found to be the best genotypic model indicating about 16.5-fold (95%CI=4.18-65.6; p=0.0006) increased risk in mutant type genotype towards lung cancer for *T>C/-6* polymorphism (**Table 5.15**).

Table 5.14: Genotypic and allelic distribution of the *MSH2* genetic variant and its association with risk of lung cancer according to tumor histology (SQCC)

SQCC													
Genotype *rs63749993/ c.2063T>G /*	**Controls (500) N(%)**	**Cases (205) N(%)**	**AOR (95% CI) [a]**	***p*[b]**	**Pcorr= *p*[b] *6**	**AIC**	**Genotype *rs17217772 Asn[127]Ser***	**Controls (500) N(%)**	**Cases (205) N(%)**	**AOR (95% CI)[a]**	***p*[b]**	**Pcorr= *p*[b] *6**	**AIC**
Co-dominant model							**Co-dominant model**						
TT	211 (42.2)	117 (57.1)	Reference				AA	474 (94.8)	171 (83.4)	Reference			
TG	226 (45.2)	76 (37.1)	0.51 (0.35-0.72)	**0.0002**	0.0012	-1544.6	AG	25 (5.0)	33 (16.1)	3.63 (2.08-6.32)	**<0.0001**	<0.0006	-1685.7
GG	63 (12..6)	12 (5.8)	0.29 (0.15-0.56)	**0.0003**	0.0018	-1002.6	GG	1 (0.2)	1 (0.5)	2.23 (0.13-36.5)	0.57	3.42	-1490.6
Dominant model							**Dominant model**						
TT	211 (42.2)	117 (57.1)	Reference				AA	474 (94.8)	171 (83.4)	Reference			
TG+GG	289 (57.8)	88 (42.9)	0.47 (0.33-0.66)	**<0.0001**	<0.0006	-1692.0	AG+GG	26 (5.2)	34 (16.6)	3.58 (2.07-6.20)	**<0.0001**	<0.0006	-1692.0
Recessive model							**Recessive model**						
TT+TG	437 (87.4)	193 (94.2)	Reference				AA+AG	499 (99.8)	204 (99.5)	Reference			
GG	63 (12.6)	12 (5.8)	0.41 (0.21-0.79)	0.0078	0.046	-1692.0	GG	1 (0.2)	1 (0.5)	-	-	-	-
Genotype *rs2303426/ IVS1+9G>C*	**Controls (500) N(%)**	**Cases (205) N(%)**	**AOR (95% CI)[a]**	***p*[b]**	**Pcorr= *p*[b] *6**	**AIC**	**Genotype *rs2303428/* T>C/-6**	**Controls (500) N(%)**	**Cases (205) N(%)**	**AOR (95% CI)[a]**	***p*[b]**	**Pcorr= *p*[b] *6**	**AIC**
Co-dominant model							**Co-dominant model**						
GG	70 (14.0)	53 (25.9)	Reference				TT	409 (81.8)	140 (68.3)	Reference			
GC	252 (50.4)	95 (46.3)	0.49 (0.32-0.76)	**0.001**	0.006	-1163.1	TC	88 (17.6)	55 (26.8)	1.93 (1.30-2.86)	**0.001**	0.006	-1637.9
CC	178 (35.6)	57 (27.8)	0.43 (0.26-0.69)	**0.0006**	0.0036	-875.39	CC	3 (0.6)	10 (4.9)	9.60 (2.57-35.4)	**0.0008**	0.0048	-1296.1
Dominant model							**Dominant model**						
GG	70 (14.0)	53 (25.9)	Reference				TT	409 (81.8)	140 (68.3)	Reference			
GC+CC	430 (86.0)	152 (74.1)	0.47 (0.31-0.72)	**0.0004**	0.0024	-1692.0	TC+CC	91 (18.2)	65 (31.7)	2.18 (1.5-3.19)	**<0.0001**	<0.0006	-1692.0
Recessive model							**Recessive model**						
GG+GC	322 (64.4)	148 (72.2)	Reference				TT+TC	497 (99.4)	195(95.1)	Reference			
CC	178 (35.6)	57 (27.8)	0.72 (0.50-1.03)	0.07	0.42	-1692.0	CC	3 (0.6)	10 (4.9)	8.14 (2.19-30.19)	**0.0017**	0.0102	-1692.0

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. P corr indicates P-value after Bonferroni correction. AIC= Akaike's Information Criterion

Table 5.15: Genotypic and allelic distribution of the *MSH2* genetic variant and its association with risk of lung cancer according to tumor histology (SCLC)

SCLC													
Genotype *rs63749993/ c.2063T>G*	**Controls (500) N(%)**	**Cases (84) N(%)**	**AOR (95% CI)^a**	**p^b**	**Pcorr= p^b*6**	**AIC**	**Genotype *rs2303425 118T>C***	**Controls (500) N(%)**	**Cases (84) N(%)**	**AOR (95% CI)^a**	**p^b**	**Pcorr= p^b*6**	**AIC**
Co-dominant model							**Co-dominant model**						
TT	211 (42.2)	39 (46.4)	Reference				TT	350 (70.0)	42 (50.0)	Reference			
TG	226 (45.2)	43 (51.2)	0.86 (0.52-1.41)	0.55	3.3	-897.82	TC	136 (27.2)	40 (47.6)	2.18 (1.32-3.60)	**0.002**	0.012	-929.9
GG	63 (12..6)	2 (2.4)	0.16 (0.03-0.69)	**0.01**	0.06	-479.22	CC	14 (2.8)	2 (2.4)	1.11 (0.24-5.08)	0.89	5.34	-550.1
Dominant model							**Dominant model**						
TT	211 (42.2)	39 (46.4)	Reference				TT	350 (70.0)	42 (50.0)	Reference			
TG+GG	289 (57.8)	45 (53.6)	0.71 (0.43-1.16)	0.17	1.02	-954.11	TC+CC	150 (30.0)	42 (50.0)	2.11 (1.29-3.45)	**0.002**	0.012	-954.1
Recessive model							**Recessive model**						
TT+TG	437 (87.4)	82 (97.6)	Reference				TT+TC	486 (97.2)	82	Reference			
GG	63 (12.6)	2 (2.4)	0.17 (0.04-0.72)	**0.01**	0.06	-954.11	CC	14 (2.8)	2	0.94 (0.21-4.29)	0.94	5.64	-954.1
Genotype *rs2303428*/T >C/-6	**Controls (500) N(%)**	**Cases (84) N(%)**	**AOR (95% CI)^a**	**p^b**	**Pcorr= p^b*6**	**AIC**	**Genotype *rs17217772 Asn^{127}Ser***	**Controls (500) N(%)**	**Cases (84) N(%)**	**AOR (95% CI)^a**	**p^b**	**Pcorr= p^b*6**	**AIC**
Co-dominant model							**Co-dominant model**						
TT	409 (81.8)	40 (47.6)	Reference				AA	474 (94.8)	76 (90.5)	Reference			
TC	88 (17.6)	36 (42.9)	4.05 (2.40-6.83)	**<0.0001**	<0.0006	-889.01	AG	25 (5.0)	8 (9.5)	2.36 (1.00-5.56)	**0.045**	0.27	-953.4
CC	3 (0.6)	8 (9.5)	5.01 (2.51-9.98)	**<0.0001**	<0.0006	-607.54	GG	1 (0.2)	0	0	0		-876.2
Dominant model							**Dominant model**						
TT	409 (81.8)	40 (47.6)	Reference				AA	474 (94.8)	76 (90.5)	Reference			
TC+CC	91 (18.2)	44 (52.4)	4.37 (2.82-6.7)	**<0.0001**	<0.0006	-954.11	AG+GG	26 (5.2)	8 (9.5)	2.30 (0.98-5.39)	0.05	0.30	-954.1
Recessive model							**Recessive model**						
TT+TC	497 (99.4)	76 (90.5)	Reference				AA+AG	499 (99.8)	84 (100)	Reference			
CC	3 (0.6)	8 (9.5)	16.5 (4.18-65.6)	**0.0001**	0.0006	-954.11	GG	1 (0.2)	0 (0)	-	-	-	-

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. P corr indicates P-value after Bonferroni correction. AIC= Akaike's Information Criterion

5.8.3 Effect of smoking status on the association of *MSH2* polymorphisms and lung cancer risk

Smoking is a well-established risk factor for the occurrence of lung cancer; hence in this study, we have also evaluated the risk of lung cancer development due to *MSH2* polymorphic variants based on the smoking habits of recruited subjects (**Table 5.16**, **Table 5.17**). In the case of for the *2063T>G* polymorphic variant subjects who were smokers and carrying either the heterozygous (*TG*) and mutant type genotype (*GG*), a decreased propensity towards lung cancer induced by smoking was visible for both genotypes (AOR=0.51; 95%CI=0.38-0.68; **p<0.0001**) and (AOR= 0.21; 95%CI=0.12-0.38; **p<0.0001**) respectively. Similarly, in both the dominant (*TG+GG*; AOR=0.44; 95%CI=0.33-0.58; **p<0.0001**) and recessive (*GG*; AOR=0.30; 95%CI=0.17-0.52; **p<0.0001**) models were also showing a decreased disposition towards smoking-induced lung cancer as depicted in **Table 5.16**. Similarly, in the co-dominant model's *IVS1+9G>C* polymorphism, individuals carrying a single mutant allele also showed a reduced risk of developing lung cancer (AOR=0.62; 95%CI=0.43-0.91; **p=0.01**). Moreover, in the dominant model, to increase the power of association, both heterozygous and homozygous variant genotypes were combined (*GC+CC*); this single subgroup also reported decreased propensity towards lung cancer induced by smoking (AOR=0.65; 95%CI=0.45-0.92; **p=0.01**). For *T>C/-6* polymorphism, individuals harboring a single copy of mutant allele (*TC*) reported having 2.65-fold (AOR=2.65; 95%CI=1.90-3.70; **p<0.0001**) increased risk towards smoking-induced lung cancer in comparison to subjects who were carrying the wild alleles (*TT*). Similarly, in the dominant and recessive models, combined genotype (*TC+CC*) and mutant type (*CC*) genotype reported a 3-fold (AOR=2.98; 95%CI=2.16-4.12; **p<0.0001**) and 9-fold increased tendency towards smoking-induced lung cancer (AOR=8.85; 95%CI=2.62-29.8; **p=0.0004**). In the case of *Asn^{127}Ser* polymorphism of *MSH2* gene, for the co-dominant model, heterozygous type genotype showed a 3-fold increase risk of developing lung cancer (AOR=2.96; 95%CI=1.80-4.89; **p<0.0001**), similar results were also reported in the combined genotype is dominant (*AG+GG*) model, 3-fold increased risk of developing lung cancer due to smoking (AOR=2.95; 95%CI=1.80-4.80; **p<0.0001**). We could not find any association between smoking-induced lung cancer and *MSH2 118T>C* and *1032G>A* polymorphisms (**Table 5.16**).

Table 5.16: Relationship of different *MSH2* genotypes with smokers of cases and control

SMOKERS									
rs6374999/ c.*2063T>G*	**Controls (449) N(%)**	**Cases (398) N(%)**	**AOR (95% CI) [a]**	***p*[b]**	***rs17217772 Asn^{127}Ser***	**Controls (449) N(%)**	**Cases (398) N(%)**	**AOR (95% CI) [a]**	***p*[b]**
Co-dominant model					Co-dominant model				
TT	166 (37.0)	223 (56.0)	Reference		AA	423 (94.2)	341 (85.7)	Reference	
TG	220 (49.0)	157 (39.4)	0.51 (0.38-0.68)	**<0.0001**	AG	25 (5.6)	55 (13.8)	2.96 (1.80-4.89)	**<0.0001**
GG	63 (14.0)	18 (4.5)	0.21 (0.12-0.38)	**<0.0001**	GG	1 (0.2)	2 (0.5)	2.08 (0.16-26.4)	0.57
Dominant model					Dominant model				
TT	166 (37.0)	223 (56.0)	Reference		AA	423 (94.2)	341 (85.7)	Reference	
TG+GG	283 (63.0)	175 (44.0)	0.44 (0.33-0.58)	**<0.0001**	AG+GG	26 (5.8)	57 (14.3)	2.95 (1.80-4.80)	**<0.0001**
Recessive model					Recessive model				
TT+TG	386 (86.0)	380 (95.5)	Reference		AA+AG	448 (99.8)	396 (99.5)	Reference	
GG	63 (14.0)	18 (4.5)	0.30 (0.17-0.52)	**<0.0001**	GG	1 (0.2)	2 (0.5)	-	-
rs2303426/ IVS1+9G>C	**Controls (449) N(%)**	**Cases (398) N(%)**	**AOR (95% CI) [a]**	***p*[b]**	***rs2303428*/T> C/-6**	**Controls (449) N(%)**	**Cases (398) N(%)**	**AOR (95% CI) [a]**	***p*[b]**
Co-dominant model					Co-dominant model				
GG	70 (15.6)	87 (21.9)	Reference		TT	372 (82.9)	246 (61.8)	Reference	
GC	231 (51.4)	181 (45.5)	0.62 (0.43-0.91)	**0.01**	TC	74 (16.5)	129 (32.4)	2.65 (1.90-3.70)	**<0.0001**
CC	148 (33.0)	130 (32.7)	0.68 (0.46-1.02)	0.06	CC	3 (0.6)	23 (5.8)	11.56 (3.41-39.1)	**0.0001**
Dominant model					Dominant model				
GG	70 (15.6)	87 (21.9)	Reference		TT	372 (82.9)	246 (61.8)	Reference	
GC+CC	379 (84.4)	311 (78.1)	0.65 (0.45-0.92)	**0.01**	TC+CC	77 (17.1)	152 (38.2)	2.98 (2.16-4.12)	**<0.0001**
Recessive model					Recessive model				
GG+GC	301 (77.0)	268 (67.3)	Reference		TT+TC	446 (99.4)	375 (94.2)	Reference	
CC	148 (33.0)	130 (32.7)	0.95 (0.71-1.27)	0.74	CC	3 (0.6)	23 (5.8)	8.85 (2.62-29.8)	**0.0004**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

Furthermore, we have also analyzed our data for non-smoking subjects (**Table 5.17**), and we concur that by leaving out the *MSH2 2063T>G*, none of the other *MSH2* polymorphisms had any impact on modulating the susceptibility towards lung cancer. Contrary to the above results, where we observed a reduced risk towards lung cancer for the *2063T>G* polymorphism for smokers, the patients who were non-smokers and carrying a single polymorphic allele (*TG*) had a 5-fold increased risk of developing lung cancer (AOR=5.53; 95%CI=2.03-15.02; **p=0.0008**) (**Table 5.17**).

Table 5.17: Relationship of different *MSH2* genotypes with non-smokers of cases and controls

rs6374999/ c.2063T>G /	**Controls (51) N(%)**	**Cases (102) N(%)**	**AOR (95% CI) [a]**	***p*[b]**
Co-dominant model				
TT	45 (88.2)	54 (53.0)	Reference	
TG	6 (11.8)	39 (38.2)	5.53 (2.03-15.0)	**0.0008**
GG	0 (0)	9 (8.8)	1.25E	-
Dominant model				
TT	45 (88.2)	54 (53.0)	Reference	
TG+GG	6 (11.8)	48 (47.0)	7.42 (2.70-19.4)	**0.0001**
Recessive model				
TT+TG	51 (100)	93 (91.2)	Reference	
GG	0 (0)	9 (8.8)	9.8E	0.9936

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

When we further analyzed light smokers (pack years <=20) (**Table 5.18**), for *2063T>G* polymorphism a weak association towards lung cancer susceptibility was observed for heterozygous (*TG*) and mutant type genotype (*GG*) (AOR=0.64; 95%CI=0.42-0.98; **p=0.04**) and (AOR=0.43; 95%CI=0.20-0.94; **p=0.03**). A similar result was reported in the dominant model for combined genotype (TG+GG; AOR=0.60; 95%CI=0.40-0.89; **p=0.012**). In the case of *IVS1+9G>C* polymorphism, we observed that this polymorphic variant did not increase the risk of lung cancer in individuals carrying one mutant allele (AOR=0.51; 95%CI=0.29-0.90; **p=0.02**). Contrary to the above results, when we analyzed *MSH2 T>C/-6* polymorphism, all the genetics models applied for *T>C/-6* were associated with an increased susceptibility for lung cancer risk for light smokers, as shown in **Table 5.18**. Based on our data for light smokers, we could not find any significant association between *118T>C*, *1032G>A*, and *Asn*^{127}*Ser MSH2* polymorphism and risk of developing lung cancer.

In the case of heavy smokers, for *MSH2 2063T>G* polymorphism, we observed that subjects carrying either heterozygous (AOR=0.32; 95%CI=0.20-0.53; **p<0.0001**) or mutant genotype (AOR=0.08; 95%CI=0.03-0.21; **p<0.0001**) seemed to have a reduced risk towards lung cancer as shown in **Table 5.19**. A similar trend of reduced risk for lung cancer was also visible in the case of the *MSH2 118T>C* polymorphism, in both the co-dominant (AOR=0.47; 95%CI=0.26-0.85; **p=0.01**) and dominant models (*TC+CC*; AOR=0.56; 95%CI=0.38-0.83; **p=0.004**) where a decrease in the propensity towards lung cancer in heavy smokers was observed. Our study did not find any association of *IVS1+9G>C* and *1032G>A* polymorphism with the risk of developing heavy smoking-induced lung cancer. However, when we additionally evaluated *T>C/-6* polymorphism, we observed that the subjects who were heavy smokers had the same comparable risk for lung cancer susceptibility as the subjects who smoked fewer pack-years (**Table 5.19**). Lung cancer subjects who were either having the heterozygous (*TC*) or mutant genotype (*CC*) reported a two and a 7-fold increase in risk associated with developing lung cancer in the com-dominant model (AOR=2.83; 95%CI=1.35-4.21; p=**0.0026** and AOR= 7.19; 95%CI=0.91-56.6; p=0.061). In the dominant model for the same polymorphism, we reported a 2-fold increase in the propensity towards lung cancer for combined genotype (TC+CC) (AOR=2.63; 95%CI=1.52-4.56; **p=0.0006**) (**Table 5.19**).

Table 5.18: Relationship of different *MSH2* genotypes with light smokers of cases and controls

LIGHT SMOKERS									
rs6374999/ c.*2063T>G /*	**Controls (323) N (%)**	**Cases (155) N (%)**	**AOR (95% CI)[a]**	***p*[b]**	***rs2303425 118T>C***	**Controls (323) N (%)**	**Cases (155) N (%)**	**AOR (95% CI)[a]**	***p*[b]**
Co-dominant model					Co-dominant model				
TT	131 (40.6)	80 (51.6)	Reference		TT	227 (70.3)	99 (63.9)	Reference	
TG	152 (47.1)	65 (41.9)	0.64 (0.42-0.98)	**0.04**	TC	91 (28.2)	53 (34.2)	1.32 (0.86-2.03)	0.19
GG	40 (12.4)	10 (6.5)	0.43 (0.20-0.94)	**0.03**	CC	5 (1.5)	3 (1.9)	1.45 (0.33-6.28)	0.61
Dominant model					Dominant model				
TT	131 (40.6)	80 (51.6)	Reference		TT	227 (70.3)	99 (63.9)	Reference	
TG+GG	192 (59.4)	75 (48.4)	0.60 (0.40-0.89)	**0.012**	TC+CC	96 (29.7)	56 (36.1)	1.33 (0.87-2.02)	0.17
Recessive model					Recessive model				
TT+TG	283 (87.6)	145 (93.5)	Reference		TT+TC	318 (98.5)	152 (98.1)	Reference	
GG	40 (12.4)	10 (6.5)	0.53 (0.25-1.11)	0.09	CC	5 (1.5)	3 (1.9)	1.32 (0.30-5.7)	0.70
rs2303426/ IVS1+9G>C	**Controls (323) N(%)**	**Cases (155) N(%)**	**AOR (95% CI) [a]**	***p*[b]**	***rs2303428*/T> C/-6**	**Controls (323) N (%)**	**Cases (155) N (%)**	**AOR (95% CI) [a]**	***p*[b]**
Co-dominant mode					Co-dominant model				
GG	42 (13.0)	31 (20.0)	Reference		TT	266 (82.4)	86 (55.5)	Reference	
GC	171 (52.9)	64 (41.3)	0.51 (0.29-0.90)	**0.02**	TC	55 (17.0)	58 (37.4)	3.53 (2.23-5.58)	**<0.0001**
CC	110 (34.1)	60 (38.7)	0.73 (0.41-1.30)	0.29	CC	2 (0.6)	11 (7.1)	20.4 (4.41-94.5)	**0.0001**
Dominant model					Dominant model				
GG	42 (13.0)	31 (20.0)	Reference		TT	266 (82.4)	86 (55.5)	Reference	
GC+CC	281 (87.0)	124 (80.0)	0.60 (0.35-1.02)	0.06	TC+CC	57 (17.6)	69 (44.5)	4.08 (2.62-6.34)	**<0.0001**
Recessive model					Recessive model				
GG+GC	213 (65.9)	95 (61.3)	Reference		TT+TC	321 (99.4)	144 (92.9)	Reference	
CC	110 (34.1)	60 (38.7)	1.20 (0.79-1.80)	0.37	CC	2 (0.6)	11 (7.1)	13.04 (2.83-59.9)	**0.001**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table

Table 5.19: Relationship of different *MSH2* genotypes with heavy smokers of cases and controls

HEAVY SMOKERS									
rs6374999/ c.*2063T>G* /	**Controls (126) N(%)**	**Cases (243) N(%)**	**AOR (95% CI) [a]**	***p*[b]**	***rs2303425 118T>C***	**Controls (126) N(%)**	**Cases (243) N(%)**	**AOR (95% CI) [a]**	***p*[b]**
Co-dominant model					Co-dominant model				
TT	35 (27.8)	143 (58.8)	Reference		TT	72 (57.1)	168 (69.1)	Reference	
TG	68 (53.9)	92 (37.9)	0.32 (0.20-0.53)	**<0.0001**	TC	45 (35.8)	70 (28.8)	0.62 (0.38-1.00)	0.05
GG	23 (18.3)	8 (3.3)	0.08 (0.03-0.21)	**<0.0001**	CC	9 (7.1)	5 (2.1)	0.47 (0.26-0.85)	**0.01**
Dominant model					Dominant model				
TT	35 (27.8)	143 (58.8)	Reference		TT	72 (57.1)	168 (69.1)	Reference	
TG+GG	91 (72.2)	100 (41.2)	0.26 (0.16-0.42)	**<0.0001**	TC+CC	54 (42.9)	75 (30.8)	0.56 (0.38-0.83)	**0.004**
Recessive model					Recessive model				
TT+TG	103 (81.7)	235 (96.7)	Reference		TT+TC	117 (92.9)	238 (97.9)	Reference	
GG	23 (18.3)	8 (3.3)	0.15 (0.06-0.36)	**<0.0001**	CC	9 (7.1)	5 (2.1)	0.28 (0.09-0.87)	**0.02**
***rs2303428*/T> C/-6**	**Controls (126) N(%)**	**Cases (243) N(%)**	**AOR (95% CI) [a]**	***p*[b]**	***rs17217772 Asn^{127}Ser***	**Controls (126) N(%)**	**Cases (243) N(%)**	**AOR (95% CI) [a]**	***p*[b]**
Co-dominant model					Co-dominant model				
TT	106 (84.1)	160 (65.9)	Reference		AA	122 (96.8)	203 (83.5)	Reference	
TC	19 (15.1)	71 (29.2)	2.83 (1.35-4.21)	**0.0026**	AG	4 (3.2)	39 (16.1)	6.06 (2.11-17.4)	**0.008**
CC	1 (0.88)	12 (4.9)	7.19 (0.91-56.6)	**0.061**	GG	0 (0)	1 (0.4)	-	-
Dominant model					Dominant model				
TT	106 (84.1)	160 (65.9)	Reference		AA	122 (96.8)	203 (83.5)	Reference	
TC+CC	20 (15.9)	83 (34.1)	2.63 (1.52-4.56)	**0.0006**	AG+GG	4 (3.2)	40 (16.5)	6.27 (2.18-18.03)	**0.0007**
Recessive model					Recessive model				
TT+TC	125 (99.2)	231 (95.1)	Reference		AA+AG	126 (100)	242 (99.6)	Reference	
CC	1 (0.8)	12 (4.9)	5.85 (0.74-46.0)	0.093	GG	0 (0)	1 (0.4)	-	-

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age and gender. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.8.4 Relation between *MSH2* polymorphic variants Genotypic Distribution based on gender

The stratification analysis for both cases and controls was further completed based on gender; results are depicted in **Table 5.20**. The mutation of *MSH2 2063T>G* in males showed significantly decreased association towards genetic predisposition of lung cancer in the co-dominant (*GG* vs. *TT*; AOR=0.25; 95%CI=0.14-0.43; **p<0.0001**) & (*TG* vs. *TT*; AOR=0.60; 95%CI=0.45-0.80; p=0.0006) and the dominant model (*TG+GG* vs. *TT*; AOR=0.52; 95%CI=0.40-0.69; p<0.0001). Similarly, the *IVS1+9G>C* polymorphism also was reporte to diminish the likelihood of lung cancer in males significantly, both in the co-dominant (*GC* vs. *GG*, AOR=0.59; 95%CI=0.41-0.86; **p=0.006**) & (*CC* vs. *GG*; AOR=0.66; 95%CI=0.44-0.98; **p=0.04**) and dominant models (*GC+CC* vs. *GG*; AOR=0.59; 95%CI=0.41-0.86; **p=0.009**). On the contrary, when we analyzed the *T>C/-6* variant, we reported there were increased association between *T>C/-6* and lung cancer susceptibility in males in the dominant (*TC+CC* vs. *TT*, AOR=2.98; 95%CI=2.17-4.11; **p<0.0001**), recessive (*CC* vs. *TT+TC*, AOR=14.41; 95%CI=3.38-61.4; **p=0.003**) and co-dominant models (*TC* vs. *TT*, AOR=2.58; 95%CI=1.85-3.58; **p<0.0001**) & (*CC* vs. *TT*; AOR=18.89; 95%CI=4.41-80.8; **p=0.0001**). Moreover, there was a noteworthy relationship between $Asn^{127}Ser$ and lung cancer risk, where male subjects harboring heterozygous genotype (*AG)* reported a 3-fold higher risk of developing lung cancer (AOR=2.98; 95%CI=1.81-4.89; p<0.0001) in the co-dominant model. A similar trend was also visible for male subjects in the dominant model (*AG+GG* vs. *AA*; AOR=2.94; 95%CI=1.79-4.78; **p<0.0001**) (**Table 5.20**).

Table 5.20: Relationship of different *MSH2* polymorphism based on the male population

MALE									
rs6374999/ c.*2063T>G*	**Controls (455) N (%)**	**Cases (401) N (%)**	**AOR (95% CI) [a]**	***p*[b]**	***rs2303428*/T> C/-6**	**Controls (455) N(%)**	**Cases (401) N(%)**	**AOR (95% CI)[a]**	***p*[b]**
Co-dominant model					Co-dominant model				
TT	179 (39.3)	224 (55.9)	Reference		TT	377 (82.9)	248 (61.9)	Reference	
TG	213 (46.8)	157 (39.2)	0.60 (0.45-0.80)	**0.0006**	TC	76 (16.7)	128 (31.9)	2.58 (1.85-3.58)	**<0.0001**
GG	63 (13.9)	20 (4.9)	0.25 (0.14-0.43)	**<0.0001**	CC	2 (0.4)	25 (6.2)	18.89 (4.41-80.8)	**0.0001**
Dominant model					Dominant model				
TT	179 (39.3)	224 (55.9)	Reference		TT	377 (82.9)	248 (61.9)	Reference	
TG+GG	276 (60.7)	177 (44.1)	0.52 (0.40-0.69)	**<0.0001**	TC+CC	78 (17.1)	153 (38.1)	2.98 (2.17-4.11)	**<0.0001**
Recessive model					Recessive model				
TT+TG	392 (86.1)	381 (95.1)	Reference		TT+TC	453 (99.6)	376 (93.8)	Reference	
GG	63 (13.9)	20 (4.9)	9.8E	0.99	CC	2 (0.4)	25 (6.2)	14.41 (3.38-61.4)	**0.003**
rs2303426/ IVS1+9G>C	**Controls (455) N(%)**	**Cases (401) N(%)**	**AOR (95% CI)[a]**	***p*[b]**	***rs4987188/ 1032G>A***	**Controls (455) N(%)**	**Cases (401) N(%)**	**AOR (95% CI)[a]**	***p*[b]**
Co-dominant model					Co-dominant model				
GG	69 (15.2)	87 (21.7)	Reference		GG	448 (98.5)	394 (98.3)	Reference	
GC	237 (52.1)	185 (46.1)	0.59 (0.41-0.86)	**0.006**	GA	7 (1.5)	7 (1.7)	1.08 (0.37-3.15)	0.88
CC	149 (32.7)	129 (32.2)	0.66 (0.44-0.98)	**0.04**	AA	0 (0)	0 (0)	-	-
Dominant model					Dominant model				
GG	69 (15.2)	87 (21.7)	Reference		GG	448 (98.5)	394 (98.3)	Reference	
GC+CC	386 (84.8)	314 (78.3)	0.62 (0.44-0.89)	**0.009**	GA+AA	7 (1.5)	7 (1.7)	1.08 (0.37-3.15)	0.88
Recessive model					Recessive model				
GG+GC	306 (67.3)	272 (67.8)	Reference		GG+GA	455 (100)	401 (100)	Reference	
CC	149 (32.7)	129 (32.2)	0.96 (0.72-1.28)	0.81	AA	0	0 (0)	-	-

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age and smoking status. (b)-Two-sided χ2 test for either genotype distribution allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table.

5.8.5 Combinatorial analysis of the risk associated with different polymorphic sites of *MSH2* gene

Further, SNP-SNP interaction of different genotypic combinations of *MSH2* gene was analyzed to determine a significant difference between lung cancer subjects and controls recruited for this study. Since we have investigated six SNPs in this study, we considered wild-type genotype as a reference and compared combined genotype of heterozygous and mutant type with it for all six SNPs. The genotypic frequencies and respective odds ratios are represented in **Table 5.21**,**Table 5.22**, **Table 5.23**. When combined genotype of {*MSH2 118T>C* (H) + *MSH2* 2063*T>G* (H)}+{*MSH2 118T>C* (M)+ *MSH2 2063T>G* (M)} was compared with *MSH2 118T>C* (W) +*MSH2 2063T>G* (W), we reported a slight decrease in likelihood of manifesting lung cancer (AOR=0.77; 95%CI=0.53-1.14; p=0.20), similar low risk was observed for patients suffering from ADCC (AOR=0.65; 95%CI=0.36-1.18; p=0.15), SQCC (AOR= 0.52; 95%CI=0.30-0.88; **p=0.01**; Fdr p=0.012). However, an increased risk is observed in patients suffering from SCLC (AOR=1.57; 95%CI=0.82-3.01; p=0.16). Further, combined genotype of {*118T>C* (H) + *IVS1+9G>C* (H)} + {*118T>C* (M) + *IVS1+9G>C* (M)} was compared with *118T*>C (W) + *IVS1*+9G>C (W) genotype, it reported a low possibility of lung cancer development (AOR=0.12; 95%CI=0.05-0.26; **p<0.0001**; Fdr p=0.0003), similar results were reported with patients suffering from ADCC (AOR=0.21; 95%CI=0.07-0.60; **p=0.003**; Fdr p=0.004), SQCC (AOR=0.08; 95%CI=0.03-0.22; **p<0.0001**; Fdr p=0.0003) and SCLC (AOR=0.09; 95%CI=0.01-0.42; **p=0.002**; Fdr p=0.003) (**Table 5.21**).

However, when we analyzed {*118T*>C (H) + *T>C/-6* (H)} + {*118T*>C (M) + *MSH2 T>C/-6* (M)} and compared it with {*118T*>C (W) + *T>C/-6* (W)} genotype, a 2-fold higher risk towards lung cancer development was clearly evident for patients when considered without stratifying on the basis of histology (AOR=2.59; 95%CI=1.64-1.40; **p<0.0001**; Fdr p=0.0003) and further patients suffering from ADCC (AOR=2.07; 95%CI=1.07-4.01; **p=0.02**; Fdr p=0.02). Moreover, subjects which were histo-pathologically confirmed to have SCLC also reported a 5-fold increase in risk towards lung cancer development (AOR=5.58; 95%CI=2.9-10.7; **p<0.0001**; Fdr p=0.0003) (**Table 5.22**)

Table 5.21: Combinatorial analysis of the risk associated with different polymorphic sites of *MSH2* gene

	SQCC					**SCLC**				
	Controls (246) N(%)	Cases (112) N(%)	AOR (95% CI) [a]	p^b	FDR p-value	Controls (246) N(%)	Cases (47) N(%)	AOR (95% CI) [a]	p^b	FDR p-value
MSH2 118T>C (W) +MSH2 2063T>G (W)	175 (71.1)	88 (78.6)	Reference			175 (71.1)	23 (48.9)	Reference		
{MSH2 118 T>C (H) + MSH2 2063 T-G (H)}+{MSH2 118T>C (M)+ MSH2 2063T>G (M)}	71 (28.9)	24 (21.4)	0.52 (0.30-0.88)	**0.01**	0.012	71 (28.9)	24 (51.1)	1.57 (0.82-3.01)	0.16	
	OVERALL					**ADCC**				
	Controls (64) N(%)	Cases (91) N(%)	AOR (95% CI) [a]	p^b	FDR p-value	Controls (64) N(%)	Cases (37) N(%)	AOR (95% CI) [a]	p^b	FDR p-value
MSH2 118T>C (W) +MSH2 IVS1+9G>C(W)	10 (15.6)	60 (65.9)	Reference			10 (15.6)	20 (54.1)	Reference		
{MSH2 118T>C (H) + MSH2 IVS1+9G>C (H)}+{MSH2 118 T-C (M)+MSH2 IVS1+9G>C (M)}	54 (84.4)	31 (34.1)	0.12 (0.05-0.26)	**<0.0001**	0.0003	54 (84.4)	17 (45.9)	0.21 (0.07-0.60)	**0.003**	0.004
	SQCC					**SCLC**				
	Controls (64) N(%)	Cases (42) N(%)	AOR (95% CI) [a]	p^b	FDR p-value	Controls (64) N(%)	Cases (12) N(%)	AOR (95% CI) [a]	p^b	FDR p-value
MSH2 118T>C (W) +MSH2 IVS1+9G>C (W)	10 (15.6)	31 (73.8)	Reference			10 (15.6)	9 (75)	Reference		
{MSH2 118T>C (H) + MSH2 IVS1+9G>C(H)}+{MSH2 118T>C (M)+MSH2 IVS1+9G>C (M)}	54 (84.4)	11 (26.2)	0.08 (0.03-0.22)	**<0.0001**	0.0003	54 (84.4)	3 (25)	0.09 (0.01-0.42)	**0.002**	0.003

a)-Adjusted Odds ratios, 95% confidence intervals and their corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender and smoking status. (b)-Two-sided χ2 test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. FDR: False Discovery Rate

Table 5.22: Combinatorial analysis of the risk associated with different polymorphic sites of *MSH2* gene

	OVERALL					**ADCC**				
	Controls (325) N(%)	**Cases (299)N(%)**	**AOR (95% CI) [a]**	p^b	**FDR p-value**	**Controls (325) N(%)**	**Cases (110)N(%)**	**AOR (95% CI) [a]**	p^b	**FDR p-value**
MSH2 118T>C (W) +MSH2 T>C/-6(W)	293 (90.1)	233 (77.9)	Reference			293 (90.1)	94 (85.5)	Reference		
{MSH2 118T>C (H) + MSH2 T>C/-6	32 (9.9)	66 (22.1)	2.59 (1.64-1.40)	**<0.0001**	0.0003	32 (9.9)	16 (14.5)	2.07 (1.07-4.01)	**0.02**	0.02

(H)}+{MSH2 118T>C (M)+MSH2 T>C/-6 (M)}										
	SQCC					**SCLC**				
	Controls (325) N(%)	**Cases (127) N (%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**	**Controls (325) N(%)**	**Cases (50) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**
MSH2 118T>C (W) +MSH2 T>C/-6(W)	293 (90.1)	107 (84.3)	Reference			293 (90.1)	30 (60.0)	Reference		
{MSH2 118T>C (H) + MSH2 T>C/-6 (H)}+{MSH2 118T>C (M)+MSH2 T>C/-6 (M)}	32 (9.9)	20 (15.7)	1.62 (0.9-2.92)	0.107		32 (9.9)	20 (40)	5.58 (2.9-10.7)	**<0.0001**	0.0003

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender, and smoking status. (b)-Two-sided $\chi2$ test for either genotype distribution or allelic frequencies between the cases and controls. The number in bold indicates the significant values in the table. FDR: False Discovery Rate

Table 5.23: Combinatorial analysis of the risk associated with different polymorphic sites of *MSH2* gene

	OVERALL					**ADCC**				
	Controls (85) N(%)	**Cases (115) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**	**Controls (85) N(%)**	**Cases (43) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**
MSH2 Asn^{127}Ser (W)+ MSH2 IVS1+9G>C(W)	70 (82.4)	87 (75.6)	Reference			70 (82.4)	28 (65.1)	Reference		
{MSH2 Asn^{127}Ser (H) + MSH2 IVS1+9G>C (H)}+{MSH2 Asn^{127}Ser (M)+MSH2 IVS1+9G>C (M)}	15 (17.6)	28 (24.3)	1.60 (0.78-3.28)	0.19		15 (17.6)	15 (34.9)	3.66 (1.37-9.80)	**0.009**	0.011
	OVERALL					**ADCC**				
	Controls (387) N(%)	**Cases (277) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**	**Controls (387) N(%)**	**Cases (119) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**
MSH2 Asn^{127}Ser (W)+ MSH2 T>C/-6 (W)	385 (99.5)	258 (93.1)	Reference			385 (99.5)	108 (90.8)	Reference		
{MSH2 Asn^{127}Ser (H) + MSH2 T>C/-6 (H)}+{MSH2 Asn^{127}Ser (M)+MSH2 T>C/-6 (M)}	2 (0.5)	19 (6.9)	15.2 (3.49-66.6)	**0.0003**	0.0008	2 (0.5)	11 (9.2)	25.0 (5.25-119.0)	**0.0001**	0.0003
	SQCC					**SCLC**				
	Controls (387) N(%)	**Cases (122) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**	**Controls (387) N(%)**	**Cases (32) N(%)**	**AOR (95% CI)** ^{a}	$\boldsymbol{p^b}$	**FDR p-value**
MSH2 Asn^{127}Ser (W)+ MSH2 T>C/-6 (W)	385 (99.5)	114 (93.4)	Reference			385 (99.5)	32 (100)	Reference		
{MSH2 Asn^{127}Ser (H) + MSH2 T>C/-6(H)}+{MSH2 Asn^{127}Ser (M)+MSH2 T>C/-6 (M)}	2 (0.5)	8 (6.6)	13.3 (2.78-63.9)	**0.001**	0.0024	2 (0.5)	0 (0)	0	0.99	

In case of {*Asn*^{127}*Ser* (H) + *IVS1+9G>C*(H)}+{*Asn*^{127}*Ser* (M)+ *IVS1+9G>C*(M)} combinatorial analysis with {*Asn*^{127}*Ser* (W)+ *IVS1+9G>C*(W)} as reference, a 3-fold increase in risk associated with lung cancer development was reported for ADCC lung cancer subjects (AOR=3.66; 95%CI=1.37-9.80; **p=0.009**; Fdr p=0.011). Further we have also investigated {*Asn*^{127}*Ser* (H) + *T>C/-6*(H)}+{*Asn*^{127}*Ser* (M)+ *T>C/-6* (M)} taking *MSH2* {*Asn*^{127}*Ser* (W)+ *T>C/-6* (W)}as reference and reported a 15-fold increase in propensity of developing lung cancer (AOR=15.2; 95%CI=3.49-66.6; **p=0.0003**; Fdr p=0.0008), 25-fold increase for subjects having ADCC (AOR=25.0; 95%CI=5.25-119.0; **p=0.0001**; Fdr p=0.0003) and 13-fold increase for patients having SQCC (AOR=13.3; 95%CI=2.78-63.9; **p=0.001**; Fdr p=0.0024) (**Table 5.23**).

5.8.6 Haplotypes and Linkage Disequilibrium analysis of *MSH2* polymorphism and their association with lung cancer susceptibility

The six polymorphisms of the MSH2 gene were present on the same gene; hence to find any genetic association of SNPs, linkage disequilibrium, haplotype analysis was done. The frequencies of different haplotypes are represented in **Table 5.24, Table 5.25**. In both cases and controls, frequencies less than 0.03 are excluded from the analysis and not considered for their relationship with lung cancer. Global chi-square value was calculated ($\chi2$=162.10, df =7, and p=1.16e-031) to compare the frequencies of cases and controls. The Hap block 2 (C G G A G T) showed the least frequency for cases whereas, and importantly, this haplotype was significantly related with reduced likelihood of lung cancer (OR=0.078; 95%CI=0.029-0.208; p=7.94e-011). After applying Bonferroni correction, this P value was statistically noteworthy (Pcorr=4.76e-010). Hap block 3 (C G T A G T) showed the least frequency for controls than other blocks. Hap block 5 (having mutant allele for 2063T>G and wild allele for rest five SNPs) exhibited a 2-fold (OR= 1.94; 95%CI=1.37-2.73; p=0.0001; Pcorr=0.0006) increased risk of developing lung cancer. The hap block 6 showcased a high frequency concerning the control group and was strongly associated with a four-fold risk towards lung cancer (OR= 3.89; 95%CI=2.60-5.83; p=2.50e-012) (**Table 5.24**). Further, we have also analyzed linkage disequilibrium (D') and r2 values for all six polymorphic sites of the MSH2 gene using the SHEsis program (**Table 5.25**). The LD analysis showed that two out of five SNPs were LD (D' > 0.80). A firm D' was observed for 1032G>A/Asn127Ser (D'=0.990 and r2= 0.006) and for IVS1+9G>C/T>C/-6 (D'=0.870 and r2 = 0.008), but other LDs composed of other polymorphisms were relatively weak (**Figure 5.22**).

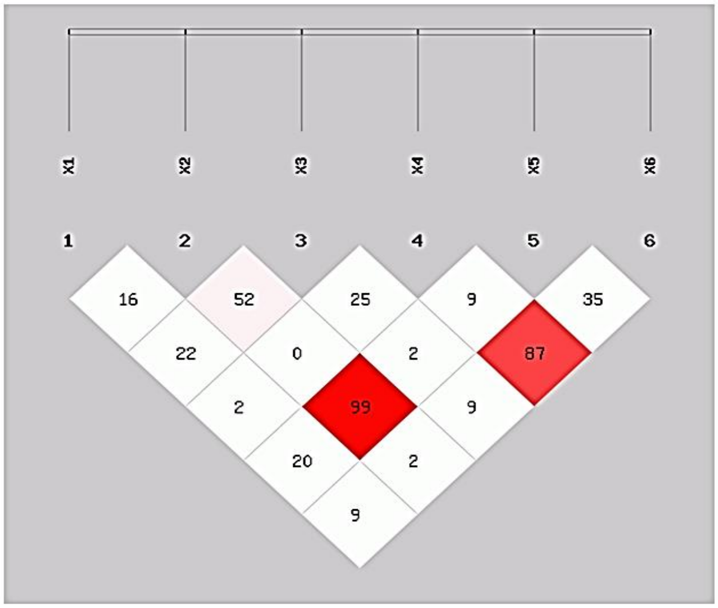

Figure 5.22: Pairwise linkage disequilibrium (D') between the six variants of MSH2 gene, red color shows the maximum LD, and white depicts the minimum LD between the polymorphic sites (X1:118T>C; X2: 1032G>A; X3: 2063T>G; X4: $Asn^{127}Ser$; X5: IVS1+9G>C; X6: T>C/-6)

Table 5.24: Haplotype frequencies and lung cancer risk associated with *MSH2* haplotypes

Haplotype	Haplotype construction (*118T>C, 1032G>A , 2063T>G , $Asn^{127}Ser$, IVS1+9G>C* , T>C/-6)	Case frequency (n (%))	Control frequency (n (%))	OR (95% CI)[a]	p [b]	P_{corr}
HAP1	C G G A C T	9.66(0.010)	31.36(0.031)	0.337 (0.163-0.698)	0.002	0.0012
HAP2	C G G A G T	4.33(0.004)	58.50(0.058)	0.078 (0.029-0.208)	7.94e-011	4.76e-010
HAP3	C G T A G T	55.04(0.055)	31.08(0.031)	2.058 (1.311-3.229)	0.001	0.01
HAP4	T G G A C T	57.10(0.057)	159.25(0.159)	0.355(0.258-0.489)	6.13e-011	3.6e-010
HAP5	T G G A G T	94.46(0.094)	57.83(0.058)	1.942 (1.379-2.736)	0.0001	0.0006
HAP6	T G T A C C	104.07(0.104)	33.17(0.033)	3.895 (2.602-5.831)	2.50e-012	1.50e-010
HAP7	T G T A C T	254.88(0.255)	321.72(0.322)	0.835 (0.682-1.022)	0.079	0.474
HAP8	T G T A G T	214.08(0.214)	196.35(0.196)	1.304 (1.043-1.629)	0.019	0.011

OR: odds ratios; CI: confidence interval. [a] logistic regression analysis calculated the odds ratios, 95% CI, and corresponding p values. [b]Two-sided $\chi 2$ test for either haplotypic distribution or frequencies between the cases and controls. P_{corr} value after applying Bonferroni correction

Table 5.25: Showing pairwise *D'* (r2) values for the six polymorphic sites among cases and controls

MSH2 SNP	1032G>A	2063T>G	Asn127Ser	IVS1+9G>C	T>C
118T>C	0.168 (0.001)	0.226 (0.024)	0.022 (0.000)	0.201 (0.011)	0.092 (0.008)
1032G>A	-	0.529 (0.001)	0.006 (0.000)	0.990 (0.006)	0.027 (0.000)
2063T>G	-	-	0.253 (0.002)	0.029 (0.001)	0.095 (0.001)
Asn127Ser	-	-	-	0.092 (0.000)	0.870 (0.008)
IVS1+9G>C	-	-	-	-	0.360 (0.017)

5.8.7 Association of high-order interactions among genetic variants with lung cancer risk by CART

Classification and regression tree analysis, a data mining tool, was used to understand high order interaction among genetic variants affecting lung cancer predisposition. In our CART analysis it resulted in a decision tree having eleven terminal nodes, the root node *MSH2 IVS1+9G>C* (M)/*T>C/-6* (M) was considered as the parent node, and hence first split was performed here, which indicates it to be the strongest risk factor for lung cancer. It was considered a reference since Terminal node 4 had the lowest case rate (17.6%). The node 8 (*MSH2 118T>C* (W)/ *IVS1+9G>C* (W)/ *Asn^{127}Ser* (W)/ *T>C/-6*(W) showed 9- fold increased risk towards lung cancer (OR=9.0; 95%CI=21.96-3.68; **p**<**0.00001)** (**Table 5.26**). Node 9 (*MSH2* 2063 T>G (W)/ *118T>C* (M)/ *IVS1+9G>C* (W)/ *Asn^{127}Ser* (W)/ *T>C/-6*(W)), node 2 (*MSH2 118T>C* (W)/*IVS1+9G>C* (W)/*T>C/-6* (M)), node 3 (*MSH2 2063T>G* (W)/*118T>C* (M)/ *IVS1+9G>C* (W)/*T>C/-6* (M)) and node 1 (*MSH2 IVS1+9G>C* (M)/ *T>C/-6* (M)) also showed 3.5-fold (OR=3.48; 95%CI=7.90-1.53; **p**=**0.002**), 6.5-fold (OR=6.53; 95%CI=19.96-2.13; **p**=**0.0009**), 2-fold (OR=2.43; 95%CI=7.78-0.76; **p**<**0.00001**), and 4-fold (OR=4.12; 95%CI=5.76-2.95; **p**<**0.00001**) increased risk towards lung cancer. Further, individuals with the genotype of nodes 4, 5, and 6 did not show any association with lung cancer (**Table 5.26**).

Table 5.26: Overall, risk estimates based on classification and regression tree analysis (MSH2 variants).

Terminal Node	**Genotype for each terminal node**	**Control (n)**	**Case (n)**	**Case rate#**	**OR (95% CI)**	**p-value$**
Node 1	*MSH2 IVS 1+9G>C (M)/ T>C/-6 (M)*	68	161	70.3	4.12 (5.76-2.95)	<0.00001
Node 2	*MSH2 118T>C (W)/IVS 1+9G>C (W)/T>C/-6 (M)*	4	15	78.9	6.53 (19.9-2.13)	0.0009
Node 3	*MSH2 2063 T>G (W)/118T>C (M)/ IVS 1+9G>C (W)/T>C/-6 (M)*	5	7	58.3	2.43 (7.78-0.76)	0.132
Node 4	*MSH2 2063 T>G (M)/ 118T>C (M)/ IVS 1+9G>C (W)/ T>C/-6 (M)*	14	3	17.6	0.37 (1.31-0.10)	0.125
Node 5	*MSH2 118T>C (M)/ Asn^{127}Ser (M)/ T>C /-6(W)*	0	20	100	-	-
Node 6	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ Asn^{127}Ser (M)/ T>C /-6(W)*	0	14	100	-	-
Node 7	*MSH2 IVS 1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (M)/ T>C /-6(W)*	24	22	47.8	1.59 (2.92-0.87)	0.129
Node 8	*MSH2 118T>C (W)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ T>C /-6(W)*	6	31	83.8	9.0 (21.9-3.68)	<0.00001
Node 9	*MSH2 2063 T>G (W)/ 118T>C (M)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ T>C /-6(W)*	9	18	66.7	3.48 (7.90-1.53)	0.002
Node 10	*MSH2 2063 T>G (M)/ 118T>C (M)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ T>C /-6(W)*	32	15	31.9	0.81 (1.54-0.43)	0.534

The #Case rate is the percentage of LC cases among all individuals in each terminal node case/ (case+ control) ×100.
$p<0.05 is considered statistically significant.
Note: The bold values in the table are highly significant and are a positive indicator for any association or not.

5.8.7.1 Association of high–order interactions among genetic variants with lung cancer risk by CART according to histology (ADCC)

We further stratified samples on the basis of histology, for the subjects having ADCC; the first split was observed at *MSH2 T>C/-6* making this polymorphism as the parent node and most important risk factor for ADCC development. The parent node further showed a 43-fold increased risk towards ADCC (OR=43.5; 95%CI=222.1-8.50; **p=0.00001**). Node 3 (*MSH2 IVS1+9G>C (M)/ 2063 T>G (W)/ 118T>C (M)/Asn^{127}Ser (W)/ T>C/-6 (M)*) and 14 (*MSH2 IVS1+9G>C (W)/ 118T>C (W)/ 2063 T>G (M)/ Asn^{127}Ser (W) / T>C/-6 (W)*) showed 23-fold and 14-fold increased risk towards developing ADCC (OR=23.7; 95%CI=132.3-4.25; **p=0.0003** and OR=13.84; 95%CI=54.9-3.48; **p=0.0001**). Node 11 (*MSH2 IVS1+9G>C (W)/ 2063T>G (W)/ Asn^{127}Ser (W) / T>C-6 (W)*) and node 2 (*MSH2 118T>C (W)/ Asn^{127}Ser (W)/ T>C/-6 (M)*) exhibited a 5.5 and 7-fold increased risk towards development of ADCC (OR=5.75; 95%CI=17.3-1.90; **p=0.0019** and OR=7.19; 95%CI=14.9-3.44; **p<0.00001**). Node 12 (*MSH2 118T>C (M)/ IVS1+9G>C (M)/ 2063T>G (W)/ Asn^{127}Ser (W) / T>C/-6 (W)*) and 13 (*MSH2*

118T>C (W)/ IVS1+9G>C (M)/ 2063T>G (W)/ Asn^{127}*Ser (W) / T>C/-6 (W)*) also showed increased risk towards ADCC (OR=3.95; 95%CI=10.58-1.47; **p=0.0062** and OR=2.42; 95%CI=5.0-1.17; **p=0.016**) (**Table 5.27**). We did not find any association of remaining node with risk of developing ADCC.

5.8.7.2 Association of high–order interactions among genetic variants with lung cancer risk by CART according to histology (SCLC)

The CART analysis from SCLC showed the first split was observed at *MSH2 T>C/-6* making this genetic variant as most important risk factor for development of SCLC. Parent Node reported to pose highest risk for SCLC development (OR=27.2; 95%CI=57.4-12.8; **p**<**0.00001**). Node 13 (*MSH2 118T>C (W)/ IVS1+9G>C (W)/ Asn*^{127}*Ser (W)/ T>C/-6 (W)*) and 4 (*MSH2 IVS1+9G>C (W)/ 118T>C (W)/ T>C/-6(M)*) exhibited a 11.7-fold and 7-fold increased risk towards SCLC development (OR=11.73; 95%CI=41.2-3.33; **p=0.0001** and OR=7.04; 95%CI=40.4-1.22; **p=0.02**). Along the lines of previous results, node 2 (*MSH2 2063T>G (W)/ IVS1+9G>C (W)/ 118T>C (M)/ T>C/-6 (M)*) and 5 (*MSH2 Asn*^{127}*Ser (W)/ 2063T>G (W)/ IVS 1+9G>C (M)/118T>C (W)/ T>C/-6(M)*) also reported to have 5.5-fold and 3-fold increased risk towards development of SCLC (OR=5.63; 95%CI=30.5-1.03; **p=0.04** and OR=3.79; 95%CI=9.62-1.49; **p=0.005**) (**Table 5.28**).No association of the remaining nodes was observed with risk of developing SCLC in this study.

Table 5.27: Risk estimates based on classification and regression tree analysis (*MSH2* variants) based on histology (ADCC)

ADCC						
Terminal Node	**Genotype for each terminal node**	**Control (n)**	**Case (n)**	**Case rate[#]**	**OR (95% CI)**	**p-value[$]**
Node 1	*MSH2 Asn^{127}Ser (M)/ T>C/-6 (M)*	2	11	84.6	43.3 (222.4-8.50)	0.00001
Node 2	*MSH2 118T>C (W)/ Asn^{127}Ser (W)/ T>C/-6 (M)*	55	50	47.6	7.19 (14.9-3.44)	<0.0000
Node 3	*MSH2 IVS 1+9G>C (M)/ 2063 T>G (W)/ 118T>C (M)/Asn^{127}Ser (W)/ T>C/-6 (M)*	2	6	75.0	23.7 (132.3-4.25)	<0.0003
Node 4	*MSH2 IVS 1+9G>C (W)/ 2063 T>G (W)/118T>C (M)/ (M)/Asn^{127}Ser (W)/ T>C/-6 (M)*	5	1	16.7	1.58 (14.8-0.16)	0.687
Node 5	*MSH2 2063 T>G (M)/ 118T>C (M)/ Asn^{127}Ser (W)/ T>C/-6 (M)*	27	6	18.2	1.75 (5.19-0.59)	0.308
Node 6	*MSH2 118 T>C (M)/ Asn^{127}Ser (M)/ T>C-6 (W)*	0	8	100	-	0.979
Node 7	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ Asn^{127}Ser (M)/ T>C-6 (W)*	0	4	100	-	0.985
Node 8	*MSH2 2063T>G (M)/1032 G>A (W)/ IVS 1+9G>C (M)/118T>C (W)/ Asn^{127}Ser (M)/ T>C-6 (W)*	6	3	33.3	3.95 (18.1-0.86)	0.076
Node 9	*MSH2 2063 T>G (W)/ 1032 G>A (M)/ IVS 1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (M)/ T>C-6 (W)*	16	6	27.3	2.96 (9.16-0.95)	0.059
Node 10	*MSH2 1032 G>A (M)/ IVS 1+9G>C (M)/118T>C (W)/ Asn^{127}Ser (M)/ T>C-6 (W)*	2	0	0	-	0.992
Node 11	*MSH2 IVS 1+9G>C (W)/ 2063 T>G (W)/ Asn^{127}Ser (W) / T>C-6 (W)*	11	8	42.1	5.75 (17.3-1.90)	0.001
Node 12	*MSH2 118T>C (M)/ IVS 1+9G>C (M)/ 2063 T>G (W)/ Asn^{127}Ser (W) / T>C/-6 (W)*	20	10	33.3	3.95 (10.5-1.47)	0.006
Node 13	*MSH2 118T>C (W)/ IVS 1+9G>C (M)/ 2063 T>G (W)/ Asn^{127}Ser (W) / T>C/-6 (W)*	127	39	23.5	2.42 (5.0-1.17)	0.016
Node 14	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ 2063 T>G (M)/ Asn^{127}Ser (W) / T>C/-6 (W)*	4	7	63.6	13.8 (54.9-3.48)	0.0001
Node 15	*MSH2 IVS 1+9G>C (M)/ 118T>C (W)/ 2063 T>G (W)/ Asn^{127}Ser (W) / T>C/-6 (W)*	136	32	19.0	1.86 (3.88-0.89)	0.098
Node 16	*MSH2 1032 G>A (M)/ 118T>C (M)/ 2063 T>G (W)/ Asn^{127}Ser (W) / T>C/-6 (W)*	0	1	100	-	0.992

The #Case rate is the percentage of LC cases among all individuals in each terminal node case/ (case+ control) ×100.
$p<0.05 is considered statistically significant.
Note: The bold values in the table are highly significant and is a positive indicator for any association or not

Table 5.28: Risk estimates based on classification and regression tree analysis (*MSH2* variants) based on histology (SCLC)

SCLC						
Terminal Node	**Genotype for each terminal node**	**Control (n)**	**Case (n)**	**Case rate#**	**OR (95% CI)**	**p-value$**
Node 1	*MSH2 IVS 1+9G>C (M)/ 118T>C (M)/ T>C/-6 (M)*	15	29	65.9	27.2 (57.4-12.8)	<0.00001
Node 2	*MSH2 2063T>G (W)/ IVS 1+9G>C (W)/ 118T>C (M)/ T>C/-6 (M)*	5	2	28.6	5.63 (30.5-1.03)	0.045
Node 3	*MSH2 2063T>G (M)/ IVS 1+9G>C (W)/ 118T>C (M)/ T>C/-6(M)*	14	1	6.7	1.00 (7.97-0.12)	0.995
Node 4	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ T>C/-6(M)*	4	2	33.3	7.04 (40.4-1.22)	0.028
Node 5	*MSH2 Asn^{127}Ser (W)/ 2063T>G (W)/ IVS 1+9G>C (M)/118T>C (W)/ T>C/-6(M)*	26	7	21.2	3.79 (9.62-1.49)	0.005
Node 6	*MSH2 Asn^{127}Ser (M)/ 2063T>G (W)/ IVS 1+9G>C (M)/ 118T>C (W)/ T>C/-6(M)*	2	0	0	-	0.993
Node 7	*MSH2 2063T>G (W)/ IVS 1+9G>C (M)/ 118T>C (W)/ T>C/-6(M)*	25	3	10.7	1.69 (6.0-0.47)	0.416
Node 8	*MSH2 118T>C (M)/ Asn^{127}Ser (M)/ T>C/-6 (W)*	0	3	100	-	0.986
Node 9	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ Asn^{127}Ser (M) / T>C/-6 (W)*	0	2	100	-	0.989
Node 10	*MSH2 1032G>A (W)/ 2063T>G (W)/ IVS 1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (M) / T>C/-6 (W)*	16	3	15.8	2.64 (9.69-0.71)	0.143
Node 11	*MSH2 1032 G>A (M)/ 2063T>G (W)/ IVS 1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (M) / T>C/-6 (W)*	1	0	0	-	0.995
Node 12	*MSH2 2063T>G (M)/ IVS 1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (M) / T>C/-6 (W)*	7	0	0	-	0.985
Node 13	*MSH2 118T>C (W)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ T>C/-6 (W)*	6	5	45.5	11.7 (41.2-3.33)	0.0001
Node 14	*MSH2 118T>C (M)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ T>C/-6 (W)*	41	3	6.8	1.03 (3.57-0.29)	0.962

The #Case rate is the percentage of LC cases among all individuals in each terminal node case/ (case+ control) ×100.
$p<0.05 is considered statistically significant.
Note: The bold values in the table are highly significant and are a positive indicator for any association or not.

5.8.7.3 Association of high–order interactions among genetic variants with lung cancer risk by CART according to histology (SQCC)

To observe the association of *MSH2* polymorphic variant with the risk of developing SQCC, the first split was observed at *MSH2 Asn^{127}Ser* rendering this polymorphism as the strongest risk factor for SQCC. Node 7 (*MSH2 118T>C (W)/ IVS1+9G>C (W)/T>C/-6 (W)/ Asn^{127}Ser (W)*) exhibited a 11-fold increased risk towards development of SQCC (OR=11.2; 95%CI=29.7-4.27; **p<0.00001**). Further node 1 (*MSH2 Asn^{127}Ser (M)), 5 (MSH2 118T>C (W)/ IVS 1+9G>C (W)/ 2063T>G (M)/ T>C/-6 (M)/ Asn^{127}Ser (W)*) and 8 (*MSH2 2063T>G (W)/ 118T>C (M)/*

IVS1+9G>C (W)/ T>C/-6 (W)/ Asn^{127}Ser (W)) showed a 5-fold increased risk towards risk of developing SQCC (OR=5.52; 95%CI=9.72-3.13; p<0.00001, OR=5.63; 95%CI=25.6-1.23; p=0.02 and OR=5.16; 95%CI=12.8-2.07; p=**0.0004**). Node 2 (*MSH2 2063T>G (W)/ T>C/-6 (M)/ Asn^{127}Ser (W)*) and 3 (*MSH2 118T>C (M)/ IVS1+9G>C (M)/ 2063T>G (M)/ T>C/-6 (M)/ Asn^{127}Ser (W)*) also showed similar increased risk towards development of SQCC (3-fold) (OR=3.97; 95%CI=6.82-2.31; p<0.00001 and OR=3.90; 95%CI=8.86-1.71; **p=0.001**). We did not find any association of remaining nodes with the risk of developing SQCC (**Table 5.29**).

Table 5.29: Risk estimates based on classification and regression tree analysis (*MSH2* variants) based on histology (SQCC)

SQCC						
Terminal Node	**Genotype for each terminal node**	**Control (n)**	**Case (n)**	**Case rate#**	**OR (95% CI)**	**p-value$**
Node 1	*MSH2 Asn^{127}Ser (M)*	26	34	56.7	5.52 (9.72-3.13)	<0.00001
Node 2	*MSH2 2063T>G (W)/ T>C/-6 (M)/ Asn^{127}Ser (W)*	34	32	48.5	3.97 (6.82-2.31)	<0.00001
Node 3	*MSH2 118T>C (M)/ IVS 1+9G>C (M)/ 2063T>G (M)/ T>C/-6 (M)/ Asn^{127}Ser (W)*	13	12	48.0	3.90 (8.86-1.71)	0.001
Node 4	*MSH2 118T>C (W)/ IVS 1+9G>C (M)/ 2063T>G (M)/ T>C/-6 (M)/ Asn^{127}Ser (W)*	25	9	26.5	1.52 (3.38-0.68)	0.304
Node 5	*MSH2 118T>C (W)/ IVS 1+9G>C (W)/ 2063T>G (M)/ T>C/-6 (M)/ Asn^{127}Ser (W)*	3	4	57.1	5.63 (25.6-1.23)	0.025
Node 6	*MSH2 118T>C (W)/ IVS 1+9G>C (W)/ 2063T>G (M)/ T>C/-6 (M)/ Asn^{127}Ser (W)*	14	0	0	-	
Node 7	*MSH2 118T>C (W)/ IVS 1+9G>C (W)/T>C/-6 (W)/ Asn^{127}Ser (W)*	6	16	72.7	11.2 (29.7 -4.27)	<0.00001
Node 8	*MSH2 2063T>G (W)/ 118T>C (M)/ IVS 1+9G>C (W)/ T>C/-6 (W)/ Asn^{127}Ser (W)*	9	11	55.0	5.16 (12.8-2.07)	0.0004
Node 9	*MSH2 2063T>G (W)/ 118T>C (M)/ IVS 1+9G>C (W)/ T>C/-6 (W)/ Asn^{127}Ser (W)*	32	7	17.9	0.92 (2.16-0.39)	0.856

The #Case rate is the percentage of LC cases among all individuals in each terminal node case/ (case+ control) ×100.
$p<0.05 is considered statistically significant.
Note: The bold values in the table are highly significant and are a positive indicator for any association or not.

5.8.7.4 Association of high–order interactions among genetic variants with lung cancer risk by CART according to smoking status

The high order interaction by CART was further applied on smokers and non-smokers. In case of smokers, the first split was observed at *MSH2 T>C/-6* making it the strongest risk factor for developing lung cancer. Node 1 (*MSH2 2063T>G (W)/T>C/-6(M)*) and 14 (*MSH2 IVS1+9G>C (W)/ 118T>C (W)/ Asn^{127}Ser (W)/ 2063T>G (M)/ T>C/-6 (W)*) reported a 17 and 20-fold increased risk of developing smoking induced lung cancer (OR=17.2; 95%CI=34.2-8.67; **p**<**0.00001** and OR=20.4; 95%CI=68.3-6.17; **p**<**0.00001**). Further node 3 (*MSH2 118T>C (M)/*

IVS 1+9G>C (M)/Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)), node 5 (*MSH2 118T>C (W)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)*) expressed 13, 12 and 13-fold increased risk towards development of smoking induced lung cancer (OR=13.01; 95%CI=29.3-5.76; p=0.00000, OR=12.8; 95%CI=53.3-3.11; p=**0.0004 and** OR=13.6; 95%CI=32.0-5.79; **p**<**0.00001**). Similarly node 4 (*MSH2 118T>C (W)/ IVS1+9G>C (M)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)*), 9 (*MSH2 Asn^{127}Ser (W)/ 118T>C (M)/ IVS1+9G>C (M)/ 2063T>G (W)/ T>C/-6 (W)*), 10 (*MSH2 118T>C (M)/ IVS1+9G>C (M)/ 2063T>G (W)/ T>C/-6 (W)*) and 15 (*MSH2 IVS1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (W)/ 2063T>G (M)/ T>C/-6 (W)*) also exhibited increased risk towards developing smoking induced lung cancer (OR=3.62; 95%CI=8.02-1.63; p=**0.001,** OR=3.14; 95%CI=7.44-1.32**; p=0.009,** OR=3.75; 95%CI=6.70-2.10; **p=0.00001** and OR=2.02; 95%CI=3.75-1.13; p=**0.017**) (**Table 5.30**). When we further analyze non-smokers, first split was observed at *MSH2 2063T>G* (M) making it the strong risk factor for developing lung cancer in non-smokers. Only one node, the parent node (*MSH2 2063T>G* (M)), showed a 13-fold increased risk towards developing lung cancer (OR=13.3; 95%CI=35.2-5.03; p<0.00001) (**Table 5.31**).

Table 5.30: Risk estimates based on classification and regression tree analysis (*MSH2* variants) based on smoking status

SMOKERS						
Terminal Node	**Genotype for each terminal node**	**Control (n)**	**Case (n)**	**Case rate[#]**	**OR (95% CI)**	**p-value[$]**
Node 1	*MSH2 2063T>G (W)/T>C/-6(M)*	23	82	78.1	17.2 (34.2-8.67)	<0.00001
Node 2	*MSH2 Asn^{127}Ser (M)/ 2063 T>G (M)/ T>C/-6 (M)*	0	6	100	-	-
Node 3	*MSH2 118T>C (M)/ IVS 1+9G>C (M)/Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)*	13	35	72.9	13.0 (29.3-5.76)	<0.0000
Node 4	*MSH2 118T>C (W)/ IVS 1+9G>C (M)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)*	24	18	42.9	3.62< (8.02-1.63)	0.0014
Node 5	*MSH2 118T>C (W)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)*	3	8	72.7	12.8 (53.3-3.11)	0.0004
Node 6	*MSH2 118T>C (M)/ IVS 1+9G>C (W)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (M)*	14	3	17.6	1.03 (3.98-0.26)	0.952
Node 7	*MSH2 IVS 1+9G>C (W)/ 2063 T>G (W)/ T>C/-6 (W)*	11	31	73.8	13.6 (32.0-5.79)	0.0000
Node 8	*MSH2 Asn^{127}Ser (M)/118T>C (M)/ IVS 1+9G>C (M)/ 2063 T>G (W)/ T>C/-6 (W)*	0	10	100	-	-
Node 9	*MSH2 Asn^{127}Ser (W)/ 118T>C (M)/ IVS 1+9G>C (M)/ 2063 T>G (W)/ T>C/-6 (W)*	20	13	39.4	3.14 (7.44-1.32)	0.0093
Node 10	*MSH2 118T>C (M)/ IVS 1+9G>C (M)/ 2063 T>G (W)/ T>C/-6 (W)*	112	87	43.7	3.75 (6.70-2.10)	0.0001
Node 11	*MSH2 IVS 1+9G>C (W)/ Asn^{127}Ser (M)/ 2063 T>G (M)*	0	6	100	-	-
Node 12	*MSH2 118T>C (M)/ IVS 1+9G>C (M)/ Asn^{127}Ser (M))/ 2063 T>G (M)/ T>C/-6 (W)*	0	3	100	-	-
Node 13	*MSH2 118T>C (W)/ IVS 1+9G>C (M)/ Asn^{127}Ser (M))/ 2063 T>G (M)/ T>C/-6 (W)*	7	5	41.7	3.45 (12.1-0.98)	0.052
Node 14	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (W)*	4	17	81.0	20.5 (68.3-6.17)	0.0000
Node 15	*MSH2 IVS 1+9G>C (M)/ 118T>C (W)/ Asn^{127}Ser (W)/ 2063 T>G (M)/ T>C/-6 (W)*	131	56	29.9	2.06 (3.75-1.13)	0.0170

The #Case rate is the percentage of LC cases among all individuals in each terminal node case/ (case+ control) ×100.
$p<0.05 is considered statistically significant.
Note: The bold values in the table are highly significant and are a positive indicator for any association or not.

Table 5.31: Risk estimates based on classification and regression tree analysis (*MSH2* variants) based on smoking status

NON-SMOKERS						
Terminal Node	**Genotype for each terminal node**	**Control (n)**	**Case (n)**	**Case rate[#]**	**OR (95% CI)**	**p-value[$]**
Node 1	*MSH2 2063T>G (M)*	6	48	88.9	13.3 (35.2-5.03)	0.0000
Node 2	*MSH2 Asn^{127}Ser (M)/ 2063T>G (W)*	0	15	100	-	-
Node 3	*MSH2 118T>C (M)/ Asn^{127}Ser (W)/ 2063T>G (W)*	0	9	100	-	-
Node 4	*MSH2 IVS 1+9G>C (W)/ 118T>C (W)/ Asn^{127}Ser (W)/ 2063T>G (W)*	0	3	100	-	-

The #Case rate is the percentage of LC cases among all individuals in each terminal node case/ (case+ control) ×100.
$p<0.05 is considered statistically significant.
Note:The bold values in the table are highly significant and are a positive indicator for any association or not

5.9 Survival analysis of *MLH1* gene

5.9.1 Association of *MLH1* genetic variants with overall survival of lung cancer patients based on the chemotherapeutic regimen

Out of 500 lung cancer patients almost, 354 patients were given chemotherapeutic drugs, so we wanted to estimate if *MLH1 -93G>A* polymorphism exerts any modulating effect and if overall survival is associated with different chemotherapeutic drugs used in the treatment. All patients received first-line chemotherapeutic drugs *viz.* carboplatin/cisplatin and the second line of chemotherapy drugs, including paclitaxel, pemetrexed, irinotecan, and docetaxel. 105 (29.66%) were given cisplatin/carboplatin + Pemetrexed, 82 (23.16%) were provided with cisplatin/carboplatin + Paclitaxel, cisplatin/carboplatin + Irinotecan was administered to 45 (12.71%) patients, 37 (10.45%) were given cisplatin/carboplatin + Docetaxel, 29 (8.19%) were given cisplatin/carboplatin + Gemcitabine whereas 48 (13.55%) patients were given other chemotherapeutic drugs. When data from patients receiving docetaxel was analyzed, a 3-fold increase in H.R. and significantly reduced MST (8.4 vs. 13.33 months, H.R.: 3.13, 95% CI= 0.90-10.9, **p=0.04**) for both heterozygous type (*GA*) and combination of heterozygous and mutant type genotype (*GA+AA*) (**Figure 5.23**) was observed (**Table 5.32**)

Table 5.32: Association of *MLH1* genotype on survival in lung cancer cases based on the chemotherapeutic regimen

Docetaxel								
***rs1800734 -93G>A* (*MLH1*)**	**CASES n(%), N=36**	**DEAD n(%), N=28**	**ALIVE n(%), N=8**	**Univariate analysis**			**Multivariate analysis**	
				MST (months)	**Log-rank *p***	**Unadjusted HR[a]**	**Adjusted HR[b] (95% CI)**	***P***
GG	9 (25.0)	5 (17.8)	4 (50.0)	13.33				
GA	24 (66.6)	21 (75.0)	3 (37.5)	8.4	0.072	2.32 (1.04-5.18)	3.13 (0.90-10.9)	**0.04**
AA	3 (8.3)	2 (7.14)	1 (12.5)	14.56	0.983	0.98 (0.19-5.03)	-	
GA+AA	27 (75.0)	23 (82.1)	4 (50.0)	8.56	0.120	2.08 (0.93-4.63)	3.15(0.90-11.0)	**0.04**

[a]Unadjusted Hazards ratio for Kaplan Meier analysis, [b]hazards ratio for age, sex, smoking, histology, stage, KPS, ECOG

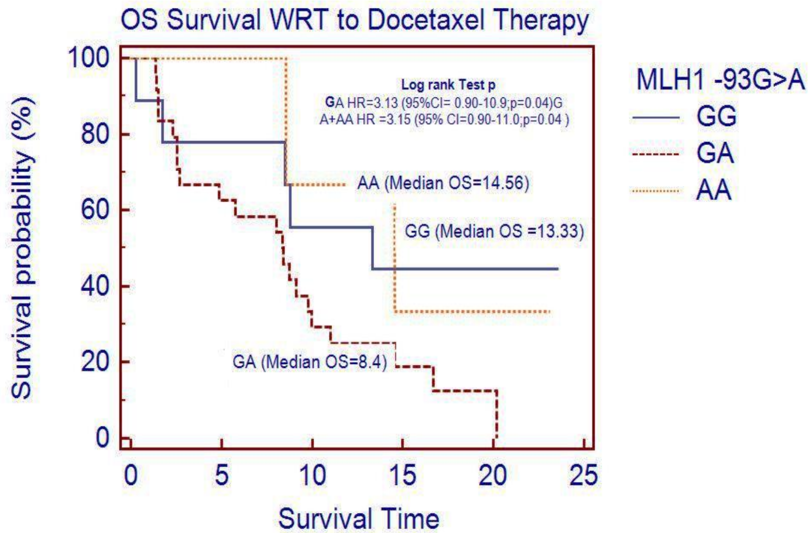

Figure 5.23: OS Survival WRT to Docetaxel Therapy

5.10 Survival analysis of *MSH6 557G>T* genetic polymorphism

5.10.1 Association of *MSH6 557G>T* genetic variants with overall survival of lung cancer patients

Survival analysis and association with *MSH6 557G>T* polymorphism of 475 lung cancer cases is shown in **Table 5.33**. Univariate analysis was performed by the Kaplan-Meier method, and multivariate analysis was evaluated by the Cox-regression method for analyzing any association between *MSH6 557G>T*, polymorphism, and survival of lung cancer patients (**Table 5.33**). Out of 475 lung cancer patients, 356 (74.9%) were dead at the end of the study, whereas 119 (25.1%) were alive. For *MSH6 557G>T* polymorphism, we reported a significant reduction in MST for candidates harboring heterozygous type genotype (*GT*) and a corresponding increase in unadjusted and adjusted HR (8.66 vs. 5.93 months; HR=1.35; 95%CI=0.95-1.87; log rank **p=0.03**).

Table 5.33: Association of *MSH6* polymorphism on overall survival in lung cancer cases

				Overall Survival				
MSH6 rs3136228 557G>T	**Cases N(%), N= 475**	**Dead N(%), N= 356**	**Alive N(%), N= 119**	**Univariate analysis**			**Multivariate analysis**	
				MST (months)	**Log rank *p***	**Unadjusted hr[a]**	**Adjusted hr[b] (95% CI)**	***P***
TT	410 (86.3)	302 (84.8)	108 (90.7)	8.66		1.00 (Reference)		
GT	65 (13.7)	54 (15.2)	11 (9.3)	5.93	**0.03**	1.35 (0.95-1.87)	1.23 (0.91-1.67)	0.17
GT+GG	65 (13.7)	54 (15.2)	11 (9.3)	5.93	0.03	1.35 (0.95-1.87)	1.23 (0.91-1.67)	0.17

[a]Unadjusted Hazards ratio for Kaplan Meier analysis, [b]hazards ratio for age, sex, smoking, histology, stage, KPS, ECOG

5.10.2 Association of *MSH6 557G>T* genetic variants with overall survival of lung cancer patients based on the chemotherapeutic regimen

All lung cancer patients selected for this study were given platinum-based doublet chemotherapy (carboplatin/ cisplatin) as first-line therapy and certain other chemotherapeutic agents that come under second-line of therapy paclitaxel, pemetrexed, irinotecan, and docetaxel. We wanted to evaluate the relationship of MSH6 557G>T polymorphism towards the overall survival of subjects treated with different chemotherapeutic regimens. The results of MSH6 polymorphism on overall survival according to the chemotherapy regimen are shown in Error! Not a valid bookmark self-reference.. Out of 36 lung cancer patients undergoing a docetaxel regimen, 28 (77.8%) were dead, and 8 (22.2%) were alive at the end of the study. We further reported a significant decrease in MST (9.13 vs. 4.9 months) and a corresponding increase in the hazard ratio (HR= 2.28; 95%CI=0.59-8.82; p=0.03) for patients harboring heterozygous type genotype (GT) (**Figure 5.24**).

Table 5.34: Association of *MSH6* polymorphism with survival in lung cancer cases based on the chemotherapeutic regimen

MSH6 rs3136228 557G>T	**CASES n(%), N=36**	**DEAD n(%), N=28**	**ALIVE n(%), N=8**	**Docetaxel Univariate analysis**			**Multivariate analysis**	
				MST (months)	**Log rank *p***	**Unadjusted HR[a]**	**Adjusted HR[b] (95% CI)**	***P***
TT	30 (83.3)	22 (78.5)	8 (100)	9.13		1.00 (Reference)		
GT	6 (16.7)	6 (21.5)	0	4.9	0.02	2.59 (0.73-9.15)	2.28 (0.59-8.82)	**0.03**
GT+GG	6 (16.7)	6 (21.5)	0	4.9	0.02	2.59 (0.73-9.15)	2.28 (0.59-8.82)	0.23

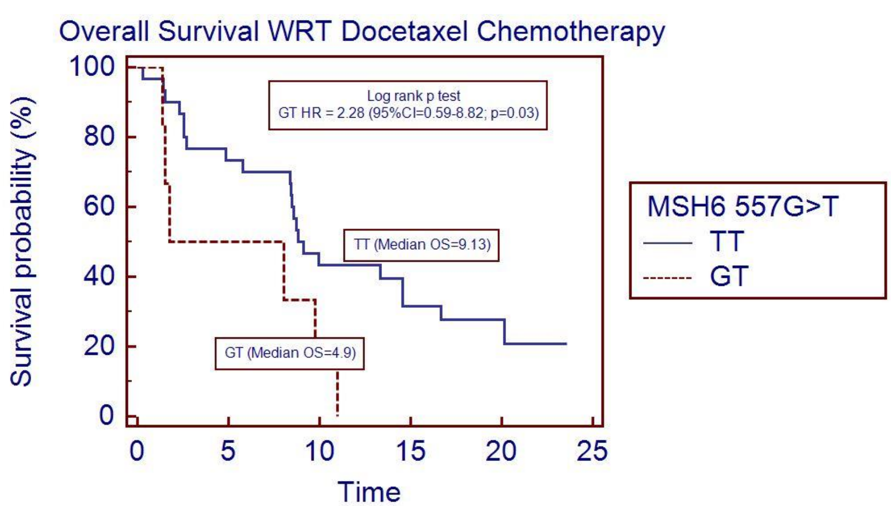

Figure 5.24: Survival of *MSH6 557G>T* WRT to Docetaxel chemotherapy

5.11 Survival Analysis for polymorphic sites of *MSH6* (*rs1800932-Pro^{92}Pro* and *rs1042821- Gly^{39}Glu*) gene

5.11.1 Association of *MSH6* (*rs1800932*) *Pro^{92}Pro* and (*rs1042821*) *Gly^{39}Glu* genetic variants with overall survival of lung cancer patients

As shown in **Table 5.35**, we stratified patients based on histology, our results showed that for *Pro^{92}Pro* (*rs1800932*) polymorphism, for SCLC patients who were carrying a single copy of the variant allele (*AG*) reported a significant increase in MST (20.6 vs. 7.6, log-rank **p=0.03**) and a favorable death ratio (HR=0.4) as compared to those patients who were carrying both the alleles for the wild genotype (*AA*). For *MSH6 Gly^{39}Glu (rs1042821*) polymorphism, we did not find any significant association with overall survival (data not shown).

Table 5.35: Association of *MSH6 A>G* polymorphism on overall survival in lung cancer cases based on histology

				SCLC				
***MSH6 rs1800932* A>G**	**Cases N(%), N= 75**	**Dead N(%), N= 64**	**Alive N(%), N= 11**	**Univariate analysis**			**Multivariate analysis**	
				MST (months)	**Log rank *p***	**Unadjusted hr^{a}**	**Adjusted hr^{b} (95% CI)**	***P***
GG	69 (92.0)	60 (93.8)	9 (81.8)	7.6				
AG	6 (8.0)	4 (6.2)	2 (18.2)	20.6	**0.03**	0.4 (0.20-0.82)	0.40 (0.12-1.30)	0.09

[a]Unadjusted Hazards ratio for Kaplan Meier analysis, [b]hazards ratio for age, sex, smoking, histology, stage, KPS,ECOG

5.12 Survival analysis of *MSH2* polymorphism

For *MSH2 IVS1+9G>C* polymorphism, the subjects with T3 and T4 stage and wild-type genotypes are 24 (28.2%) and 65 (21.5%), respectively. The subjects carrying mutant-type genotypes for T3 and T4 stages are 19 (22.4%) and 104 (34.3%), respectively (**Table 5.36**). logistic regression for *rs2303426* polymorphism revealed that mutant type (*CC*) genotype had an approximately 2-fold increased risk of developing T3 or T4 stage lung cancer (AOR=1.97; 95% CI=0.99-3.94; **p=0.03**). For the same polymorphism, mutant type genotype (*CC*) has 22 (21.6%) and 124 (33.2%) lung cancer patients belonging to NX+N0+N1 and N2+N3+N4. Further, the mutant genotype reported a 2.25-fold increased risk for lymph node invasion than patients with the wild genotype (AOR=2.25; 95%CI=1.20-4.23; **p=0.01**). The combined genotype (*GC+CC*) has 72 (70.5%) and 299 (80.1%) subjects suffering from NX+N0+N1 and N2+N3+N4 categories. The combined genotype also reported a 1.68-fold increased risk of developing lung cancer (AOR=1.68; 95%CI=1.02-2.76;

p=0.04). Moreover, mutant type genotype (*CC*) reported 59 (28.0%) of lung cancer subjects with no metastasis, and 87 (32.8%) reported metastasis. Mutant type genotype was further reported to have a 1.40-fold increased risk of metastatic lung cancer, but the results were not statistically significant (AOR=1.40; 95%CI=0.81-2.39; p=0.21). For *MSH2 1032G>A* polymorphism, the number of lung cancer patients harboring wild-type genotype and suffering from stage III and IV were 191 (98.9%) and 252 (98.1%), respectively. Further, no subject was carrying mutant type genotype (*AA*) and suffering from stage III or IV lung cancer. The heterozygous type genotype (*GA*) reported a 1.65-fold increased risk of developing lung cancer, but the results were insignificant. Moreover, the subjects carrying heterozygous type genotype were deficient; hence the results are insignificant (AOR=1.65; 95%Ci=0.29-9.23; p=0.56). For the same polymorphism, subjects harboring wild-type genotype and showing no metastasis and some degree of metastasis were 207 (98.6%) and 260 (98.1%), respectively. Further, the heterozygous type genotype (*GA*) reported a 1.19-fold increased risk of developing lung cancer, but the results are not significant due to a minimal number of patients in this category (**Table 5.36**).

Table 5.36: Relationship of *MSH2* (*rs2303426*) polymorphism and different genotypes with the clinical-pathological parameters

	Clinical stage				**Primary tumor extension**			
rs2303426/ IVS1+9G>C	**III 193 N (%)**	**IV 257 N (%)**	**AOR (95% CI)** [a]	p^b	**T3 85 N (%)**	**T4 303 N (%)**	**AOR (95% CI)** [a]	p^b
GG	49 (25.3)	50 (19.4)	Reference		24 (28.2)	65 (21.5)	Reference	
GC	91 (47.3)	121 (47.1)	1.25 (0.76-2.06)	0.37	42 (49.4)	134 (44.2)	1.15 (0.64-2.08)	0.62
CC	53 (27.4)	86 (33.5)	1.57 (0.90-2.75)	0.10	19 (22.4)	104 (34.3)	1.97 (0.99-3.94)	**0.03**
GC+CC	144 (74.6)	207(80.5)	1.35 (0.84-2.16)	0.20	61 (71.7)	238 (78.5)	1.39 (0.80-2.42)	0.23
	Lymph Node Invasion				**Metastasis**			
rs2303426/ IVS1+9G>C	**$N_x+N_0+N_1$ 102 N (%)**	**$N_2+N_3+N_4$ 373 N (%)**	**AOR (95% CI)**	p^b	**No 210 N (%)**	**Yes 265 N (%)**	**AOR (95% CI)**	p^b
GG	30 (29.4)	74 (19.9)	Reference		50 (23.8)	52 (19.6)	Reference	
GC	50 (49.0)	175 (46.9)	1.42 (0.83-2.43)	0.18	99 (47.2)	126 (47.6)	1.20 (0.74-1.96)	0.44
CC	22 (21.6)	124 (33.2)	2.25 (1.20-4.23)	**0.01**	59 (28.0)	87 (32.8)	1.40 (0.81-2.39)	0.21
GC+CC	72 (70.5)	299 (80.1)	1.68 (1.02-2.76)	**0.04**	158 (75.2)	213 (80.3)	1.27 (0.80-2.01)	0.30

a)-Adjusted Odds ratios, 95% confidence intervals and their corresponding *p*-values were calculated by logistic regression analysis after adjusting for age, gender and smoking habit. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies the cases and controls. The number in bold indicates the significant values in the table.

5.12.1 Association of *MSH2* genetic variants with overall survival of lung cancer patients

As depicted in **Table 5.37**, our survival data shows that none of the polymorphic variants for the *MSH2* gene, *i.e., rs63749993* (*Met^{688}Arg*), *rs2303426* (*IVS1+9G>C*), *rs4987188* (*Gly^{322}Asp*), and *rs2303428* showed no significant association to overall survival. However, our results revealed that patients harboring the mutant type genotype (*CC*) for *rs2303425* polymorphism had significantly reduced survival time as compared to the lung cancer subjects carrying the wild genotype (*TT*) (MST=3.13 vs. 7.91 months; HR=1.91; Log-rank **p=0.03**) (**Table 5.37**). Furthermore, after applying the cox regression model, it was evident that patients with the mutant genotype had a high death ratio (HR=1.86, **p=0.04**), thus suggesting a poor prognosis. In addition, our data also observed that for the *rs17217772* (*Asn^{127}Ser*) polymorphism; subjects who were harboring a single copy of the variant allele (*AG*) (*Asn^{127}Ser*) reported a significant decrease in median survival time and a higher hazards ratio in comparison to those patients who had the wild genotype (*AA*, *Asn/Asn*) (4.4 vs. 8.8 months; HR=1.41; log-rank **p=0.012**) as shown in **Table 5.37**.

Table 5.37: Association of *MSH2* polymorphism on overall survival in lung cancer cases and based on histology

Overall Survival								
	Cases	**Dead**	**Alive**	**Univariate analysis**			**Multivariate analysis**	
rs2303425 118T>C	**N(%), N= 475**	**N(%), N= 356**	**N(%), N= 119**	**MST (months)**	**Log rank *p***	**Unadjusted hr[a]**	**Adjusted hr[b] (95% CI)**	***P***
TT	328 (69.1)	239 (67.4)	89 (74.8)	7.91				
TC	135 (28.4)	106 (29.6)	29 (24.4)	9.7	0.39	0.90 (0.72-1.13)	0.80 (0.62-1.02)	0.07
CC	12 (2.5)	11 (3.0)	1 (0.8)	3.13	0.03	1.91 (0.84-4.33)	1.86 (0.98-3.52)	**0.04**
TC+CC	147 (30.9)	117 (32.8)	30 (25.2)	8.83	0.69	0.95 (0.76-1.19)	0.85 (0.67-1.08)	0.19
	Cases	**Dead**	**Alive**	**Univariate analysis**			**Multivariate analysis**	
rs17217772 Asn^{127}Ser	**N(%), N= 475**	**N(%), N= 356**	**N(%), N= 119**	**MST (months)**	**Log rank *p***	**Unadjusted hr[a]**	**Adjusted hr[b] (95% CI)**	***P***
AA	400 (84.2)	292 (82.0)	108 (90.7)	8.8				
AG	73 (15.4)	62 (17.4)	11 (9.3)	4.4	**0.012**	1.4 (1.03-1.91)	1.25 (0.94-1.66)	**0.03**
GG	2 (0.42)	2 (0.6)	0 (0)	12.5	0.86	1.12 (0.25-4.94)	0.60 (0.14-2.57)	0.50
AG+GG	75 (15.8)	64 (17.9)	11 (9.3)	4.4	0.015	1.39 (1.02-1.88)	1.22 (0.92-1.61)	0.16

[a]Unadjusted Hazards ratio for Kaplan Meier analysis, [b] Hazards ratio for age, sex, smoking, histology, stage, KPS, ECOG

We further stratified our OS data based on histological sub-types to estimate the MST of patients based on the different histology of lung cancer concerning the polymorphic variants of the *MSH2* gene. As shown in **Table 5.38**, the patients who were diagnosed with adenocarcinoma and harboring a single copy of the variant allele (*TG*) for *rs63749993* polymorphism reported a significant increase in MST and a corresponding decrease in HR (12.6 vs. 8.23 months; log-rank p=0.06) as compared to those adenocarcinoma subjects who had the wild genotype (*TT*). Furthermore, after cox regression analysis, the adenocarcinoma subjects carrying the genotype mentioned above had a significantly lower death ratio (HR=0.76, 95%CI= 0.51-1.13, p=0.18) not significant. Further, for *118T>C* polymorphism, we observed that the individuals carrying mutant allele (*CC*) have a much lower MST (12.43 vs. 3.36 months; log-rank **p=0.008**) as compared to wild type genotype (*TT*). Further, when we applied Cox regression analysis, we reported a corresponding increase in the hazard ratio (HR=2.93; 95%CI=1.19-7.17; **p=0.01**), showing a bad prognosis (**Table 5.38**) (**Figure 5.25**). Our results showed no association for the remaining *MSH2* polymorphisms with survival for patients diagnosed with adenocarcinoma.

Table 5.38: Association of *MSH2* polymorphism on overall survival in lung cancer cases and based on histology (ADCC)

ADCC								
				Univariate analysis			**Multivariate analysis**	
rs2303425 118T>C	**Cases N(%), N= 191**	**Dead N(%), N= 131**	**Alive N(%), N= 60**	**MST (months)**	**Log rank *p***	**Unadjusted hr[a]**	**Adjusted hr[b] (95% CI)**	***P***
TT	137 (71.7)	94 (71.7)	43 (71.7)	8.8				
TC	48 (25.1)	31 (23.7)	17 (28.3)	12.43	0.16	0.75 (0.51-1.10)	0.63 (0.41-0.98)	**0.03**
CC	6 (3.2)	6 (4.5)	0 (0)	3.36	**0.008**	2.87 (0.74-11.0)	2.93 (1.19-7.17)	**0.01**
TC+CC	54 (28.3)	37 (28.2)	17 (28.3)	11.13	0.42	0.85 (0.59-1.24)	0.75 (0.50-1.12)	0.16

[a] Unadjusted Hazards ratio for Kaplan Meier analysis, [b] Hazards ratio for age, sex, smoking, histology, stage, KPS, ECOG

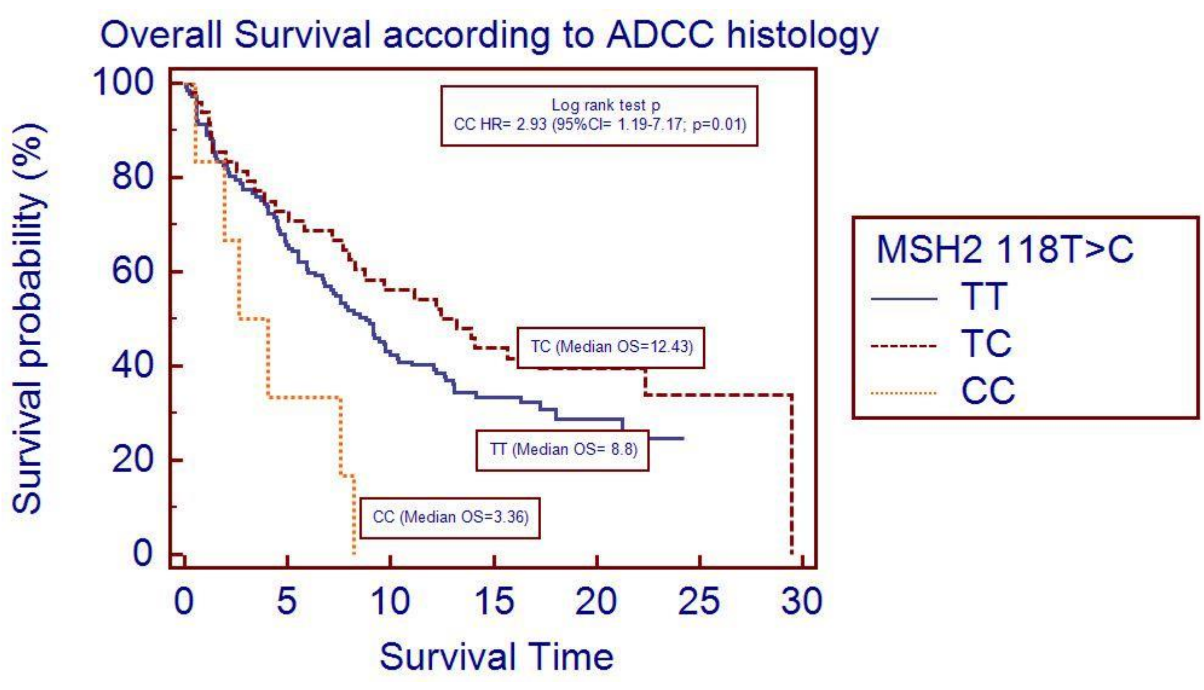

Figure 5.25: Overall survival of *MSH2 118T>C* variant based on ADCC histology

5.12.2 Association of *MSH2* genetic variants with overall survival of lung cancer patients based on the chemotherapeutic regimen

We have also evaluated the association between *MSH2* polymorphisms and the overall survival of patients based on different chemotherapeutic regimens administered. As represented in **Table 5.39**, when we compared the role of *MSH2* polymorphic variants in patients undergoing irinotecan-based platinum therapy, it was observed that for *rs63749993* (*Met*^{688}*Arg*) polymorphism, patients harboring heterozygous (*Met/Arg*) genotype reported a significant decrease in MST and overall survival in comparison to the patients carrying both the copies of the wild alleles for the *MSH2* (*Met*^{688}*Arg*) polymorphism (6.06 vs. 13.26; 95%CI=1.17-6.75; Log-Rank **p=0.001**) despite having the same type of chemotherapeutic agent administered. Further, the cox model reported a 5-fold increase in the hazard ratio for such patients (HR=5.37, 95%CI=2.11-13.66, **p=0.0004**), thus suggesting a higher likely hood of death and a poor prognosis for such patients based upon the *MSH2* variant allele and type of chemotherapeutic agent given. In the *rs2303426* polymorphism, heterozygous and mutant type genotype reported a significant increase in survival time (5.3 vs 15.1, 10.4, HR1= 0.39, HR2 =0.20; log-rank **p=0.05**, and **p=0.02**).

Table 5.39: Association of *MSH2* polymorphic variants on survival in lung cancer cases based on the chemotherapeutic regimen

IRINOTECAN								
rs63749993/ c.2063T>G /	**Cases N(%), N= 41**	**Dead N(%), N= 33**	**Alive N(%), N= 8**	**Univariate analysis MST (months)**	**Log rank *p***	**Unadjusted hr[a]**	**Multivariate analysis Adjusted hr[b] (95% CI)**	***P***
TT	27 (65.9)	19 (57.6)	8 (100)	13.26				
TG	14 (34.1)	14 (42.4)	0 (0)	6.06	0.001	2.82 (1.17-6.75)	5.37 (2.11-13.66)	**0.0004**
GG	0 (0)	0 (0)	0 (0)	-	-	-	-	-
TG+GG	14 (34.1)	14 (42.4)	0 (0)	6.06	0.001	2.82 (1.17-6.75)	5.37 (2.11-13.66)	0.**0004**
rs2303426/ IVS1+9G>C	**Cases N(%), N= 41**	**Dead N(%), N= 33**	**Alive N(%), N= 8**	**Univariate analysis MST (months)**	**Log rank *p***	**Unadjusted hr[a]**	**Multivariate analysis Adjusted hr[b] (95% CI)**	***P***
GG	6 (14.6)	6 (18.3)	0 (0)	5.3				
GC	14 (34.2)	9 (27.3)	5 (62.5)	15.1	0.05	0.39 (0.11-1.37)	-	
CC	21 (51.2)	18 (54.4)	3 (37.5)	10.4	0.06	0.43 (0.12-1.44)	0.20 (0.05-0.77)	**0.02**
GC+CC	35 (85.4)	27 (81.8)	8 (100)	10.4	0.03	0.4 (0.11-1.39)	0.27 (0.1-0.75)	**0.01**

[a]Unadjusted Hazards ratio for Kaplan Meier analysis, [b]hazards ratio for age, sex, smoking, histology, stage, KPS, ECOG

5.12.3 High-order SNP-SNP interactions of *MSH2* gene and prognosis of lung cancer patients using survival tree analysis (STREE)

Multiple interaction studies, including different SNPs of the *MSH2* gene involved in the MMR pathway, can add a new dimension to find out the high-order parametric interactions among them and their effect on the overall survival of these interactions the patients. The current study uses the recursive partitioning method to generate a survival tree that shows the involvement of different SNPs in prognosis. The tree was generated by STREE software, as shown in **Figure 5.26**. The log-rank method created the tree, which had nine terminal nodes. The factor which contributed most significantly to the prognosis of patients was *MSH2 rs17217772* (*Asn^{127}Ser*) which served as a root node. Terminal node 7 has the highest Median survival time (MST) of 15.4 months and was taken as reference. MST and HR of all the terminal nodes are shown in **Figure 5.26**. shows the comparison among survival curves of all nodes, showing a notable difference between MST of different nodes compared with the reference node (**Table 5.40**).

Further, Cox regression analysis was also employed to calculate the adjusted hazard ratio for each node after considering confounding factors such as gender, smoking, age, ECOG, and tumor stage. Node 6 depicts the lowest death rate (HR=0.59; **p=0.30**) and the same MST (14.9 vs.151.4 months) as the reference node. Node 5 showed the highest death rate and lowest MST (HR=6.58; **p=0.002**). All the remaining nodes showed a similar trend of good prognosis (**Figure 5.26**).

Table 5.40: High order gene-gene interaction table

Name	**MST (months)**	**Unadjusted hr[a] (95% CI)**	**Log rank p**	**Adjusted hr[b] (95% CI)**	**p**
Reference Node 7	15.4	1.00		1.00	
Node 1	8.4	1.70 (1.05-2.75)	0.07	1.85 (1.00-3.40)	0.04
Node 2	8.4	1.80 (0.92-3.54)	0.09	1.99 (0.89-4.42)	0.09
Node 3	10.5	1.86 (1.02-3.40)	0.06	2.25 (1.10-4.59)	0.02
Node 4	5.2	-	-	-	-
Node 6	14.9	1.00 (0.41-2.41)	0.99	0.59 (0.22-1.60)	0.30
Node 5	1.9	4.48 (1.43-14.05)	0.0001	6.58 (2.00-21.6)	0.002
Node 8	6.1	2.30 (0.85-6.23)	0.03	-	-
Node 5+8	6.8	3.29 (1.55-6.97)	<0.0001	3.6 (1.77-7.49)	**0.0005**

[a]Unadjusted Hazards ratio for Kaplan Meier analysis, [b]hazards ratio for age, sex, smoking, histology, stage, KPS, ECOG

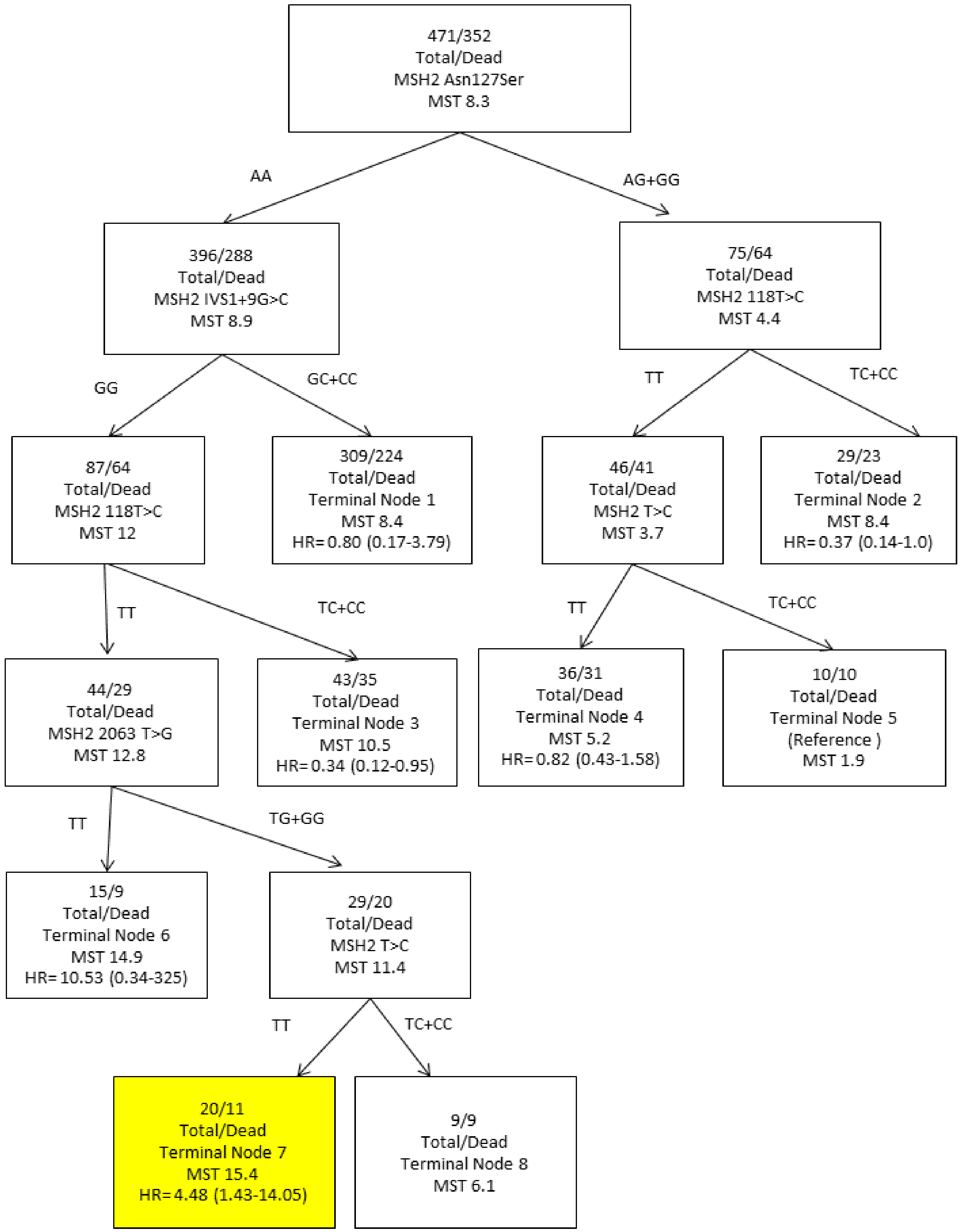

Figure 5.26: Survival tree analysis of MSH2 variants; Yellow color shows reference terminal node.

5.13 Toxicity analysis

5.13.1 Association between *MSH2* polymorphisms and toxicity:

One of the outcomes of interest in the current study was investigating the toxicity in lung cancer subjects undergoing platinum-based doublet chemotherapy. Univariate logistic regression method was used to determine the odds ratio (OR) and 95% confidence interval (95%CI) to estimate the risk of developing various toxicities after chemotherapy in patients with each specific genotype. Further multivariate logistic regression method was used to determine adjusted ORs, considering toxicity grades as the dependent variable. In the multivariate model, genotype, age, gender, and performance status were taken as confounding variables to evaluate if these variables have any effect. We performed three types of analysis to investigate the impact of polymorphisms on different grades of toxicities. In the first analysis, we compared the presence of any toxicity grade (grade 1-4) with the absence of toxicity (grade 0) in the second analysis. Then we compared the patients manifesting intermediate/severe toxicity (grade 2-4) with no /low grade of toxicity (grade 0-1), and in the third analysis, we compared patients manifesting severe toxicity (grade 3-4) with no/low/intermediate toxicity (grade 0-2) to thoroughly evaluate whether different polymorphism of *MSH2* gene was associated with severe toxicities. Analysis of severe neutrophil toxic events (grade 2-4) showed less frequent toxicities for patients with both copies of mutant (*CC*) allele 70 (24.1%) than for patients homozygous for wild type (*GG*) allele 5 (71.4%) in *MSH2 IVS1+9G>C* polymorphism. However, when logistic regression was employed, patients homozygous for mutant type (*CC*) allele showed 2.47-fold increased neutropenia when compared to patients homozygous to wild type (*GG*) allele, but this increased risk does not reach the threshold of statistical significance (AOR=2.47; 95%CI=0.82-7.33; p=0.10) (**Table 5.41**). Further, for *MSH2 T>C/-6* polymorphism, 16 (5.7%) and 3 (17.6%) subjects were harboring mutant type genotype (*CC*) and belonged to toxicity group 0 and 1-4 respectively. Lung cancer subjects carrying both copies of the mutant allele (*CC*) reported a 2.2-fold higher risk of developing neutropenia (ANC) as compared to subjects carrying both copies of the wild allele (*TT*) (AOR=2.2; 95%CI=1.07-4.66; **p=0.03**). When we compared the patients manifesting neutropenia of grade 0-1 and grade 2-4, then our results further concluded that subjects harboring mutant type genotype (*CC*) reported a 2.4-fold higher risk of developing neutropenia as compared to wild type genotype (*TT*) (AOR=2.4; 95%CI=0.98-6.24; p=0.05) (**Table 5.41**).

Table 5.41: Neutropenia toxicity (grade 0 vs.1-4; grade 0-1 vs. 2-4) associated with MSH2 genetic variants in multivariate analysis

MSH2 T>C/-6	**Grade of toxicity 0, n(%) 280**	**Grade of toxicity 1-4, n(%) 17**	**OR (95% CI)**	***P*[b]**	**AOR (95% CI) [a]**	***P*[b]**
TT	198 (70.7)	11 (64.7)	Reference			
TC	66 (23.6)	3 (17.6)	0.81 (0.22-3.02)	0.76	0.75 (0.20-2.81)	0.67
CC	16 (5.7)	3 (17.6)	1.83 (0.92-3.65)	0.08	**2.23 (1.07-4.66)**	**0.03**
TC+CC	82 (29.3)	6 (35.3)	1.31 (0.47-3.68)	0.59	1.35 (0.48-3.80)	0.56
MSH2 T>C/-6	**Grade of toxicity 0-1, n(%) 290**	**Grade of toxicity 2-4, n(%)7**	**OR (95% CI)**	***P*[b]**	**AOR (95% CI) [a]**	***P*[b]**
TT	204 (70.3)	5 (71.4)	Reference			
TC	69 (23.8)	0 (0.0)	-	-	-	-
CC	17 (5.9)	2 (28.6)	2.10 (0.93-5.15)	0.07	**2.47 (0.98-6.24)**	**0.05**
TC+CC	86 (29.7)	2 (28.6)	0.94 (0.18-4.98)	0.95	0.94 (0.17-5.00)	0.94

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age, gender, and ECOG. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies the cases and controls. A p-value <0.05 was considered statistically significant. The number in bold indicates the significant values in the table.

5.13.2 Association between *MSH2* polymorphisms and gastrointestinal toxicity

Further, for MSH2 118T>C polymorphism, the subjects having mutant type genotype and manifesting grade 1 and grade 2-4 diarrhea were: 4 (8.6%) and 15 (30.0%), respectively. They have also reported a 4.3-fold increased risk of developing severe grade diarrhea (AOR=4.33; 95%CI=1.30-14.4; p=0.01). The subjects with a combined genotype also reported a 2.5-fold increased risk of suffering from a severe grade of diarrhea (AOR=2.54; 95%CI=0.90-7.16; p=0.04). Our results further reported that for MSH2 2063T>G polymorphism, individuals harboring a single copy of mutant allele manifested a 2.9-fold increased risk of developing grade 2-4 diarrhea (AOR=2.93; 95%CI=1.16-7.52; p=0.02). Subjects having the same genotype also reported a 3.64- fold increased risk of developing severe grade diarrhea (AOR=3.6; 95%CI=1.17-11.3; p=0.02) (**Table 5.42**). For MSH2 IVS1+9G>C polymorphism, 13 (27.7%) and 8 (16.0%) patients were suffering from grade 1 and 2-4 diarrhea and were carrying mutant type genotype (CC) reported a reduced risk of developing grade 2-4 diarrhea as compared with those homozygous for the G allele (AOR=0.53; 95%CI=0.28-1.01; p=0.04). Following the same trend, the mutant type and combined genotype (GC+CC) reported a reduced risk of developing severe grade diarrhea (AOR=0.41; 95%CI=0.18-0.96; p=0.04, AOR=0.39; 95%CI=0.17-0.87; p=0.02). Further, for MSH2 Asn127Ser polymorphism, our results reported no association with gastrointestinal toxicity, for MSH2 T>C/-6 polymorphism, 18 (38.3%) subjects harboring combined genotype (TC+CC) belonged to grade 1 diarrhea, whereas 11 (22.2%) subjects were suffering from grade 2-4 diarrhea our results reported a diminished risk of developing grade 2-4 diarrhea for individuals carrying combined genotype (TC+CC) (AOR=0.42; 95%CI=0.18-0.96; p=0.03).

Table 5.42: Groups of graded gastrointestinal toxicities (Diarrhea) and the relationship with platinum–induced toxicities in lung cancer patients for *MSH2* polymorphisms.

Diarrhea						
MSH2 118T>C	**Grade of toxicity 1, n(%) 47**	**Grade of toxicity 2-4, n(%) 50**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
TT	42 (89.3)	35 (70.0)				
TC	4 (8.6)	15 (30.0)	4.5 (1.36-14.8)	0.01	4.33 (1.30 -14.4)	**0.01**
CC	1 (2.1)	0 (0)	-	-	-	-
TC+CC	5 (10.6)	15 (30.0)	2.66 (0.96-7.36)	0.05	2.54 (0.90-7.16)	**0.04**
MSH2 2063T>G	**Grade of toxicity 1, n(%) 47**	**Grade of toxicity 2-4, n(%) 50**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
TT	31 (68.9)	21 (42.0)				
TG	12 (25.5)	24 (48.0)	2.95 (1.2-7.16)	0.01	2.93 (1.16-7.52)	**0.02**
GG	4 (8.5)	5 (10.0)	1.35 (0.66-2.77)	0.40	1.25 (0.57-2.73)	0.56
TG+GG	16 (34.0)	29 (58.0)	1.84 (0.97-3.48)	0.06	1.81 (0.91-3.57)	0.08
MSH2 2063T>G	**Grade of toxicity 1-2, n(%) (77)**	**Grade of toxicity 3-4, n(%) (20)**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
TT	46 (59.7)	6 (30.0)				
TG	24 (31.2)	12 (60.0)	3.83 (1.27-11.4)	0.01	3.64 (1.17-11.3)	**0.02**
GG	7 (9.1)	2 (10.0)	1.48 (0.60-3.61)	0.38	1.45 (0.52-4.05)	0.47
TG+GG	31 (40.3)	14 (70.0)	1.95 (0.94-4.01)	0.06	1.88 (0.85-4.16)	0.11
MSH2 *IVS1+9G>C*	**Grade of toxicity 1, n(%) 47**	**Grade of toxicity 2-4, n(%) 50**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
GG	8 (17.0)	16 (32.0)				
GC	26 (55.3)	26 (52.0)	0.50 (0.18-1.36)	0.17	0.57 (0.20-1.62)	0.29
CC	13 (27.7)	8 (16.0)	0.55 (0.30-1.0)	0.05	0.53 (0.28-1.01)	**0.04**
GC+CC	39 (82.9)	34 (68.0)	0.55 (00.30-1.0)	0.05	0.54 (0.28-1.01)	**0.04**
MSH2 *IVS1+9G>C*	**Grade of toxicity 1-2, n(%) (77)**	**Grade of toxicity 3-4, n(%) (20)**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
GG	15 (19.5)	9 (45.0)				
GC	43 (55.8)	9 (45.0)	0.34 (0.11-1.04)	0.05	0.40 (0.12-1.29)	0.12
CC	19 (24.7)	2 (10.0)	0.41 (0.18-0.96)	**0.04**	0.43 (0.18-1.04)	0.06
GC+CC	62 (80.5)	11 (55.0)	0.39 (0.17-0.87)	**0.02**	0.43 (0.19-0.96)	**0.04**
MSH2 T>C/-6	**Grade of toxicity 1, n(%) 47**	**Grade of toxicity 2-4, n(%) 50**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
TT	29 (61.7)	39 (78.0)				
TC	14 (29.8)	11 (22.0)	0.58 (0.23-1.47)	0.25	0.55 (0.21-1.41)	0.21
CC	4 (8.5)	0 (0)	-	-	-	-
TC+CC	18 (38.3)	11 (22.2)	0.42 (0.19-0.93)	0.03	0.42 (0.18-0.96)	**0.03**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age, gender, and ECOG. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies the cases and controls. A p-value <0.05 was considered statistically significant. The number in bold indicates the significant values in the table.

5.13.3 Association between *MLH1* polymorphisms and gastrointestinal toxicity

As depicted in **Table 5.43**, for *MLH1 -93G>A* polymorphism, for gastrointestinal toxicity, we reported that a total of 38 patients were suffering from grade 1 nausea/vomiting, and 41 were suffering from grade 2-3 nausea/vomiting. The individuals harboring heterozygous type genotype and suffering from grade 1 toxicity were 17 (44.7%) compared to 13 (37.1%) patients suffering from grade 2-3. They also showed a protective effect towards severe grade nausea/ vomiting development, but the results were not statistically significant (AOR=0.46; 95%CI=0.16-1.3; p=0.14). Further, 50 patients suffered from grade 2-4 diarrhea, and 47 from grade 1 diarrhea. Similarly, when patients were further stratified into grades 1-2 and 3-4, our results showed that 77 belonged to the former category and 20 belonged to the latter. Almost 50% of subjects carried heterozygous type genotype for severe grade diarrhea. For grade 1-2 diarrhea, 39% of subjects were carrying combined genotype. Our results reported a 3.2-fold increased risk of developing severe grade diarrhea (AOR=3.24; 95%CI=1.08-9.71; **p=0.03**) for individuals carrying heterozygous type genotype. Further, we reported no association between *MLH1* polymorphism and the risk of developing constipation and anorexia. (**Table 5.43**).

Table 5.43: Gastrointestinal toxicity grade 1-2 vs. 3-4 (Diarrhea), associated with MLH1 genetic variant in multivariate analysis

			Diarrhea			
MLH1 -93G>A	**Grade of toxicity 1-2, n(%) (77)**	**Grade of toxicity 3-4, n(%) (20)**	**OR (95% CI)[a]**	**P[b]**	**AOR (95% CI)[a]**	**P[b]**
GG	47 (61.0)	9 (45.0)				
GA	20 (26.0)	10 (50.0)	2.61 (0.92-7.38)	0.07	3.24 (1.08-9.71)	**0.03**
AA	10 (13.0)	1 (5.0)	0.72 (0.24-2.14)	0.55	0.62 (0.19-1.99)	0.42
GA+AA	30 (39.0)	11 (55.0)	1.17 (0.58-2.36)	0.64	1.24 (0.61-2.53)	0.54

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age, gender, and ECOG. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies the cases and controls. A p-value <0.05 was considered statistically significant. The number in bold indicates the significant values in the table

5.13.4 Association between *MSH3* polymorphisms and toxicity

As shown in **Table 5.44**, 23 (23.7%), 56 (23.2%) lung cancer patients were carrying heterozygous type genotype (*GA*) and belonged to anemia toxicity grade 0 and grade 1-4, respectively. 68 (70.1%), 182 (75.5%) lung cancer patients were harboring mutant type genotype (*AA*) and belonged to anemia toxicity grade 0 and grade 1-4, respectively. Further, our results concluded that subjects carrying one copy of variant allele (*GA*) manifest a 7.23-old increased risk of developing anemia (AOR=7.23; 95%CI=1.51-34.6; **p=0.013**) as compared to the subjects harboring both copies of variant allele (*AA*) that reported a 2.48-fold increased risk of developing anemia (AOR=2.48; 95%CI=1.21-5.08; **p=0.012**). Our results also suggested that individuals carrying the combined genotype (*GA+AA*) reported a 6.39-fold increased risk of developing anemia (AOR=6.39; 95%CI=1.53-26.6; p=0.010). When we analyzed *MSH3 G>A* polymorphism to gastrointestinal toxicity, we observed that the frequency of individuals carrying a single copy of mutant allele (*GA*) having severe toxicity (grade 3-5) is less compared to patients manifesting low-grade toxicity. Further, the risk of developing high-grade toxicity is 2.4-fold increased compared to patients with wild-type genotype (*GG*) (AOR=2.40; 95%CI=0.49-11.6; p=0.27). Following the same trend, the combined genotype (*GA+AA*) subjects have a higher frequency of toxicity and manifested a 2.8-fold increased risk of developing gastrointestinal toxicity (AOR=2.81; 95%CI=0.60-13.0; p=0.18) (**Table 5.44**).

Table 5.44: Hematological, toxicity (grade 0 vs. 1-4 (Anemia); associated with *MSH3* genetic variant in multivariate analysis

			Anemia			
MSH3 G>A	**Grade of toxicity 0, n (%) 97**	**Grade of toxicity 1-4, n(%) 241**	**OR (95% CI)**	**P^b**	**AOR (95% CI) ^{a}**	**P^b**
GG	6 (6.2)	3 (1.2)	Reference			
GA	23 (23.7)	56 (23.2)	4.86 (1.12-21.1)	0.03	7.23 (1.51-34.6)	**0.013**
AA	68 (70.1)	182 (75.5)	5.35 (1.30-22.0)	0.02	2.48 (1.21-5.08)	**0.012**
GA+AA	91 (93.8)	238 (98.8)	5.23 (1.28-21.3)	0.02	6.39 (1.53-26.6)	**0.010**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age, gender, and ECOG. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies the cases and controls. A p-value <0.05 was considered statistically significant. The number in bold indicates the significant values in the table

5.13.5 Association between *MSH6 C>T* polymorphisms and toxicity

As shown in **Table 5.45**, for *MSH6 C>T* polymorphism, 79 (35.3%) and 51 (44.7%) lung cancer patients carrying one copy of mutant allele (*CT*) belonged to anemia toxicity grade 0-1 and grade 2-4, respectively. Further, 100 (44.6%) and 64 (56.1%) lung cancer patients were carrying combined genotype (*CT+TT*) belonged to anemia toxicity grade 0-1 and grade 2-4, respectively. Our results concluded that subjects with combined genotype reported a 1.58-fold increased risk of developing anemia compared to patients carrying wild-type genotype (AOR=1.58; 95%CI=1.0-2.49; **p=0.04**). We further evaluated the lung cancer patients for thrombocytopenia, and as shown in table 5.55, 148 (50.3%) and 13 (36.1%) were carrying combined genotype (*CT+TT*) and belonged to toxicity grade 0 and grade 1-2, respectively. Lung cancer patients having combined genotype (*CT+TT*) reported a decreased risk of developing thrombocytopenia as compared to subjects carrying both copies of wild allele (*CC*) (AOR=0.50; 95%CI=0.27-0.93; **p=0.02**). We did not find any *MSH6 C>T* polymorphism association with neutropenia and nephrological toxicity (**Table 5.45**).

Table 5.45: Hematological, toxicity (grade 0-1 vs. 2-4 (Anemia); grade 0 vs. 1-2 (Thrombocytopenia); associated with *MSH6 C>T* genetic variant in multivariate analysis

			Anemia			
MSH6 C>T	**Grade of toxicity 0-1, n(%) 224**	**Grade of toxicity 2-4, n(%) 114**	**OR (95% CI)**	***P*[b]**	**AOR (95% CI) [a]**	***P*[b]**
CC	124 (55.4)	50 (43.9)	Reference			
CT	79 (35.3)	51 (44.7)	1.60 (0.98-2.59)	0.05	1.55 (0.94-2.54)	0.08
TT	21 (9.4)	13 (11.4)	1.53 (0.71-3.30)	0.27	1.18 (0.80-1.75)	0.39
CT+TT	100 (44.6)	64 (56.1)	1.58 (1.00-2.49)	**0.04**	1.52 (0.95-2.42)	0.07
			Thrombocytopenia			
MSH6 C>T	**Grade of toxicity 0, n(%) 294**	**Grade of toxicity 1-2, n(%) 36**	**OR (95% CI)**	***P*[b]**	**AOR (95% CI) [a]**	***P*[b]**
CC	146 (49.7)	23 (63.9)	Reference			
CT	114 (38.8)	13 (36.1)	0.72 (0.35-1.49)	0.38	0.73 (0.35-1.52)	0.40
TT	34 (11.6)	0 (0)	-	-	-	-
CT+TT	148 (50.3)	13 (36.1)	0.51 (0.27-0.94)	0.11	0.50 (0.27-0.93)	**0.02**

a)-Adjusted Odds ratios, 95% confidence intervals, and corresponding p-values were calculated by logistic regression analysis after adjusting for age, gender, and ECOG. (b)-Two-sided $\chi 2$ test for either genotype distribution or allelic frequencies the cases and controls. A p-value <0.05 was considered statistically significant. The number in bold indicates the significant values in the table

5.14 Bioinformatics analysis of SNPs taken in experimental studies:

We did polymorphism study of 11 SNPs of four genes (*MLH1*, *MSH2*, *MSH3* and *MSH6*) to show the association towards the lung cancer susceptibility in previous objectives, now bioinformatics analysis were carried out in order to see the effect of these SNPs on the structure and function of proteins. Five out of 11 SNPs were present either in UTR or intronic region which will not result into amino acid change hence only 6 SNPs were considered for computational analysis. For MSH2 protein, *rs63749993* and *rs17217772* polymorphism was shown to be deleterious by SIFT, PROVEAN, PANTHER, Polyphen-2, PhD-SNP and SNP&GO whereas I-mutant and MuPRO predicted it to have decreased stability. *rs4987188* polymorphism was predicted to have damaging effect by PANTHER, Polyphen-2 only and I-mutant and MuPro predicted it to have decreased stability. *rs1800932*, *rs1042821* and *rs26279* was shown to have damaging effect by PANTHER, Polyphen-2 whereas only *rs1800932*, *rs1042821* were predicted to have increased stability by I-mutant and MuPRO (
Table 5.46).

Table 5.46: Computational analysis of the SNPs present in the coding region and selected for wet lab experimentation.

S. No.	rs IDs	Amino Acid Substitution	Gene	SIFT	PROVEAN	PANTHER	Polyphen	PhD SNP	SNP & GO	I-mutant	Mu PRO
1	rs63749993	M668R	MSH2	Deleterious	Deleterious	Invalid substitution	Probably damaging	Disease	Disease	Decrease	Decrease
2	rs4987188	G322D	MSH2	Tolerated	Neutral	Possibly damaging	Possibly damaging	Neutral	Neutral	Decrease	Decrease
3	rs17217772	N127S	MSH2	Deleterious	Deleterious	Probably damaging	Possibly damaging	Disease	Neutral	Decrease	Decrease
4	rs1800932	P92P	MSH6	Tolerated	Neutral	Invalid substitution	Probably damaging	Disease	Neutral	Increase	-
5	rs1042821	G39E	MSH6	Tolerated	Neutral	Possibly damaging	Benign	Neutral	Neutral	Increase	Increase
6	rs26279	A1045T	MSH3	Tolerated	Neutral	Probably benign	Benign	Neutral	Incorrect mutation	Decrease	Not found

5.15 Structural and functional analysis of Mismatch repair (MMR) system SNPs based on different computational tool

In this part, we have explored others SNPs present in NCBI SNPs database of *MSH2, MSH6, MSH3* and *MLH1* to find the functional SNPs which can be further selected for experimental validation.

5.15.1 MSH2 protein

The Human MSH2 gene contains a total number of 80098 bp and MSH2 protein has 934 amino acids. SNP database has 68322 SNPs of human MSH2 gene in May 2020, out of which 1570 belongs to the category of clinical significance which was considered for analysis. 174 entries out of 1570 were found to be redundant hence were deleted so the final numbers which were taken for analysis were 1396.

5.15.1.1 Identification of deleterious and damaging nsSNP's based on sequence-based homology approach

SIFT, PROVEAN and PANTHER tools which use sequence-based homology approach to predict the deleterious and damaging nsSNP's were used to analyze 1396 SNPs. Further, SIFT was showing three protein IDs for the same "rs ID" so the result of one protein id (ENSP00000233146) containing the full length (934 amino acids) was taken. After selecting single protein ID and non-synonymous mutations, 122 SNPs were left. Out of 122 SNPs, 114 rs IDs were found to have either had either one or two amino acid changes which resulted in 122 amino acid changes. SIFT predicted 76 and 46 as deleterious and tolerated respectively out of 122 SNPs. These 122 SNPs were further analyzed by PROVEAN and PANTHER for identification of SNPs having deleterious and damaging effects. PROVEAN predicted 77 and 45 as deleterious and tolerated respectively while PANTHER server predicted 107 nsSNPs to be damaging and 15 nsSNPs as probably benign. Interestingly, 70 SNPs were found to be commonly predicted as deleterious or damaging by all the three tools used (**Table 5.47**).

5.15.1.2 Identification of deleterious and damaging nsSNP's using sequence and structure based approach (PolyPhen):

Polyphen tool predicts if particular SNPs are having deleterious or benign effects. 64 and 63 SNPs out of 70 were predicted to be damaging by HumDiv and HumVar data sets of Polyphen respectively. 63 SNPs were taken into consideration which is predicted by both HumDiv and HumVar data sets (**Table 5.47**). Both datasets of Polyphen have predicted 58 SNPs to be probably damaging having a score range of 0.925 to 1.0 and 5 SNPs were predicted to be probably damaging and possibly damaging by either HumDiv or HumVar data set. Moreover, Seven SNPs were reported to be benign by Polyphen 2 server is underlined in **Table 5.47**.

Table 5.47: Deleterious SNPs commonly predicted by different tools

S.No.	rs IDs	Amino acid change	S.No.	rs IDs	Amino acid change
1	rs141711342	F19L	36	rs151244108	Q409E
2	rs17217723	Y43C	37	rs63750697	M453K
3	rs33946261	H46Q	38	rs63751403	S473L
4	rs63751429	L93F	39	rs63751207	R524P
5	rs17217772	N127S	40	rs63751207	R524L
6	rs63750126	V161D	41	rs63749910	I577T
7	rs63750624	G162R	42	rs63751140	G587R
8	rs63750214	V163D	43	rs63750657	D603N
9	rs63750582	G164R	44	rs63750657	D603Y
10	rs63750582	G164W	45	rs267607985	D603G
11	rs63750255	D167H	46	rs28929483	P622L
12	rs63750070	L173P	47	rs63750280	P622T
13	rs63750070	L173R	48	rs371776176	G628R
14	rs63750382	E177K	49	rs28929484	H639Y
15	rs151129360	N186S	50	rs267607983	P652H
16	rs63751444	L187P	51	rs63751640	G669D
17	rs63751444	L187R	52	rs63751668	G669S
18	rs63750327	E198G	53	rs63750234	G674S
19	rs63751110	C199R	54	rs63750234	G674R
20	rs63750058	S269L	55	rs267607996	G674A
21	rs34136999	A272V	56	rs267607995	G683R
22	rs63751454	A305T	57	rs63749993	M688R
23	rs63750640	L310P	58	rs63750232	G692R
24	rs202026056	G315V	59	rs63751432	G692V
25	rs63750630	L330P	60	rs267607994	P696L
26	rs63750468	C333R	61	rs63750398	C697F
27	rs63750828	C333Y	62	rs63750961	C697R
28	rs63751004	G338R	63	rs373226409	C707Y
29	rs63751147	L341P	64	rs63751477	E749K
30	rs375799148	Q344H	65	rs63751119	G751R
31	rs267607939	P349A	66	rs386833406	G759E
32	rs63751604	R359I	67	rs200252727	H785P
33	rs63751617	R359S	68	rs63750623	Q824E
34	rs17224367	L390F	69	rs63750757	A834T
35	rs146567853	R406Q	70	rs41295182	L911R

*The underline SNPs are not predicted as damaging by Polyphen server.

5.15.1.3 Disease prediction by PhD SNP and SNP&GO:

PhD SNP and SNP&GO are used to analyze that single amino acid substitution is associated with the occurrence of a disease or not. 63 nsSNPs which were predicted to be deleterious by both sequence and structure-based homology approach were taken as input for PhD SNP and SNP GO. PhD SNP and SNP GO server has predicted 59 and 40 nsSNPs, respectively to be associated with disease.

5.15.1.4 Estimating the stability of the mutated protein structures using I-mutant and MuPro server:

Among the 40 nsSNPs submitted, I-mutant estimated one nsSNPs resulted in increased stability (*rs28929484*). The remaining 39 have shown a decrease in stability by the I-mutant server. Only two nsSNPs (*rs28929483*, *rs267607994*) have reported an increase in stability by MuPro server rest 38 have reported a decrease in stability. DDG value less than 0 is associated with a decrease in stability and 37 SNPs have reported a decrease in stability by both I Mutant and MuPro server as shown in **Table 5.48**. Further analysis revealed that these 37 SNPs were belonging to 32 rs IDs in which 27 and 5 are having one and two mutation respectively.

5.15.1.5 Modeling of the complete MSH2 protein using MODELLER v9.22

In the PDB database, protein ID 2O8B which showed 100% sequence identity with MSH2 protein was considered. When protein ID 2O8B MSH2 protein was checked for complete structure it was found to have the structure of only 851 amino acid residues and the structure of 85 residues was missing. The selected SNPs were present at different positions in protein**;** hence there was a need to generate the complete structure. Further, Protein ID 2O8B is also a complex structure involving two proteins MSH2 and MSH6, So MSH6 chains were removed using PyMol. The complete structure of MSH2 was then generated by using MODELLER v9.22 which is shown in **Figure 5.27**. The generated structure was further validated via PROCHECK and energy minimization was performed using Chimera tool.

Table 5.48: Stability prediction and RMSD value of selected deleterious SNPs

S.No.	rs IDs	AAS	RMSD (Å)	I-Mutant DDG Value	MuPro server Delta G value
1	rs17217723	Y43C	0.718	-1.25	-1.12
2	rs63750624	G162R	0.216	-0.56	-0.33
3	rs63750070	L173P	0.237	-2	-1.97
4	rs63750070	L173R	0.213	-1.92	-1.66
5	rs63750382	E177K	0.243	-0.96	-1.39
6	rs63751444	L187P	0.238	-1.79	-1.66
7	rs63751444	L187R	0.248	-1.68	-1.31
8	rs63750640	L310P	0.223	-1.98	-2.27
9	rs63750630	L330P	0.234	-1.09	-1.79
10	rs63750468	C333R	0.234	-0.1	-1.94
11	rs63750828	C333Y	0.211	0.04	-1.65
12	rs63751004	G338R	0.207	-0.04	-0.14
13	rs63751147	L341P	0.718	-1.17	-1.86
14	rs63751604	R359I	0.773	-0.44	-0.13
15	rs63751617	R359S	0.728	-1.36	-0.99
16	rs63751207	R524P	0.739	-0.65	-1.36
17	rs63751207	R524L	0.728	-0.36	-0.32
18	rs63751140	G587R	0.203	-0.57	-1.37
19	rs63750657	D603N	0.327	-0.84	-0.80
20	rs63750657	D603Y	0.224	-0.28	-0.73
21	rs267607985	D603G	0.256	-1.06	-1.65
22	rs63751640	G669D	0.22	-1	-0.76
23	rs63751668	G669S	0.311	-1.22	-0.89
24	rs63750234	G674S	0.282	-1.08	-1.02
25	rs63750234	G674R	0.215	-0.29	-1.05
26	rs267607996	G674A	0.213	-0.76	-1.18
27	rs267607995	G683R	0.25	-0.41	-0.69
28	rs63749993	M688R	0.715	-0.69	-1.13
29	rs63750232	G692R	0.239	-0.42	-0.43
30	rs63751432	G692V	0.211	-0.51	-0.03
31	rs63750398	C697F	0.236	-0.07	-0.10
32	rs63750961	C697R	0.211	-0.33	-0.46
33	rs63751477	E749K	0.307	-0.6	-0.99
34	rs63751119	G751R	0.215	-0.56	-0.73
35	rs386833406	G759E	0.753	-0.55	-0.65
36	rs63750757	A834T	0.208	-0.64	-1.12
37	rs41295182	L911R	0.692	-1.41	-1.92

*The underline SNPs is having double mutation for the same rs ID.

Figure 5.27: Complete structure of Energy Minimized MSH2 _Wild.

5.15.1.6 Root Mean Square Deviation value estimation of the modeled protein:

The structure of the mutant proteins with selected 32 nsSNPs (37 amino acid changes) was modeled by taking the previously generated MSH2 structure as a template with appropriate mutations. Further energy minimization of mutant proteins was done by Chimera tool using steepest descent and conjugate gradient algorithms. RMSD values were calculated by utilizing align facility in PYMOL.9 mutants have RMSD values between 0.69Å and 0.75 Å indicating certain structural deviation.

5.15.1.7 Mapping of the missense mutations in to MSH2 protein domains:

There are three domains in MSH2 viz. DNA binding domain, MutL homologs interaction domain, and hMSH3/ MSH6 interaction domain, these 32 SNPs with 37 amino acid changes were found to be distributed in these three domains (**Figure 5.28**). Fourteen SNPs (15 amino acid changes) were found to be MutL homologs interaction domain while ten SNPs (11 amino acid changes) were clustered in the hMSH3/ MSH6 interaction domain and five SNPs (seven amino acid changes) are located in the DNA binding domain, three SNPs (four amino acid changes) lies outside of the three domains. Amino acid positions 173 and 187 (DNA binding domain), 524 (hMSH3/ MSH6 interaction domain) 603, and 674 (MutL homologs interaction domain) are associated with two amino acid changes due to single SNPs. Amino acid positions 333, 359 (hMSH3/ MSH6 interaction domain), 669, 692, and 697 (MutL homologs interaction domain) have two amino acid changes due to different SNPs. Interestingly, 674 position (MutL homologs interaction domain) have three amino acid changes.

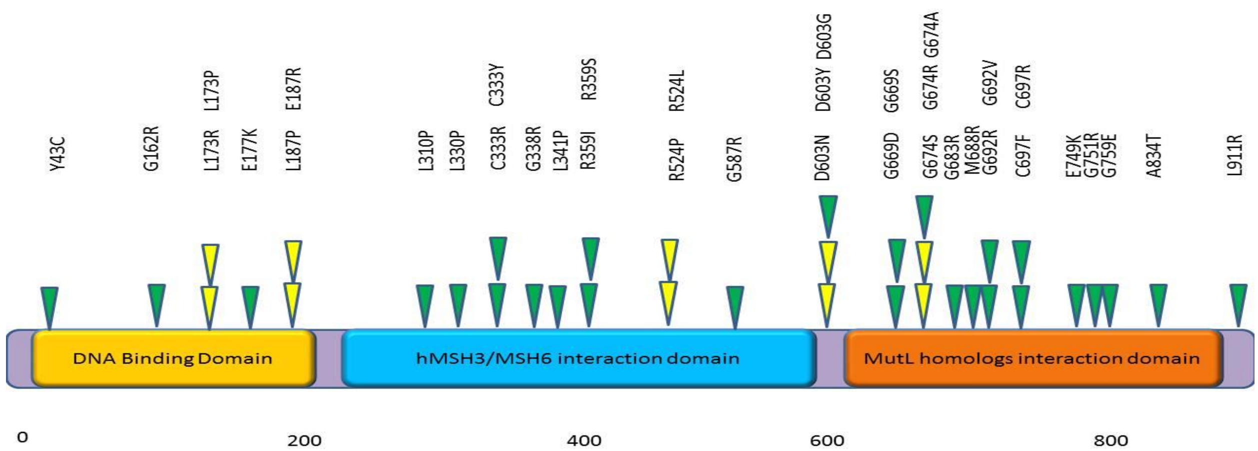

Figure 5.28: Location of different domains on MSH2 protein

5.15.1.8 SNPs affecting the interaction of MSH2 with MSH6

MSH2 interact with MSH6, hence the SNPs present in this interacting domain were selected for docking analysis. Molecular docking of MSH6 protein with native and mutants MSH2 protein showed a difference in HADDOCK score, hydrogen bonds, and Van der Waals energy (**Figure 5.29).** HADDOCK score helps in determining the binding energy of the peptide complex. Overall HADDOCK score for native MSH2 protein was -160.2 whereas the lowest HADDOCK score in mutants was -126.6 for G587R indicating a change in binding affinity of MSH2_G587R towards MSH6. Complete details of the HADDOCK score, Van der Waals energy, and Number of hydrogen bonds of mutant structures used in the docking study are shown in **Table 5.49**. We found an interesting relationship between the HADDOCK score and hydrogen bonds present in the docked structures. We have observed that as the HADDOCK score decreases the number of hydrogen bonds decreases indicating lower overall stability of the mutant proteins except for C333Y and R524L mutation. Six mutants structure (L310P, L330P, C333R, L341P, R524P, and G587R) have lower HADDOCK score, Vander Waals energy and fewer hydrogen bonds than MSH2_WILD (**Table 5.49**).

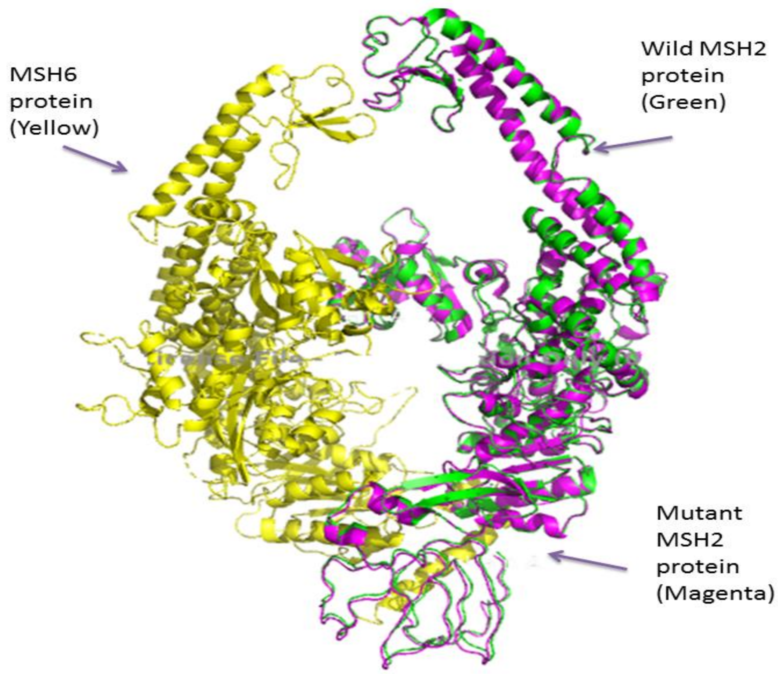

Figure 5.29: MSH2_WILD (Green) docked structure with MSH6 (Yellow) along with MSH2 mutant (Magenta) obtained using Pymol software.

Table 5.49: Docking results of the mutation present in the MSH3/MSH6 interaction domain with MSH2_WILD

Amino acid change	HADDOCK Score	Van der Waals energy	No. of Hydrogen bonds
MSH2_WILD	-160.2 +/- 31.3	-88.1 +/- 20.7	10
L310P	-142.0 +/- 10.8	-76.3 +/- 14.3	8
L330P	-141.8 +/- 12.3	-78.0 +/- 8.5	8
C333R	-150.2 +/- 8.7	-82.1 +/- 6.3	8
C333Y	-154.6 +/- 26.5	-70.9 +/- 14.2	11
G338R	-160.7 +/- 17.0	-64.6 +/- 10.2	12
L341P	-146.4 +/- 12.4	-73.4 +/- 8.9	6
R359I	-159.6 +/- 23.1	-89.2 +/- 18.6	6
R359S	-161.6 +/- 25.5	-94.3 +/- 10.9	14
R524P	-147.6 +/- 26.6	-78.8 +/- 15.5	5
R524L	-156.1 +/- 5.7	-86.6 +/- 10.3	10
G587R	-126.6 +/- 4.3	-70.2 +/- 4.9	6

5.15.2 MSH3 protein:

The Human *MSH3* gene contains a total number of 222,341 bp and MSH3 protein has 1137 amino acids and molecular weight of 127 KDa. SNP database has 79845 SNPs of human *MSH3* gene in May 2020, out of which 414 belongs to the category of clinical significance which were considered for analysis.

5.15.2.1 Identification of deleterious and damaging nsSNP's based on sequence-based homology approach

MSH3 gene was investigated using sequence-based homology techniques (SIFT, PROVEAN, and PANTHER tools) to identify damaging and deleterious nsSNPs. Initially, these SNPs were checked in SIFT where 76 entries were found. Out of 76, 24 rs IDs having non-synonymous (ns) SNPs were found to have either one or two amino acid changes which resulted in 25 SNPs. SIFT predicted 6 and 19 as deleterious and tolerated respectively out of 25 SNPs. These 25 SNPs were further analyzed by PROVEAN and PANTHER for identification of SNPs having deleterious and damaging effects. PROVEAN predicted 6 and 19 as deleterious and tolerated respectively while PANTHER server predicted 25 nsSNPs to be damaging. Interestingly, no SNPs were found to be commonly predicted as deleterious or damaging by all the three tools used (**Table 5.50**). Hence, there was no further bioinformatics analysis done for these SNPs.

Table 5.50: Results of SIFT, PROVEAN, PANTHER tools for MSH3 protein

S.No.	rs IDs	Amino acid substitution	SIFT	PROVEAN	PANTHER
			Prediction	**Prediction**	**Prediction**
1	rs26279	A1045T	Tolerated	Neutral	probably benign
2	rs184967	Q949R	Tolerated	Neutral	probably benign
3	rs1574197	A61P	Tolerated	Neutral	probably benign
4	rs1650697	I79V	Tolerated	Neutral	probably benign
5	rs34058399	E523K	Tolerated	Neutral	probably benign
6	rs35045151	R669W	Deleterious	Deleterious	probably benign
7	rs35121792	A438V	Deleterious	Neutral	probably benign
8	rs55660516	S420G	Tolerated	Neutral	probably benign
9	rs55724159	N524T	Tolerated	Neutral	probably benign
10	rs61749609	I440M	Deleterious	Neutral	probably benign
11	rs76249824	S490Y	Tolerated	Deleterious	probably benign
12	rs138339194	H827R	Deleterious	Deleterious	probably benign
13	rs140543135	F387Y	Tolerated	Neutral	probably benign
14	rs144607594	S119F	Deleterious	Deleterious	probably benign
15	rs144776112	A57P	Deleterious	Neutral	probably benign
16	rs144798521	R454Q	Tolerated	Deleterious	probably benign
17	rs144798521	R454L	Tolerated	Deleterious	probably benign
18	rs146768828	K508E	Tolerated	Neutral	probably damaging

19	rs147289289	R188C	Tolerated	Neutral	probably benign
20	rs148550291	A58P	Tolerated	Neutral	probably benign
21	rs148633216	I914V	Tolerated	Neutral	probably benign
22	rs149350323	Q106E	Tolerated	Neutral	probably benign
23	rs200612739	E725K	Tolerated	Neutral	probably benign
24	rs201149584	A58V	Tolerated	Neutral	probably benign
25	rs372048303	R669Q	Tolerated	Neutral	probably benign

5.15.3 MSH6 protein

The Human MSH6 gene contains a total number of 80098 bp and MSH6 protein has 934 amino acids. SNP database has 68322 SNPs of human MSH2 gene in May 2020, out of which 1820 belongs to the category of clinical significance which was considered for analysis. 226 entries out of 1820 were found to be redundant hence were deleted so the final numbers which were taken for analysis were 1594.

5.15.3.1 Identification of deleterious and damaging nsSNP's based on sequence-based homology approach

SIFT, PROVEAN and PANTHER tools which use sequence-based homology approach to predict the deleterious and damaging nsSNP's were used to analyze 1594 SNPs. Further, SIFT was showing six protein IDs for the same "rs ID" so the result of one protein id (ENSP00000234420) containing the full length (1360 amino acids) was taken. After selecting single protein ID and non-synonymous mutations 84 SNPs are left. Out of 84 SNPs, 80 rs IDs were found to have either one or two amino acid changes which resulted in 85 amino acid changes. 4 rs IDs were found to have two amino acid changes, which resulted in 8 amino acid changes. Out of these 84 nsSNPs, SIFT predicted 31 as deleterious and 53 as tolerated, PROVEAN predicted 32 as deleterious and 52 as tolerated whereas PANTHER server predicted 57 as damaging and 27 as probably benign. Interestingly, 25 nsSNPs were found to be commonly predicted as deleterious by all the three tools.

5.15.3.2 Identification of deleterious and damaging nsSNP's using sequence and structure based approach (PolyPhen):

Polyphen tool predicts if particular SNPs are having deleterious or benign effects. 24 and 23 SNPs out of 25 were predicted to be damaging by HumDiv and HumVar data sets of Polyphen respectively. 23 SNPs were taken into consideration which is predicted by both HumDiv and HumVar data sets (**Table 5.51**). Both datasets of Polyphen have predicted 17 SNPs to be probably damaging having a score range of 0.954 to 1.0 and 6 SNPs were predicted to be probably damaging and possibly damaging by either HumDiv or HumVar data set. Moreover, One SNPs were reported to be benign by Polyphen 2 server is underlined in **Table 5.51**.

Table 5.51: Deleterious SNPs commonly predicted by different tools

S.No.	rs IDs	Amino acid change
1	rs34374438	K854M
2	rs41295268	R468H
3	rs41295278	R1321G
4	rs61754783	M492V
5	rs63749919	Y969C
6	rs63749973	G566R
7	rs63750119	R1242L
8	rs63750138	R772W
9	rs63750253	R1095H
10	rs63750442	T1100R
11	rs63750617	R1076C
12	rs63750741	L449P
13	rs63750836	T1284M
14	rs63750949	T1219I
15	rs63750985	C765W
16	rs63751005	V509A
17	rs63751113	R976H
18	rs63751328	E1193K
19	rs63751405	L435P
20	rs147453999	T1243S
21	rs182024561	L1201F
22	rs191109849	P1082L
23	rs376243329	R1095C
24	rs63750753	P1087H
25	rs63750753	P1087R

*The underline SNPs are not predicted as damaging by Polyphen server.

5.15.3.3 Disease prediction by PhD SNP and SNP&GO

PhD SNP and SNP&GO are used to analyze that single amino acid substitution is associated with the occurrence of a disease or not. 23 nsSNPs which were predicted to be deleterious by both sequence and structure-based homology approach were taken as input for PhD SNP and SNP GO.

PhD SNP and SNP GO server has predicted 16 and 13 nsSNPs, respectively to be associated with disease.

5.15.3.4 Estimating the stability of the mutated protein structures using I-mutant and MuPro server:

Among the 13 nsSNPs commonly predicted to be disease associated by both PhD SNP and SNP&GO, I-mutant predicted all nsSNPs resulted in decreased stability. Only one nsSNPs (rs63750949) have reported an increase in stability by MuPro server rest 12 have reported a decrease in stability. DDG value less than 0 is associated with a decrease in stability and 12 SNPs have reported a decrease in stability by both I Mutant and MuPro server as shown in **Table 5.52**.

Table 5.52: Stability prediction and RMSD value of selected deleterious SNPs

S.No.	rs IDs	AAS	RMSD (Å)	I-Mutant DDG Value	Mu Pro server Delta G value
1	rs63749919	Y969C	0.243	-1.31	-1.00
2	rs63749973	G566R	0.22	-0.59	-0.97
3	rs63750138	R772W	0.232	0	-0.23
4	rs63750253	R1095H	0.248	-1.2	-0.80
5	rs63750442	T1100R	0.241	-0.6	-0.93
6	rs63750617	R1076C	0.205	-1.01	-0.55
7	rs63750741	L449P	0.249	-1.67	-1.75
8	rs63750985	C765W	0.273	0.04	-1.42
9	rs63751328	E1193K	0.24	-0.61	-1.13
10	rs63751405	L435P	0.257	-1.73	-2.03
11	rs182024561	L1201F	0.225	-0.84	-1.14
12	rs376243329	R1095C	0.211	-0.87	-0.52

5.15.3.5 Modeling of the complete MSH6 protein using MODELLER v9.22

Homology modeling approach was used to build MSH6 wild protein structure as complete MSH6 protein structure was not available in PDB database. BLAST results showed that structures (1EWQ, 1FW6, 1NNE and 2O8B) have maximum coverage and percentage similarity with the query. Hence these were selected to be used as templates for creating MSH6 structure and named as MSH6_WILD. The complete structure of MSH6 was then generated by using MODELLER v9.22 which is shown in **Figure 5.30**. Furthermore, this MSH6_WILD structure was submitted to energy minimization to find a set of coordinates representing the minimum energy conformation of the structure. For energy minimization CHIMERA tool was used which utilizes steepest descent, and conjugate gradient optimization algorithms for energy minimization. After energy minimization of MSH6_WILD protein, the energy got significantly reduced to 49992.2857 KJ/mol from 1626512.431 KJ/mol (lower is better) which signifies that all the forces on the atoms in that protein molecule are balanced. Mutant protein structures of shortlisted 12 nsSNPs were also generated by MODELLER 9.22 using previously generated energy minimized MSH6_WILD structure with appropriate mutations as template. These generated structures were also subjected to energy minimization. These structures were further subjected to Ramachandran plot analysis via PROCHECK. Ramachandran plot results for MSH6_WILD represent 96.3% residues in the allowed region.

Figure 5.30: Complete structure of Energy minimized MSH6_wild

5.15.3.6 Root Mean Square Deviation value estimation of the modeled protein:

The structure of the mutant proteins with selected 12 nsSNPs was modeled by taking the previously generated MSH6_WILD structure as a template with appropriate mutations. RMSD values were calculated by utilizing align facility in PYMOL (**Table 5.52**). It appears there is not much structural deviation due to these 12 SNPS as the RMSD value is less than 2.7 Å.

5.15.3.7 Mapping of the missense mutations in to MSH6 protein domains

There are seven domains in MSH6 viz. PCNA binding motif, PWWP domain, and domain I-V these 12 SNPs were found to be distributed in domain I-V (**Figure 5.31**). Two SNPs were found in Domain I, One SNP was found in domain II, two SNPs were located in domain III, one and five SNPs were found in domain IV and V respectively.

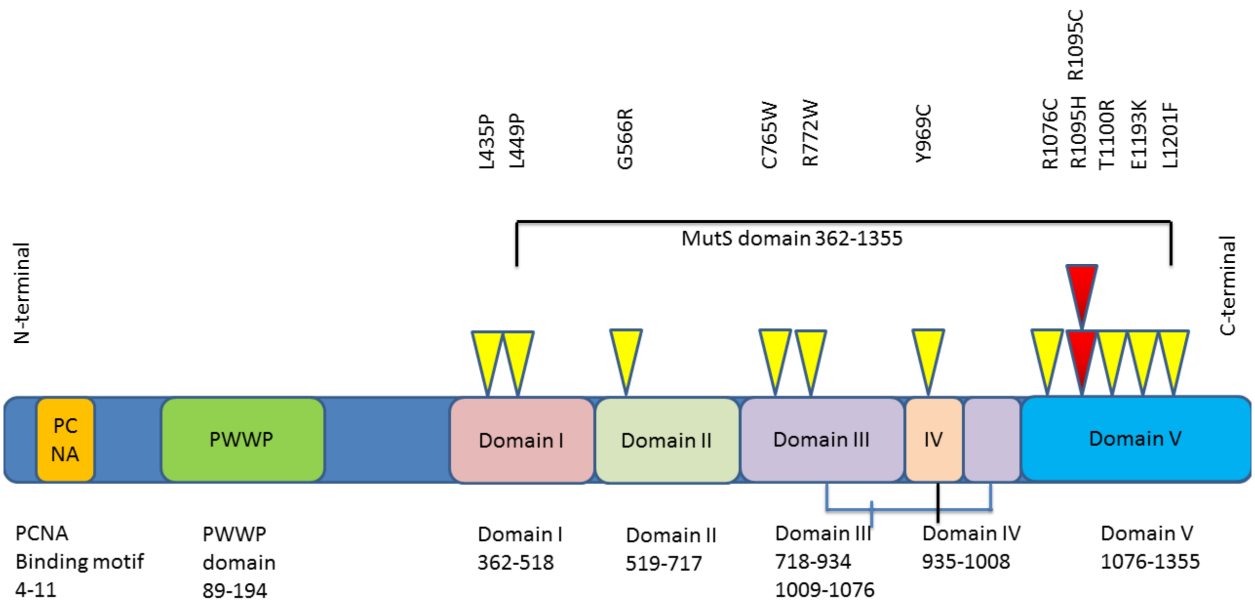

Figure 5.31: Different domains of MSH6 protein

5.15.3.8 SNPs affecting the interaction of MSH6 with MSH2

MSH6 interact with MSH2 to form MutSα in humans, and interaction sites of MSH6 were spread throughout the amino acid sequence hence complete MSH6 molecule is selected for docking analysis with MSH2 protein. When molecular docking of MSH2 was performed with wild and mutant MSH6 protein, a clear difference in HADDOCK score, no. of hydrogen bonds and Van der Waals energy was observed in few SNPs (**Figure 5.32**). Two mutant structures (R1076C and R1095C) have higher HADDOCK score, Van der Waals energy and higher no. of hydrogen bonds than wild type structure. In two mutant structure (R772W and L449P), the number of hydrogen bonds is less as compared to wild type corresponding decrease in HADDOCK score (**Table 5.53**).

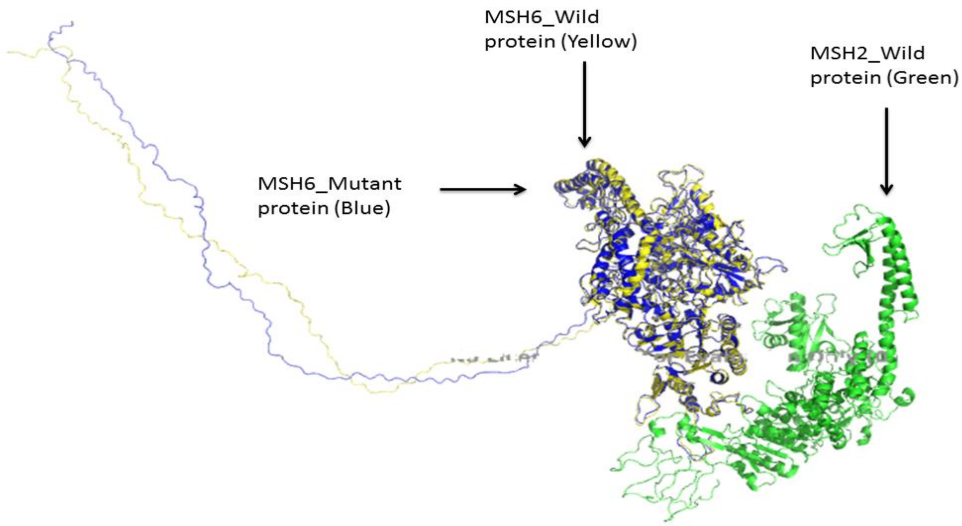

Figure 5.32:MSH6_WILD (Yellow) docked structure with MSH2_WILD (Green) along with MSH6 mutant (Blue) obtained using Pymol software.

Table 5.53: Docking results of the mutation present in the MSH3/MSH6 interaction domain with MSH2_WILD

Amino acid change	HADDOCK Score	Van der Waals energy	No. of Hydrogen bonds
MSH6_WILD	-94.8 +/- 9.6	-38.7 +/- 10.8	16
Y969C	-98.4 +/- 4.4	-41.5 +/- 5.8	16
G566R	-93.5+/-3.0	-42.9 +/- 5.6	12
R772W	-91.9 +/- 19.8	-40.3 +/- 19.4	4
R1095H	-90.2 +/- 10.9	-29.0 +/- 6.3	16
T1100R	-94.6 +/- 8.8	-34.7 +/- 7.8	16
R1076C	-113.5 +/- 23.4	-53.7 +/- 17.4	19
L449P	-91.2 +/- 7.5	-34.9 +/- 8.3	8
C765W	-98.5 +/- 4.2	-39.7 +/- 6.7	16
E1193K	-99.2 +/- 2.8	-36.9 +/- 1.9	16
L435P	-99.5 +/- 2.9	-42.5 +/- 6.0	16
L1201F	92.5+/- 2.5	-42.9 +/- 5.6	11
R1095C	-110.5 +/- 25.2	-54.1 +/- 17.0	18

5.15.4 MLH1 protein

The Human *MLH1* gene contains a total number of 57360 bp and MLH1 protein has 756 amino acids. SNP database has 15721 SNPs of human MLH1 gene in May 2020, out of which 1549 belongs to the category of clinical significance which was considered for analysis.

5.15.4.1 Identification of deleterious and damaging nsSNP's based on sequence-based homology approach

Bioinformatics tools utilizing sequence-based homology techniques (SIFT, PROVEAN and PANTHER) were used for the analysis of *MLH1* gene (1549 SNPs). Further, SIFT was showing multiple protein IDs for the same "rs ID" so the result of one protein id (ENSP00000231790) containing the full length (756 amino acids) was taken. After selecting single protein ID and non-synonymous mutations 287 SNPs were left. Out of 287 SNPs, 256 rs IDs having non-synonymous (ns) SNPs were found to have one, two or four amino acid changes. SIFT predicted 184 and 103 as deleterious and tolerated respectively out of 287 SNPs. These 287 SNPs were further analyzed by PROVEAN and PANTHER for identification of SNPs having deleterious and damaging effects. PROVEAN predicted 198 and 89 as deleterious and tolerated respectively while PANTHER server predicted 258 nsSNPs to be damaging and 29 as probably benign. Interestingly, 170 SNPs were found to be commonly predicted as deleterious or damaging by all the three tools used.

5.15.4.2 Identification of deleterious and damaging nsSNP's using sequence and structure based approach (PolyPhen):

Polyphen tool predicts if particular SNPs are having deleterious or benign effects. 160 and 155 SNPs out of 170 were predicted to be damaging by HumDiv and HumVar data sets of Polyphen respectively. 154 SNPs were taken into consideration which is predicted by both HumDiv and HumVar data sets (**Table 5.54**). Both datasets of Polyphen have predicted 127 SNPs to be probably damaging having a score range of 0.958 to 1.0 and 27 were predicted to be probably damaging or possibly damaging by either HumDiv or HumVar data set. Eight SNPs were reported to be benign by Polyphen 2 server is underlined in **Table 5.54**.

Table 5.54: Deleterious SNPs commonly predicted by different tools

S. No.	rs IDs	Amino acid substitution	S. No.	rs IDs	Amino acid substitution
1	rs2020873	H718Y	86	rs63750891	L155R
2	rs4986984	R217C	87	rs63750899	P648S
3	rs28930073	D132H	88	rs63750952	N64S
4	rs35001569	K618E	89	rs63750977	P141A
5	rs35045067	Y646C	90	rs63750993	D304V
6	rs41295280	G22A	91	rs63751012	E37K
7	rs41295282	S93G	92	rs63751021	S193P
8	rs41295284	L607H	93	rs63751049	V326A
9	rs56185292	S577L	94	rs63751069	T81I
10	rs56250509	L259S	95	rs63751094	D41G
11	rs61751644	R389W	96	rs63751109	S44F
12	rs63749792	P640S	97	rs63751191	K70N
13	rs63749803	H109Q	98	rs63751194	R265S
14	rs63749838	I25F	99	rs63751194	R265C
15	rs63749859	C77R	100	rs63751202	L653R
16	rs63749906	M35R	101	rs63751242	N158H
17	rs63749909	V506A	102	rs63751275	R687W
18	rs63749990	F80V	103	rs63751283	L260R
19	rs63750005	T82I	104	rs63751428	Q62K
20	rs63750016	A589D	105	rs63751448	R265H
21	rs63750059	L559R	106	rs63751465	G65D
22	rs63750062	I565F	107	rs63751597	H264Y
23	rs63750098	V49E	108	rs63751598	S295G
24	rs63750102	V180G	109	rs63751608	L574P
25	rs63750144	S295N	110	rs63751609	Q346R
26	rs63750144	S295T	111	rs63751616	L582P
27	rs63750206	G67R	112	rs63751630	R522L
28	rs63750206	G67W	113	rs63751664	H264R
29	rs63750211	R182G	114	rs63751665	E102D
30	rs63750216	A29G	115	rs63751682	E663A
31	rs63750242	L676R	116	rs63751682	E663G
32	rs63750266	R100Q	117	rs63751684	P581L
33	rs63750266	R100P	118	rs63751701	C39Y
34	rs63750271	N551T	119	rs63751703	R522W
35	rs63750281	I68N	120	rs63751711	R226L
36	rs63750289	L549P	121	rs63751713	L582F
37	rs63750297	S106R	122	rs104894997	N263T
38	rs63750303	G244D	123	rs138584384	R725C
39	rs63750303	G244V	124	rs138705565	R27Q
40	rs63750395	K286Q	125	rs138705565	R27P
41	rs63750430	R385H	126	rs143009528	Y379C

42	rs63750430	R385P	127	rs147342421	Q48H
43	rs63750437	C77Y	128	rs147939838	R474W
44	rs63750447	V384D	129	rs148317871	G738E
45	rs63750449	K618T	130	rs151119913	H315Y
46	rs63750453	E102K	131	rs182963667	F88L
47	rs63750465	T116K	132	rs200076893	Y126N
48	rs63750465	T116R	133	rs200830026	R497W
49	rs63750507	I107R	134	rs201931669	Y283C
50	rs63750511	Q542P	135	rs267607706	N38K
51	rs63750511	Q542L	136	rs267607707	I36N
52	rs63750514	I25T	137	rs267607707	I36T
53	rs63750517	L292P	138	rs267607713	D41H
54	rs63750539	A111V	139	rs267607725	G98S
55	rs63750559	V113D	140	rs267607725	G98R
56	rs63750575	L588P	141	rs267607726	G101S
57	rs63750580	N38H	142	rs267607727	G101D
58	rs63750580	N38D	143	rs267607730	F99L
59	rs63750584	L272S	144	rs267607741	Y126C
60	rs63750587	A586D	145	rs267607747	G147R
61	rs63750610	P648L	146	rs267607766	S191I
62	rs63750638	Q149H	147	rs267607776	I216S
63	rs63750641	K84E	148	rs267607808	P309S
64	rs63750642	L260F	149	rs267607813	V303E
65	rs63750648	I19F	150	rs267607829	L509F
66	rs63750650	E268G	151	rs267607843	A539D
67	rs63750693	L622H	152	rs267607847	Y561H
68	rs63750693	L622P	153	rs267607848	F568I
69	rs63750706	A21V	154	rs267607861	E600G
70	rs63750718	D601G	155	rs267607864	A608D
71	rs63750726	P654L	156	rs267607865	L585R
72	rs63750760	R385C	157	rs267607866	A619P
73	rs63750781	T117R	158	rs267607875	P640L
74	rs63750781	T117M	159	rs267607876	F656S
75	rs63750792	P28L	160	rs267607885	D737V
76	rs63750796	E319K	161	rs267607886	Y684C
77	rs63750809	C680R	162	rs267607887	W666R
78	rs63750809	C680G	163	rs267607894	L749Q
79	rs63750825	L636P	164	rs267607894	L749P
80	rs63750834	P169L	165	rs267607895	R755S
81	rs63750846	S627T	166	rs267607900	R755W
82	rs63750850	D63N	167	rs367654552	R18C
83	rs63750864	A681V	168	rs368208495	S106N
84	rs63750866	A128P	169	rs374380262	D748H
85	rs63750876	P603R	170	rs377241633	F626L

*The underline SNPs are not predicted as damaging by Polyphen server.

5.15.4.3 Disease prediction by PhD SNP and SNP&GO

PhD SNP and SNP&GO are used to analyze that single amino acid substitution is associated with the occurrence of a disease or not. 154 nsSNPs which were predicted to be deleterious by both sequence and structure-based homology approach were taken as input for PhD SNP and SNP GO. PhD SNP and SNP GO server has predicted 132 and 106 nsSNPs, respectively to be associated with disease.

5.15.4.4 Estimating the stability of the mutated protein structures using I-mutant and MuPro server

Among the 106 nsSNPs commonly predicted to be disease associated by both PhD SNP and SNP&GO, we have considered 31 nsSNPs which lie in the PMS1/PMS2/MLH3 interaction domain for stability analysis. Out of 31 nsSNPs I-mutant predicted only one nsSNP (*rs2020873*) with increased stability and MuPro server predicted 2 nsSNPs (*rs63750511*, *rs63750726*) with increased stability. DDG value less than 0 is associated with a decrease in stability and 28 SNPs have reported a decrease in stability by both I Mutant and MuPro server as shown in **Table 5.55**. Further analysis revealed that these 28 SNPs were belonging to 27 rs IDs in which 26 and 1 are having one and two mutation respectively (**Table 5.55**).

Table 5.55: Stability prediction and RMSD value of selected deleterious SNPs

S.No.	rs IDs	AAS	RMSD (Å)	I-Mutant DDG Value	Mu Pro server Delta G value
1	rs35045067	Y646C	0.204	-0.93	-1.17
2	rs63749792	P640S	0.188	-1.7	-1.17
3	rs63749909	V506A	0.227	-1.41	-1.37
4	rs63750059	L559R	0.185	-1.57	-1.88
5	rs63750242	L676R	0.202	-1.32	-1.71
6	rs63750289	L549P	0.199	-1.08	-1.83
7	rs63750575	L588P	0.21	-1.5	-2.56
8	rs63750587	A586D	0.197	-0.58	-1.05
9	rs63750610	P648L	0.199	-0.75	-0.26
10	rs63750825	L636P	0.209	-1.43	-1.93
11	rs63750864	A681V	0.21	-0.17	-0.74
12	rs63750899	P648S	0.191	-1.49	-0.93
13	rs63751202	L653R	0.187	-1.85	-1.81
14	rs63751616	L582P	0.201	-1.47	-1.60
15	rs138584384	R725C	0.199	-1.02	-0.73

16	rs267607829	L509F	0.209	-0.89	-0.78
17	rs267607843	A539D	0.229	-1.69	-1.09
18	rs267607847	Y561H	0.339	-0.84	-0.30
19	rs267607864	A608D	0.2	-0.97	-1.51
20	rs267607875	P640L	0.219	-1.21	-0.87
21	rs267607876	F656S	0.198	-0.31	-0.47
22	rs267607885	D737V	0.185	0.18	-0.42
23	rs267607886	Y684C	0.215	-0.98	-0.42
24	rs267607887	W666R	0.2	-0.97	-1.51
25	rs377241633	F626L	0.196	-1.01	-0.77
26	rs63750511	Q542P	0.223	-0.27	-1.08
27	rs63750693	L622H	0.206	-2.16	-1.62
28	rs63750693	L622P	0.218	-1.7	-2.02

5.15.4.5 Modeling of the complete MLH1 protein using MODELLER v9.22

Homology modeling approach was used to build MLH1 wild protein structure as complete MLH1 protein structure was not available in PDB database. BLAST results showed that structures (1B62, 3RBN, 4P7A and 5AKB) have maximum coverage and percentage similarity with the query. Hence these were selected to be used as templates for creating MLH1 structure and named as MLH1_WILD. The complete structure of MLH1 was then generated by using MODELLER v9.22 which is shown in **Figure 5.33**. Furthermore, this MLH1 _WILD structure was submitted to energy minimization to find a set of coordinates representing the minimum energy conformation of the structure. For energy minimization CHIMERA tool was used which utilizes steepest descent, and conjugate gradient optimization algorithms for energy minimization. After energy minimization of MLH1_WILD protein, the energy got significantly reduced to 28571.52 KJ/mol from 46361738.41KJ/mol (lower is better) which signifies that all the forces on the atoms in that protein molecule are balanced. This structure was further subjected to Ramachandran plot analysis via PROCHECK. Ramachandran plot results represent 93.4% residues in the allowed region.

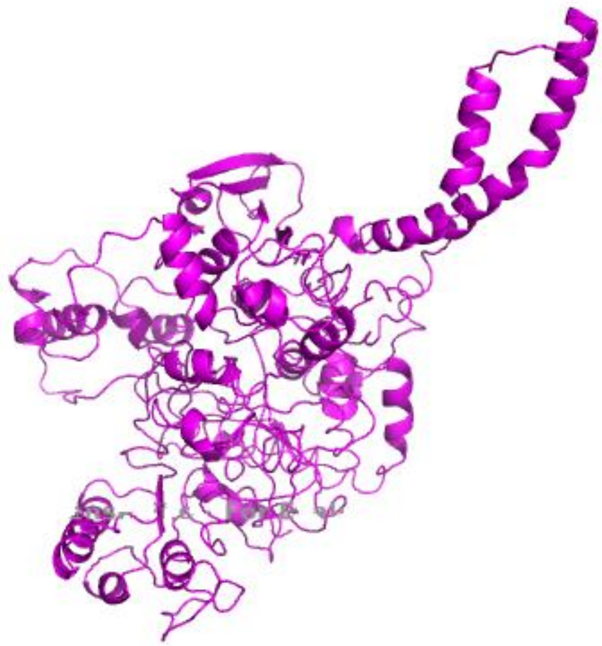

Figure 5.33: Complete structure of Energy Minimized MLH1_Wild

5.15.4.6 Root Mean Square Deviation value estimation of the modeled protein

The structure of the mutant proteins with selected 27 nsSNPs (28 amino acid change) was modeled by taking the previously generated energy minimized MLH1 structure as a template with appropriate mutations. RMSD values were calculated by utilizing align facility in PYMOL. The RMSD values of all nsSNPs ranges between 0.185 Å and 0.339 Å (**Table 5.55**). Further, energy minimization of mutant MLH1 protein was also done by Chimera tool using steepest descent and conjugate gradient algorithms.

5.15.4.7 Mapping of the missense mutations in to MLH1 protein domains

There are three domains in MLH1 viz. ATPase domain, MutS homologs interaction domain and PMS1/PMS2/MLH3 interaction domain. For this study we have shortlisted only those SNPs which lies in the PMS1/PMS2/MLH3 domain (492-742). 27 SNPs with 28 amino acid change were located in this domain as shown in **Figure 5.34**.

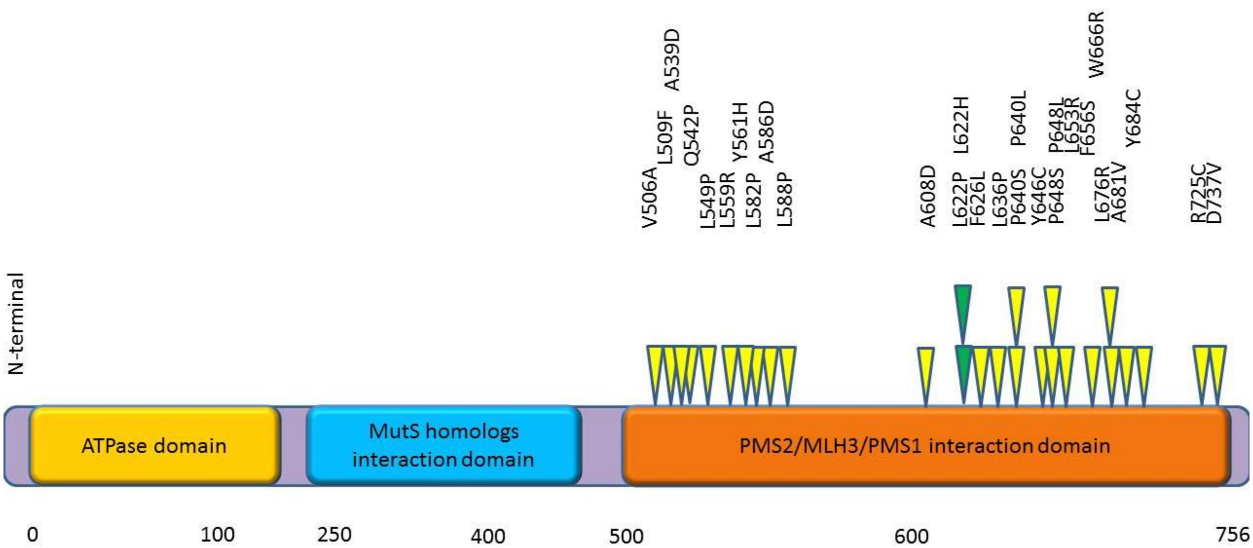

Figure 5.34: Different domains of MLH1 protein

5.15.4.8 SNPs affecting the interaction of MLH1 with MLH3

MLH1 interacts with MLH3 to form MutLα which is plays an important role in DNA mis-match repair pathways hence the docking study of SNP present in this domain was carried out. Molecular docking of MLH3 protein with wild and mutant MLH1 protein depicts a difference in HADDOCK score, hydrogen bonds and Van der Waals energy. The HADDOCK score is **a** weighted sum of a variety of energy terms including van der Waals, electrostatic, desolvation, and restraint violation energies. The docked complex of MLH1_wild and mutant with MLH3 is represented in **Figure 5.35.** Overall HADDOCK score of wild MLH1 protein was -78.4 and the lowest HADDOCK score for MLH1 mutant was -40.9 for Y561H indicating a change in binding affinity of MLH1_Y561H towards MLH3. Complete details of the HADDOCK score, Van der Waals energy, and Number of hydrogen bonds of mutant structures used in the docking study are shown in **Table 5.56**. We found a noteworthy association between the HADDOCK score and hydrogen bonds present in the docked complexes. Our results conclude that as the HADDOCK score increases the number of hydrogen bonds also increases signifying higher overall stability of the mutant proteins. P640S, V506A, L559R, L588P, A681V, L653R, R725C, A539D, A608D, P640L, D7377V, and L622P have higher HADDOCK score and more number of hydrogen bonds than the wild structure.

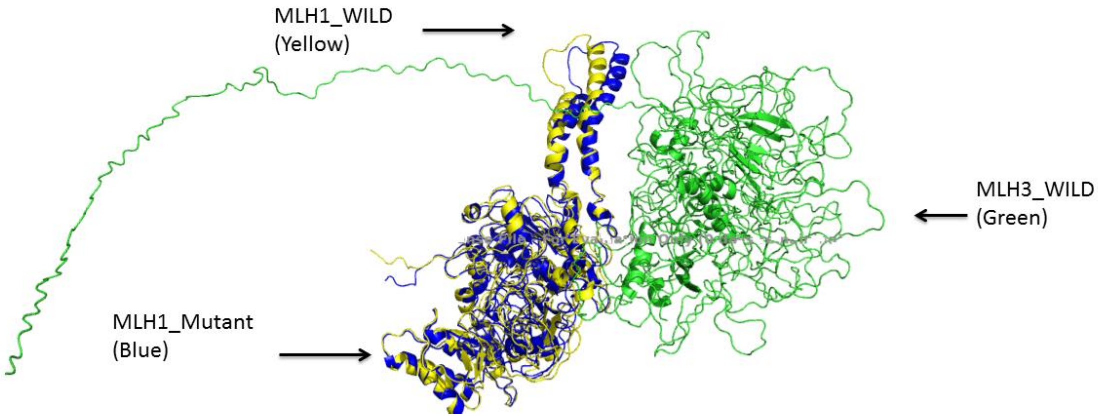

Figure 5.35: MSH6_WILD (Yellow) docked structure with MSH2_WILD (Green) along with MSH6 mutant (Blue) obtained using Pymol software.

Table 5.56: Docking results of the mutation present in the MLH3 interaction domain with MLH1_WILD

Amino acid change	HADDOCK Score	Van der Waals energy	No. of Hydrogen bonds
MLH1_WILD	-78.4 +/- 4.8	-84.8 +/- 5.5	8
Y646C	-49.1 +/- 2.2	-64.0 +/- 3.7	11
P640S	-96.6 +/- 15.6	-71.8 +/- 12.5	17
V506A	-86.3 +/- 12.2	-80.0 +/- 11.6	13
L559R	-109.0 +/- 11.2	-56.5 +/- 10.1	13
L676R	-50.7 +/- 18.5	-67.9 +/- 10.2	8
L549P	-67.9 +/- 19.8	-52.5 +/- 7.5	19
L588P	-111.2 +/- 21.2	-92.3 +/- 12.1	14
A586D	-80.3 +/- 6.1	-70.9 +/- 9.1	8
P648L	-71.3 +/- 16.7	-50.3 +/- 15.7	14
L636P	-58.8 +/- 7.2	-42.9 +/- 11.1	8
A681V	-122.0 +/- 35.0	-78.2 +/- 9.6	16
P648S	-65.5 +/- 13.3	-56.7 +/- 4.0	7
L653R	-83.5 +/- 18.2	-76.8 +/- 21.0	12
L582P	-49.6 +/- 14.4	-61.7 +/- 9.2	11
R725C	-102.7 +/- 20.9	-71.3 +/- 14.4	11
L509F	-76.2 +/- 6.8	-61.5 +/- 13.2	8
A539D	-90.4 +/- 16.4	-64.9 +/- 15.7	21
Y561H	-40.9 +/- 14.1	-45.3 +/- 11.1	12
A608D	-95.6 +/- 21.5	-83.9 +/- 9.1	13
P640L	-78.5 +/- 17.8	-54.5 +/- 13.0	20

F656S	-50.8 +/- 19.6	-50.9 +/- 8.0	13
D737V	-87.5 +/- 15.1	-66.8 +/- 4.4	18
Y684C	-59.2 +/- 8.5	-53.7 +/- 6.8	6
W666R	-57.2 +/- 8.1	-57.5 +/- 6.7	9
F626L	-76.4 +/- 7.7	-66.2 +/- 8.8	9
Q542P	-71.7 +/- 19.6	-62.1 +/- 10.6	10
L622H	-84.4 +/- 10.7	-73.4 +/- 3.1	7
L622P	-91.3 +/- 7.5	-61.9 +/- 8.2	17

Chapter 6

Discussion

Epidemiological studies performed with a large population size assist us in understanding the relationship between DNA repair genes and the risk of developing cancer. These studies further aid us to conclude that single nucleotide polymorphisms in the candidate gene influence the advent of disease regardless of environmental variations and regulatory pathways. These also help increase our understanding of various dimensions of public health due to variations in DNA repair genes.

The current study focuses on different polymorphic variants of Mismatch repair pathway genes, which change DNA repair efficiency and influence lung cancer patients' susceptibility and survival. Four significant MMR genes, i.e., *MLH1, MSH2, MSH3, and MSH6* having one, six, one, and three single nucleotide polymorphisms, are evaluated for their influence on individuals' predisposition towards lung cancer. These polymorphic variants were also investigated to evaluate their role in predicting the prognosis and survival of lung cancer patients of North Indian origin. This investigation further also evaluated whether the polymorphisms in DNA mismatch repair genes were associated with increased toxicity in lung cancer patients undergoing platinum-based doublet chemotherapy. This is the first study reporting such a large sample size in the north Indian populace evaluating the effect of polymorphisms in the Mismatch DNA repair pathway.

6.1 *MLH1* gene (*rs1800734*) and its role in lung cancer predisposition

The mismatch repair (MMR) pathway is responsible for recognizing and repairing the erroneous insertion, misincorporation, and deletion of the bases during the process of replication and recombination. Mutations in DNA repair pathways have been associated with developing various types of cancer (Winsey *et al.*, 2000, Lawania *et al.*, 2019). To date, several studies have addressed whether some genetic variation (SNPs) affects various clinical outcomes in lung cancer patients (Singh *et al.*, 2019, Slováková *et al.*, 2014). This study is focused on finding whether *MLH1* (*rs1800734*) polymorphism plays any role in modulating the risk for lung cancer, and we also attempted to evaluate the impact of the *MLH1 -93G>A* polymorphism on the outcome of lung cancer patients treated with platinum-based doublet chemotherapy.

Our study concluded no significant association between *MLH1 -93G>A* polymorphism and risk of developing lung cancer when all the three genetic models were evaluated. Our results follow earlier studies conducted that have reported no association between the *MLH1 -93G>A* genotype and the risk of prostate cancer in the Iranian population (Khooshemehri *et al.*, 2020). However, a study conducted by Slováková and colleagues on the Slovakia populace reported a 1.4-fold higher likelihood of lung cancer (Slováková *et al.*, 2014). Furthermore, when we analyzed our data based on histology, we observed that lung cancer subjects were at a reduced risk of developing ADCC (OR=0.48**, p=0.0007** & 0.88, **p=0.008**) when the codominant, dominant models were applied. On the contrary, when the recessive model was evaluated, we observed a 1.8-fold risk for ADCC (**p =0.02**). In the case of SCLC, subjects harboring a single mutant allele had a higher propensity towards SCLC (**p=0.03**). Our results were following the previous studies, which reported an association for the occurrence of LC based on histology. Shih and co-workers have reported that individuals having *MLH1* polymorphism have an about 2-fold increased risk of developing ADCC.

Cigarette smoking is another risk factor related to the development of lung cancer. Hence in order to understand if there is an association of smoking and its synergistic interaction with the *MLH1-93G>A* polymorphism for the development of lung cancer, we stratified patients into two groups: "smokers" and "non-smokers." Further 'smokers' group was divided into two subgroups: Light smokers and heavy smokers based upon the pack-years smoked. when we estimated the role between heavy and light smokers, it was evident that subjects who were smoking more than 25 pack-years and harboring a single copy of the variant allele (*GA*) for the *MLH1 -93G>A* polymorphism had a 2-fold high risk for lung cancer predisposition in the co-dominant model (***p*=0.001**). A similar observation was also evident in those subjects carrying a single mutant allele for the *-93G>A* polymorphism where a 1.8-fold increase risk was observed. Earlier studies conducted also corroborated our results. Shih and co-workers reported a 1.5-fold higher risk of developing LC. However, Shih has also reported a 4-fold increased risk for lung cancer with non-smokers for the above polymorphism (Shih *et al.*, 2010).

We have also analyzed our data after dividing the study based on gender, and our results conclude that female subjects were at low risk of developing lung cancer compared to males (***p*=0.0001**). Our results were following a previous study done in the Slovak population, where the authors have reported a significantly low risk of occurrence of LC in females (Slováková *et al.,* 2014). It is very enticing to propose that reduced risk of LC in females is an indicator of a relation between rs1800734 polymorphism, lung cancer, and hormonal stimulation. The *MLH1 -93G>A* polymorphism lies in the promoter region, with a semi-responsive estrogen element (Irving *et al.,* 2002). Further, Belcher and colleagues have shown a relationship between tumors not associated with reproductive organs and hormones, especially estrogen (Belcher *et al.,* 2009).

6.1.1 Relation of *MLH1* polymorphism (rs1800734) with clinic-pathological features and prognosis

We have also analyzed the implication of *MLH1 -93G>A* polymorphism based on various clinical-pathological features like the clinical stage of the tumor, its size, invasion in the lymph node, and metastasis. Our data does not follow previous studies. Shih and co-workers have predicted a 2-fold increased risk of developing stage III or IV lung cancer in the Taiwanese populace (Shih *et al.,* 2010).

This study also evaluated any relationship between the *MLH1 -93G>A* polymorphism and OS of lung cancer patients undergoing platinum-based chemotherapy. All the LC patients were either given docetaxel, paclitaxel, irinotecan, or pemetrexed along with carboplatin/cisplatin as advised by the clinician. Our results showed that the patients given carboplatin/cisplatin and docetaxel and carrying the single variant allele for *MLH1* polymorphism exhibited the worst prognosis (MST=8.4; HR=3.13**; *p*=0.04**).

6.1.2 *MLH1* polymorphism and its relationship with toxicity

Genetic polymorphisms in the DNA mismatch repair pathway are a binate sword for estimating chemotherapy efficiency. Proficient mismatch repair would safeguard tissues from cancer, leading to increased prognosis but may lead tumor towards chemotherapy resistance (Gazdar *et al.,* 2007). Conversely, excellent DNA repair in normal cells shall increase damage to mutagenesis due to platinum and diminished toxicity. DNA mismatch repair activity is critical to repair normal cells during chemotherapy and avert overwhelming treatment-related adverse events.

This investigation evaluated whether the polymorphisms in *MLH1* gene was associated with increased toxicity in lung cancer patients undergoing platinum-based doublet chemotherapy. We reported that variant heterozygote of *MLH1 -93G>A* polymorphism was significantly associated with severe grade gastrointestinal toxicity (diarrhea). Our results further reported no association between *MLH1 (rs1800734)* genetic polymorphic variant and haematological toxicity, which is following the previous study conducted by Noguiera and co-workers on head and neck squamous cell carcinoma patients of Brazilian origin (Nogueira *et al.,* 2018).

6.2 *MSH3* gene (*rs26279*) and its role in lung cancer predisposition

This case-control study is focused on finding whether *MSH3 Ala*^{1045}*Thr* (*rs26279*) genetic polymorphisms play any role in modulating the risk for lung cancer, and we also examined the impact of this polymorphism on the outcome of lung cancer patients with platinum-based doublet chemotherapy. Data from our study suggest a lack of any noteworthy association between *MSH3 Ala*^{1045}*Thr* polymorphism and risk of developing lung cancer. Our results here are supported by an earlier study that reported no association between *MSH3* (*rs26279*) polymorphism and risk for lung cancer (Xu *et al.,* 2015). Smith and co-workers have also reported no association between *rs26279* polymorphism and susceptibility towards breast cancer in the Caucasian population (Smith *et al.,* 2008). However, a few studies have found an association between *MSH3 Ala*^{1045}*Thr* polymorphism with an increased propensity towards colorectal and breast cancer (Berndt *et al.,* 2007, Tulupova *et al.,* 2008).

Another well-known risk factor for lung cancer is tobacco smoking; therefore, to evaluate the synergistic role of smoking and *MSH3* polymorphisms towards lung cancer susceptibility, we stratified our data based on smoking status to study the gene-environment interaction. Our data suggest that smoking status does not exhibit any association between *MSH3 Ala*^{1045}*Thr* polymorphism and the risk of developing lung cancer. However, a study conducted by Vogelsang and colleagues reported a significant increase in promoter methylation of *MSH3* (91.9%) and further concluded that the factors responsible for this increase are also responsible for increased risk of esophageal cancer (Vogelsang *et al.,* 2014). They tested 17 *MSH3* related CpG sites, and methylation level at the Cg16401290 site located in the *MSH3* promoter region reported a higher methylation level than normal tissues**.** Further, we have also analyzed if gender is associated with the risk of developing lung cancer. For *MSH3 Ala*^{1045}*Thr* polymorphism, heterozygous type genotype (*GA*) in the codominant model of females has a 2-fold increased risk of developing lung cancer. However, a previous study conducted by Conde and co-workers differed from our results and concluded that there is no relationship of *rs26279* polymorphism and susceptibility towards breast cancer in Caucasian females (Conde *et al.,* 2009).

6.2.1 *MSH3* polymorphism and its relationship with toxicity

This investigation evaluated whether the polymorphisms in *MSH3* gene was associated with increased toxicity in lung cancer patients undergoing platinum-based doublet chemotherapy. Our investigation reported that for MSH3 *rs26279* “A” allele carriers reported poor hematological toxicity tolerance in patients undergoing platinum-based doublet chemotherapy. *MSH3* promoter, if methylated, is reported to be associated with oesophageal cancer, suggesting that it is a critical factor in modulating the chemosensitivity of cells. Further, MSH3 dimerizes with MSH2 to form MutSβ heteroduplex, which serves as an ATPase critical in mismatch identification and initiation of repair. It recognizes 2-13 bp insertion-deletion loops and binds to them to start DNA repair (Doherty *et al.,* 2013). *rs26279* SNP has been found to alter the function of MSH3; nevertheless, the fundamental mechanism is still not precise.

6.3 *MSH6* (*rs3136228*) polymorphism and its role in lung cancer predisposition

For *MSH6* (*rs3136228*) polymorphism, our results suggest an increased susceptibility towards lung cancer in subjects harboring the *TT* genotype (AOR=1.43; **p=0.03**) when the recessive genetic model was applied. Our findings were further corroborated by results shown in a study conducted by Tulupova and colleagues in reporting an increase in susceptibility towards lung cancer in the Czech Republic populace (Tulupova *et al.,* 2008). We further reported an increased risk of developing ADCC in candidates carrying TT genotype (AOR=1.75; **p=0.03**).

Another well-known risk factor for lung cancer is tobacco smoking; therefore, to evaluate the synergistic role of smoking and *MSH6* polymorphisms towards lung cancer susceptibility, we stratified our data based on smoking status to study the gene-environment interaction. For *MSH6* polymorphism, our data suggest that individuals carrying *GT* genotype were observed to incur a lower risk of lung cancer development **(p=0.01**). However, we could not find any study demonstrating any association among *MSH6* variants, lung cancer susceptibility, and smoking status. Further, we have also analyzed if gender is associated with the risk of developing lung cancer. For *MSH6 557G>T* polymorphism, our data show no association of gender with lung cancer susceptibility, but a study conducted by Carrera and co-workers reported an increased risk of lung cancer in males of Spanish Caucasian origin (Carrera *et al.,* 2017).

6.3.1 Relation of *MSH6* (*rs3136228*) polymorphism with clinic-pathological features and prognosis

We also investigated the role of *MSH6 557G>T* polymorphism based on clinical-pathological features such as stage, tumor size, lymph node invasion, and metastasis. Our data suggest no association between clinical-pathological features and lung cancer concerning *MSH6 557G>T* polymorphisms. We have further analyzed the effect on overall survival (OS) due to *MSH6 557G>T* polymorphism and we reported a significant decrease in median survival time in subjects carrying *GT* genotype (**log rank p=0.03**). However, one study conducted by zanussu and co-workers on prostate cancer patients reported an increase in OS in patients with at least one T allele in the Italian population (Zanusso *et al.*, 2017). This study also evaluated the association of overall survival in lung cancer subjects treated with platinum-based doublet chemotherapy. Our results suggested that no association with OS whether the patient was given Paclitaxel, Irinotecan, or Pemetrexed along with carboplatin/cisplatin, further a 2-fold higher hazard ratio (**p=0.03**) for patients undergoing docetaxel + carbo/cisplatin combination was observed. However, we could not find any study investigating the role of *MSH6* polymorphism in the overall survival of lung cancer patients concerning chemotherapy.

6.4 MSH6 polymorphism (*rs1800932* and *rs1042821*) with clinic-pathological features and prognosis

Data from our research revealed that both *MSH6 Pro^{92}Pro* (*rs1800932*) and *Gly^{39}Glu* (*rs1042821*) polymorphisms did not show any significant association towards stage III and IV lung tumors and also for lymph node invasion and metastasis. Our study concludes that neither of the two polymorphisms showed any significant association with the overall survival of lung cancer patients. However, when classified based on histological subtypes, SCLC patients who were heterozygous for the *Pro^{92}Pro* (*rs1800932*) polymorphism reported a significant increase in SCLC patients (**p=0.03**). Our results were corroborated by a previous study, where Vymetalkova and co-workers reported no association of *rs1800932* polymorphism in colorectal cancer patients with overall survival in the Czech Republic population (Vymetalkova *et al.,* 2014). However, another study conducted by Dong and co-workers on pancreatic cancer patients contradicted our findings and reported a 2-fold increased risk and a corresponding decrease in survival time in American lung cancer patients for rs1042821 polymorphism (Dong *et al.*, 2009).

6.5 MSH2 polymorphism (*rs63749993, rs2303425, rs2303426, rs4987188, rs2303428, rs17217772*) and its role in lung cancer predisposition

Results from our study conclude that *2063T>G* polymorphism showed a reduced likelihood of developing lung cancer, which is not following earlier studies conducted by Li and colleagues where the population under study showed a 4-fold increased risk towards the development of esophageal cancer (EC) (Li *et al.*, 2012). Further, another study conducted by Arana and co-workers reported missense variant of *2063T>G* as pathogenic due to switch of polarity of the amino acid substitution in the functional protein domain (Medina-Arana *et al.*, 2006). We reported low or no risk of developing lung cancer for *IVS1+9G>C*, *118T>C*, and *1032G>A* polymorphisms. Our data contradict the earlier studies that reported a significant association of *IVS1+9G>C* with an increased risk of gall bladder cancer in the north Indian population (Srivastava *et al.*, 2010). Another study by Smolarz and colleagues reported a strong association of *1032G>A* polymorphism with the development of endometrial cancer in the Polish populace (Smolarz *et al.*, 2018). However, a previous study by Jung and co-workers on the Korean lung cancer patients reported no association of *118T>C* and *IVS1+9G>C* polymorphism with the development of lung cancer (Jung *et al.*, 2006). Moreover, for *T>C/-6* and *Asn*^{127}*Ser* polymorphism, we reported a multi-fold increase in the risk of developing lung cancer which can be associated with down-regulation of *MSH2* expression via TNFα mediated by NF- kB. NF-kB has been reported to be present in epithelial cells under inflammatory conditions and associated with inflammation-related tumor development (Eso *et al.*, 2016). Beiner and co-workers further corroborated our results for *T>C/-6* polymorphism, and they reported a strong association of individuals carrying mutant type genotype (*CC*) with the development of endometrial cancer (Beiner *et al.*, 2006).

We further reported a low risk for *2063T>G* and *IVS1+9G>C* polymorphisms of developing adenocarcinoma (ADCC) in contrast to *T>C/-6* and *Asn*^{127}*Ser* polymorphisms which are strongly implicated in the development of ADCC. An earlier study by Jung and co-workers reported an increased risk of developing ADCC for mutant type genotype for *IVS1+9G>C* polymorphism which is not following our results (Jung *et al.*, 2006). However, Jung and co-workers reported a reduced risk for the development of ADCC in individuals carrying mutant genotypes and a slightly increased risk of developing ADCC in individuals harboring heterozygous type genotype

(Jung *et al.,* 2006). The mutant genotype for *T>C/-6* polymorphism in both codominant and recessive models reported a 12 (**p=0.0003**) and 9-fold (**p=0.0009**) increased risk of developing ADCC, which might be attributed to TNF-α mediated downregulation of MSH2 protein. We could not find any study analyzing the association of *2063T>G*, *T>C/-6*, *1032G>A*, and *Asn*^{127}*Ser* polymorphisms with the risk of developing ADCC. When we further analyzed our data based on SQCC, all models of *2063T>G* and *IVS1+9G>C* polymorphism reported a reduced risk of developing lung cancer. In contrast, mutant type genotype reported a 9-fold (**p=0.0008**) increased risk towards SQCC occurrence for *T>C/-6* polymorphism. Jung and co-workers contradict our results; they have reported a 2-fold increase in the risk of developing SQCC in individuals harboring mutant and a 1.05-fold increase in risk for individuals harboring heterozygous type genotype in the Korean population for *IVS 1+9G>C* polymorphism (Jung *et al.,* 2006).

We found an exciting result in the codominant and dominant model while analyzing patients suffering from SCLC for *118T>C* polymorphism. We reported a 2-fold (**p=0.002**) increase in susceptibility towards SCLC. These results did not follow Jung and colleagues' previous study; they reported a reduction in risk for developing SCLC (Jung *et al.,* 2006). Further *2063T>G* and *IVS1+9G>C* polymorphisms reported a reduced risk towards SCLC development in codominant (**p=0.01**) and dominant model (**p=0.007**). These results were also not in line with the previous study of Jung and co-workers; they reported 1.7-fold and 1.4-fold increased risk towards the development of SCLC in the Korean population (Jung *et al.,* 2006). All the histology have reported a different rate of association with cancer development. Even though the rationale behind the histology-dependent difference in risk associated with cancer development by MSH2 polymorphic variants is not entirely understood, some studies have concluded that the carcinogenesis among various histology is the dominant factor for this discrepancy (Xinarianos *et al.,* 2000).

In this study, we further investigated the smoking habits of recruited subjects and whether there is an association between smoking and modulating the risk of lung cancer due to *MSH2* polymorphic variants. Our results concluded a reduced risk of developing lung cancer for all genetic models of *2063T>G* and *IVS1+9G>C* polymorphism. In contrast, smoking-induced lung cancer has no association with *118T>C* and *1032G>A* polymorphisms. Our results are not in accord with a previous study conducted by Li and co-workers; they reported an increased risk of endometrial cancer due to smoking in the Chinese population (Li *et al.,* 2012). Contrary to the above results, we observed that *T>C/-6* polymorphism had an 11- and 8-fold increased risk of smoking-induced lung cancer in both codominant and recessive models, respectively. When we further analyzed heavy smokers, the results followed previous results, *2063T>G* and *118T>C* polymorphisms reported a reduced risk in all genetic models, with no association with *1032G>A* and *IVS1+9G>C* polymorphism. In contrast, *T>C/-6* and $Asn^{127}Ser$ polymorphism reported a 2-fold (**p=0.002**) and 6-fold (**p=0.008**) increased risk in the codominant model. The results for *118T>C* polymorphism are not in accord with the previous study conducted by Xiao and co-workers; they reported that the combined variant was associated with increased risk of gastric cancer in younger patients and diffuse tumors in the Chinese population (Xiao *et al.,* 2012). A similarly reduced risk of lung cancer was observed for light smokers for *2063T>G* and *IVS1+9G>C*. In contrast, *T>C/-6* polymorphism reported a 20 (**p=0.0001**) and 13-fold (**p=0.001**) increased risk in the codominant and recessive models, respectively. We have also analyzed non-smokers to decipher any association with the risk of developing lung cancer and found a rather interesting result for non-smokers, *2063T>G* polymorphism, which shows reduced risk in all other categories. For non-smokers, we reported a 5-fold (**p=0.0008**) and 7-fold (**p=0.0001**) increased risk of developing lung cancer in codominant and dominant models. *IVS1+9G>C* polymorphism reported a protective effect for mutant genotype in the recessive model. In contrast, we did not find any association with the rest of the polymorphism with non-smokers and the risk of developing lung cancer.

We have also analyzed our data after dividing the study based on gender. Our results show that males have an increased risk of developing lung cancer for *T>C/-6* polymorphism (18-fold) and a 2-fold (**p<=0.0001**) increased risk in the codominant model for $Asn^{127}Ser$ polymorphism.

Haplotype analysis showed a noteworthy difference in the distribution between controls and cases. The Hap block 1 with a mutant allele of *118T>C*, *2063T>G*, and *IVS 1+9G>C* polymorphism confers a protective effect towards the occurrence of lung cancer, as compared to Hap block 3 with a mutant allele of *118T>C*, Hap block 5 with mutant allele of *2063T>G* polymorphism and Hap block 8 with a wild allele of all six polymorphism confers 2-fold, 1.9-fold, and 1.3-fold increased risk towards the development of lung cancer. These haplotypes possess higher risk alleles, which are persistently over-represented in lung cancer subjects as compared to healthy controls, indicating a major role of shortlisted polymorphisms in risk determination. We have calculated linkage disequilibrium for six SNPs of the *MSH2* gene, and our results indicate that *2063T>G* was in incomplete linkage disequilibrium with *Asn*^{127}*Ser*. Similarly, *1032G>A* was in complete linkage disequilibrium with *IVS1+9G>C* polymorphism. It has been demonstrated that interaction of polymorphisms of the same gene or other genes, by linkage disequilibrium, may be necessary for modulating the overall repair activity; thus, this might explain the influence of genetic variations in the carcinogenic process (Kiyohara *et al.*, 2010).

Also, to date, this is the first study reporting a high order regression analysis to create a decision tree and calculate the corresponding odds ratio of various genotypes using the CART technique. We also report for the first time SNP-SNP interaction among the *MSH2* genetic polymorphic variants and their effect in modulating risk towards developing lung cancer. Our results show that *MSH2 IVS 1+9G>C* (M) and *118T>C* are the SNPs playing a vital role in the interactive decision tree and dendrogram obtained from CART. *IVS1+9G>C* mutant polymorphism presented a 4-fold risk of developing lung cancer. For ADCC, *MSH2 Asn127Ser (M)/ T>C/-6 (M)* node represented a 43-fold enhanced risk of developing lung cancer as compared to *MSH2 IVS 1+9G>C (M)/ 118T>C (M)/ T>C/-6 (M)* for SCLC which represented 27-fold. In the case of SQCC, the *MSH2 Asn*^{127}*Ser* (M) represented a 5-fold increased risk of developing lung cancer. Hence, we can conclude that the independent presence of *MSH2* polymorphism can modulate the likelihood of developing lung cancer. However, SNPs interaction may also pose a severe risk of developing lung cancer.

6.5.1 Relation of *MSH2* (*rs63749993, rs2303425, rs2303426, rs4987188, rs2303428, rs17217772*) polymorphism with clinic-pathological features and prognosis

The *MSH2* gene is a tumor suppressor and, more specifically, a caretaker gene responsible for coding DNA mismatch repair protein, MSH2. Several mutations of the *MSH2* gene that predispose an individual to lung, colorectal, endometrial, and lynch syndrome-associated cancer have been well documented. Due to these mutations, the *MSH2* gene may stop expressing normal MSH2 protein and thus, produce either a truncated or modified protein or no protein at all (Li *et al.,* 2020). Aberrant MSH2 protein leads to a significant increase in unrepaired errors during cell division. This accumulation of errors over a long period results in the lack of maintenance of genomic integrity, the decrease in the propensity of repair due to damages caused by carcinogens, and thus pushing the cell towards uncontrolled cell growth, leading the cell to cancer of different organs such as colon, endometrial and lung (Sarasin, 2003). To find out the potential value of *MSH2* SNPs in lung cancer prognosis, we analyze the relationship between gene polymorphisms of *MSH2* towards overall survival and clinical response in lung cancer patients undergoing platinum-based chemotherapy. To the best of our understanding, this is the first study in the north Indian lung cancer patients to understand how DNA mismatch repair gene polymorphisms affect the survival years and clinical outcomes in lung cancer patients.

Data from our study suggest that the *MSH2 rs63749993, rs2303425, rs4987188, rs2303428,* and *rs17217772* polymorphisms did not show any significant association towards a stage of the tumor, lymph nodal invasion, and metastasis. However, it was observed that the *rs2303426* polymorphism showed a significant association among lung cancer patients who were carrying the mutant genotype and tumor size. These patients had a 2-fold higher risk for advanced-stage tumors (T4). Furthermore, the *rs2303426* polymorphism also exhibited a 2.25-fold (**p=0.01**) increased risk of lymph nodal invasion in patients with the mutant form of the gene. This observation from our data follows a previous study conducted by Zhu and colleagues in Chinese patients suffering from hepatocellular carcinoma, where they too have reported a significant correlation of rs2303428 polymorphism with TNM staging (Zhu *et al.*, 2018).

Furthermore, our study did not reveal any association between the six polymorphisms and clinical response based on how patients respond to platinum-based doublet chemotherapy. When we analyzed our data for overall survival, we found no significant association between the *MSH2 rs63749993*, *rs2303426*, *rs4987188*, and *rs2303428* polymorphism and OS. However, our data shows that the *MSH2* polymorphisms *rs2303425* and *rs17217772* were significantly associated with the overall survival of north Indian lung cancer patients undergoing platinum-based doublet chemotherapy. Our analysis observed that *MSH2* rs2303425 minor allele *C* was significantly associated with poor survival (**p=0.03**) in lung cancer patients. When we stratified our findings based on histological subtypes, adenocarcinoma patients who were heterozygous for the *MSH2 rs2303425* polymorphism were associated with more prolonged survival (MST=12.43, **p=0.03**). The *MSH2 rs2303425* polymorphism is located in the NF-Y transcription factor binding site, also known as CAAT binding factor. Studies have reported an increased risk of breast cancer with this polymorphic site. Further, it has also been reported that the C allele for this polymorphism has lower promoter activity in mammary cancer cell lines than the T allele. This suggests that the C allele has a lower transcriptional activity for the *MSH2* gene, which might diminish the MMR pathway's repair capacity, thereby conferring a role in risk towards cancer and an indirect effect towards platinum-based chemotherapy (Hsieh *et al.,* 2017).

Similarly, the *MSH2 rs17217772* SNP patients with AG genotype were shown to have a poor outcome than those with AA genotype (MST=4.4, **p=0.01**). Some earlier studies conducted in *MSH2* polymorphic variants have shown mixed results; Nogueira and colleagues have concluded that *rs2303426* polymorphism does not affect the overall survival of head and neck squamous cell carcinoma patient's Brazilian origin (Nogueira *et al.,* 2015). Zhu and co-workers have also shown a decrease in survival time for hepatocellular carcinoma patients in the Chinese (Zhu *et al.,* 2018). Another study conducted by Kantelinen and co-workers concluded that when $Asn^{127}Ser/Ala^{328}Pro$ polymorphism was present in combination, then DNA MMR repair efficiency got significantly reduced (Kantelinen *et al.,* 2012).

An earlier study evaluated in Chinese gastric cancer patients concluded that *rs2303428* polymorphism patients harboring combined genotype (*TC+CC*) have poor survival (Wang *et al.,* 2019). Contrary to the above study, we did not find any association of rs2303428 polymorphism with overall survival in north Indian patients. We have further performed survival analysis on lung cancer patients undergoing platinum-based doublet chemotherapy according to the chemotherapeutic regimen being administered to them. All the patients were given docetaxel, paclitaxel, irinotecan, or pemetrexed as a second line of chemotherapy drugs along with platinum-based drugs like carboplatin/ cisplatin. Our results concluded no significant association between all six *MSH2* polymorphisms being evaluated in this study along with docetaxel and pemetrexed chemotherapeutic regimen. In lung cancer patients administered paclitaxel, the subjects who were heterozygous carriers for the *rs17217772* polymorphism reported a significant decrease in overall survival for the patient and a higher probability of death (MST=7.96 vs. 2.63 months; **p=0.02**). Rest four polymorphisms showed no association between survival and paclitaxel chemotherapeutic regimen. When we evaluated the role of *MSH2 rs63749993* polymorphisms concerning patients administered irinotecan and platinum doublet chemotherapy, the subjects were harboring a single copy of the variant allele had a poor survival outcome and a decrease in survival time (6.06 months; **p=0.0004**). In *MSH2 rs63749993* polymorphism, T is substituting with G at codon 688 in exon 13, which resides in the Walker A motif. This motif plays an essential role in repair activity by interacting with both α and β phosphatases of ATP. It has been shown that mutation in the structural motif decreases in ATP binding and hydrolysis of MSH2-MSH6 complexes and thus results in loss or reduced MMR activity (Wu *et al.*, 1994, Drotschmann *et al.*, 1999). Instead of substituting T with G nucleotide, it confers a replacement of methionine amino acid with a positively charged arginine amino acid ($Met^{688}Arg$). This substitution is associated with an inadequate capacity to repair carcinogen-induced damage DNA.

Similarly, for the *rs2303426* polymorphism, subjects who were treated with irinotecan-based platinum therapy, our data suggest that lung cancer subjects carrying the mutant alleles (CC) for the *rs2303426* reported a significant increase in median survival time and better outcome as such patients had a low death ratio (HR=0.43 **p=0.02**). It has been postulated that the G allele of *MSH2* c.21119C>G SNP is less efficient in protein production than the wild-type allele. Thus, MSH2 expression levels of protein might alter the effects of platinum chemotherapy in lung cancer patients. However, to the best of our knowledge, no previous studies were conducted focusing on the associated SNP with clinical lung cancer aspects. A noteworthy highlight of this current study is the utilization of high-order gene interaction tools to identify the subgroups showing the best prognosis and highest hazard ratio due to various genotypic combinations, which traditional statistical tools were not capable of performing. This is the first study reporting the high order gene interaction of different *MSH2* polymorphic variants in predicting lung cancer subjects' overall survival as shown by the survival tree, wherein MSH2 $Asn^{127}Ser$ served as the most critical factor in modulating the survival time of lung cancer subjects.

As far as our knowledge is concerned, there is no such previous study reported to date, and the sample size of our study is extensive compared to the previous study, which makes it a significant strength of our investigation. The dissimilarity between the findings of the previous study and those of the current study may be assigned to differences in ethnicity, tumor characteristics, treatment, and median follow-up times. We were also unable to find many studies evaluating the effect of polymorphism under investigation in this study on overall survival (OS) and OS based on a chemotherapeutic regimen, making it another necessary strength of our study. One of the significant strengths of our study is the homogeneity of our patient population concerning treatment modality and tumor stage, which quashed out the effects that might persist in differences in DNA damage and their repair mechanisms and their effects on clinical outcomes. However, there are certain limitations to our study as well. First, although the sample size is large, the patients falling under different categories of lung cancer are more petite, resulting in biased results. Also, patients' smoking habits vary significantly, which might be another limitation.

6.5.2 MSH2 (*rs63749993, rs2303425, rs2303426, rs4987188, rs2303428, rs17217772*) polymorphism and its relationship with toxicity.

This investigation evaluated whether the polymorphisms in DNA mismatch repair - *MSH2* gene were associated with increased toxicity in lung cancer patients undergoing platinum-based doublet chemotherapy. We found that variant heterozygote of *MSH2 118T>C, 2063T>G* polymorphisms was significantly associated with grade 2-4 gastrointestinal toxicity, specifically diarrhea. Contrarily variant homozygote of *MSH2 IVS1+9G>C* polymorphism was noticeably related with reduced risk of developing severe grade of gastrointestinal toxicity (Diarrhea). An earlier investigation reported that MSH2 plays a crucial role in influencing 6-thioguanine (6-TG)- and O_6-methylguanine- induced toxicity in cells lacking DNA glycosylase, hinting that MSH2 serves a critical part in attenuating oxidative DNA damage (Cooley *et al.,* 2010). In our investigation, individuals carrying both copies of the variant allele for *rs2303426* polymorphism, which is present in the intronic region of *MSH2*, results in the translation of G to C nucleotide change, such patients reflected a poor hematological and anorexia toxicity tolerance after being treated with platinum-based doublet chemotherapy. Following the same trend, *rs2303428* polymorphism present at location -6 of 3'splice acceptor site of exon 13, leading to T/C single-nucleotide variation, also reported a poor hematological toxicity tolerance in patients undergoing chemotherapy. We hypothesized that *rs2303426* and *rs2303428* polymorphisms might alter the capacity of *MSH2* to remove platinum adducts. Previous studies have shown that increased expression of MSH3 sequesters nuclear MSH2 and support production of MutSβ heterodimer, causing a significant alteration in proportions of MutSα and MutSβ, reducing the efficiency of the mismatch repair system (Drummond *et al.,* 1997). Our results further reported no association between *MSH2* (*rs2303426*) genetic polymorphic variant and nephrotoxicity, which are following the previous study conducted by Noguiera and co-workers on head and neck squamous cell carcinoma patients of Brazilian origin (Nogueira *et al.,* 2018).

6.6 Computational analysis

Further, analyzing genetic information, understanding gene expression, evolutionary analysis, SNP analysis, molecular dynamics studies, and model generation of the desired protein to better understand the functionality of a protein can be achieved by modern-day computational tools and techniques. Kumar and co-workers utilize computational tools and techniques to unearth the effect of missense mutation responsible for causing D-2- hydroxyglutaric aciduria. Using these tools, they understood structural changes caused by these mutations that further helped them to develop novel targets for new drug therapies (Kumar *et al.,* 2018).

Mutations in the *MLH1*, *MSH2*, *MSH3* and *MSH6* gene were reported to be linked with different cancers such as lung cancer, prostate, colorectal, colon cancer, hereditary nonpolyposis colorectal cancer (HNPCC) syndrome, mismatch repair cancer syndrome (MMRCS), Muir-Torre syndrome (MRTES), and endometrial cancer (ENDMC) , breast cancer (Khooshemehri *et al.*, 2021, Whiffin *et al.,* 2011, Pardini *et al.,* 2020, Marafie *et al.,* 2009, Chen *et al.,* 2020, Yamamoto *et al.,* 2021, Smith *et al.,* 2008, Conde *et al.*, 2009). The current study has tried to screen the SNPs which need to be validated in different populations for their linkage with genetic susceptibility towards cancer development. A consensus approach that integrates multiple prediction tools is being used in identifying the functional SNPs. Four different tools for predicting deleterious SNPs (SIFT, PROVEAN, PANTHER, and PolyPhen), two tools for identifying disease association (PhD SNP and SNP&GO), and two tools for estimating stability (I-mutant and MUPro) were applied. The benefit of different tools is that it allows consideration of multiple algorithms, parameters, and databases.

For MSH2 protein, the identified SNPs by these tools were distributed in all three domains (DNA binding, MSH3/MSH6 interaction, and MutL homologs interaction domain). Since all three domains play a crucial role in the proper functioning of the DNA repair pathways, any mutation in these domains is bound to affect the efficiency of these pathways. Five nsSNPs corresponding to seven amino acid changes (*rs17217723, rs63750624, rs63750070, rs63750382*, and *rs63751444*) lies in the DNA binding domain. Two nsSNPs (*rs63750624* and *rs63751444*) out of these seven amino acid changes have been reported to be associated with the development of HNPCC due to complete loss of MMR function, whereas one SNP (*rs63750070*) have been linked with the development of Lynch syndrome (Ollila *et al.*, 2006, Fan *et al.*, 2006).

hMSH3/MSH6 domain of MSH2 has two nsSNP (*rs63750828* and *rs63751207*) out of 11, which have reported complete loss of MMR function, and hence they are implicated in the development of HNPCC. Interestingly MutL homologs binding domain has 15 nsSNPs out of which 6 (*rs63750398, rs63751477*, *rs63750234, rs63750232*, *rs63750961,* and *rs267607996*) are reported to be linked with HNPCC (Table 4) (Heinen *et al.,* 2002, Isidro *et al.,* 2000, Lützen *et al.,* 2008, Ollila *et al.,* 2008). A significant increment in the aggregation of proteins, misfolding, and degradation has also been reported due to decreased protein stability (Du *et al.*, 2005, Singh *et al.*, 2010).

For MSH6 protein, the identified SNPs by all computational tools were distributed in five domains (domain I-V). In domain I, two SNPs L435P (*rs63751405*) and L449P (*rs63750741*) were located. *rs63750741* is a missense variant and is implicated in HNPCC (https://www.disgenet.org/browser/1/1/2/2956/diseaseid__C1333990-source__ALL/). Only one SNP was located in domain II, G566R (*rs63749973*). This polymorphic variant was reported to change highly conserved amino acids leading to reduced mismatch repair capacity, and loss of protein expression in familial non-HNPCC (Kolodner *et al.,* 1999). Two SNPs were located in domain III, C765W (*rs63750985*) and R772W (*rs63750138*) which were reported to implicated in the development of HNPCC (Plaschke *et al.*, 2004 and https://cutt.ly/SPgNXd7). Only one SNP Y969C (*rs63749919*) was present in IVth domain. Further domain V has six SNPs (*rs63750617***,** *rs63750253, rs376243329, rs63750442*, *rs63751328***,** and *rs182024561*). Out of these six SNPs, one (*rs63750617*) is implicated in endometrium and rectum cancer (Plaschke *et al.,* 2004) whereas for *rs63751328*, E1193K, the amino acid E1193 is located in a highly conserved region between the ATP and Mg^{2+} binding sites in *MSH6* and accordingly can be expected to be sensitive to alterations (Iaccarino et al, 1998, 2000; Dufner et al, 2000) Further in a study conducted by Kariola and co-workers it is reported to be E1193K mutation is present in endometrial or colorectal tumor patients that have high MSI of central Ohio origin (Kariola *et al.*, 2004).

For MLH1 protein, since the most important interaction domain was MLH3 interaction domain hence all the SNPs located in it were considered for analysis. Out of 27 SNPs (28 amino acid changes), *rs35045067* (Y646C), *rs63749909* (V506A), *rs63750289* (L549P), *rs63750575* (L588P), *rs63750610* (P648L), *rs63750825* (L636P), *rs63750899* (P648S), *rs63750511* (Q542P), *rs63750693* (L622H, L622P) are reported in Hereditary Non-polyposis colorectal cancer (HNPCC) (Belvederesi *et al.,* 2006, Lliu *et al.,* 1996, Hardt *et al.,* 2011, Raevaara *et al.,* 2004). *rs63749792* (P640S) is an MLH1 missense variant resulting from Proline to Serine substitution at codon 646 and is reported in hereditary endometrial cancer. *rs63750242* (L676P) is present in exon 18 of *MLH1* gene, results from T to C amino acid change at nucleotide position 2027. Leucine present at 676 position is replaced by an amino acid which have similar properties (Proline). Functional studies based on L676P have reported impaired expression of MLH1 in HNPCC/Lynch syndrome families (Hardt *et al.,* 2011). *rs63750587* (A586D) this sequence mutation removes alanine with aspartic acid at codon position 586 of MLH1 protein.

Mutations in the interacting domain can affect protein-protein information, a docking study can help understand protein-protein interaction and identify essential mutant affecting the interaction (Gromiha *et al.,* 2017). Seven, five and four mutants of MSH2, MSH6 and MLH1 protein were found to have lower HADDOCK score and fewer hydrogen bonds compared to wild type whereas two, three and twelve mutants of MSH2, MSH6 and MLH1 protein were found to have higher HADDOCK score and more hydrogen bonds revealing that these SNPs affect the interaction of MSH2, MSH6 and MLH1 with MSH6, MSH2 and MLH3 respectively. Docking results indicate these SNPs are affecting the interaction of MSH2/MSH6, MSH6/MSH2 and MLH1/MLH3 but for deep understanding in functional and structural change due to these SNPs need to be done using molecular dynamics in the future. Although some of these *in silico* screened SNPs are reported to be associated with cancer predisposition, further to confirm its precise role, there is a need to validate the SNPs in in-vitro experimental system.

6.7 Strength and weakness of the study

The strength of the current study are a large sample size compared to previous study done in the Indian population, which could enhance the reliability of the results.Furthermore, a particular strength of this study lies in the analysis of three endpoints, *viz*. OS, chemotherapy response, and clinic-pathological features to the polymorphic variants of gene involved in mismatch repair pathways. As far as our knowledge is concerned, no other study has evaluated the role of these polymorphisms to OS and platinum-based doublet chemotherapy and the combinatorial analysis of these SNPs. Our study have several strengths, including that many patients receiving platinum-based chemotherapy were enrolled, and all the patients were treated in the same hospital.

Secondly, the recruitment of patients and the clinical parameters data collection was carried out independently without knowing the polymorphism's studied. The stratified analysis was also performed for smoking and histology to find out high risk posing subgroups. Our study is the first attempt in DNA mismatch repair pathway genes for lung cancer in the Indian population to use a multifactorial approach (CART and MDR) to explore the complex interaction between these genes and their relationship towards lung cancer risk. This is the first study that involves evaluating the association of these SNPs with chemotherapy related toxicities. The present findings with a large sample size for the overall and stratified groups would help develop biomarkers both for diagnosis of lung cancer and prognosis of lung cancer.

6.8 Limitation of the study

The limitation of this study includes: more population-based study in different population of India needs to be carried out to reduce the extent of selection bias. Our research has so far been limited to North Indian population. As genetic polymorphism frequently differs between ethnic groups, this should be expanded to diverse ethnic population. Secondly, to fulfill the personalized therapeutic system, developing comprehensive and reliable SNP and haplotype analyses of diverse ethnic populations, like Indian, shall be highly beneficial, allowing for predicted genotype -drug efficacy in each individual. But further study needs to understand the role of these genes in immunotherapy and development of oncopanles for detection.

Chapter 7

Conclusion

This preliminary study provides an insight into the potential of genetic differences in MMR genes to modulate lung cancer susceptibility. *MLH1* (*rs1800734*), *MSH6* (*rs3136228*), *MSH2* (*rs2303425, rs4987188, rs2303428* and *rs17217772*) and *MSH3* (*rs26279*) polymorphism were associated with modulating the risk of lung cancer in North Indian population. *MSH6* (*rs1800932*) independently affects patients' overall survival, whereas *MSH6* (*rs1042821*) polymorphism affects the OS of patients undergoing docetaxel and carbo/cisplatin combination chemotherapeutic regimen. Computational analysis of nsSNPs selected for this study showed that they play a key role in altering the structure of proteins, thereby altering the stability of proteins associated with the development of various cancers. Further docking analysis of deleterious nsSNPs of *MLH1*, *MSH2*, *MSH3* and *MSH6* gene revealed that as docking score increases, number of hydrogen bonds increases implying an increase in overall stability of the protein.

However, since our study only included 500 lung cancer patients and 500 control subject which despite being huge sample size is still not enough because the subjects were subdivided into smaller groups based on gender, smoking status, lung cancer subtypes. Hence such studies needs to be conducted on a large scale covering entire India to deduce any logical conclusion.

Chapter 8

Bibliography

- Abdelraheem, N. E., Osman, M. M., Elgemaabi, O. M., Alla, A. A. F., Ismail, M. M., Osman, S. A., Ibrahim, A. I., Babekir, N. E., Mekki, S. O., & 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.
- Aihara, T., Kim, Y., & Takatsuka, Y. (2002). Phase II study of weekly docetaxel in patients with metastatic breast cancer. *Annals of oncology*, *13*(2), 286-292
- Aubrey, B. J., Kelly, G. L., Janic, A., Herold, M. J., & Strasser, A. (2018). How does p53 induce apoptosis and how does this relate to p53-mediated tumour suppression?. *Cell Death & Differentiation*, *25*(1), 104-113
- Augusto-Pinto, L., Da Silva, C. G., Lopes, D. D. O., Machado-Silva, A., & Machado, C. R. (2003). Escherichia coli as a model system to study DNA repair genes of eukaryotic organisms. *Genetics and Molecular Research*, *2*(1), 77-91
- Ban, C., & Yang, W. (1998). Crystal structure and ATPase activity of MutL: implications for DNA repair and mutagenesis. *Cell*, *95*(4), 541-552.
- Ban, C., & Yang, W. (1998). Structural basis for MutH activation in E. coli mismatch repair and relationship of MutH to restriction endonucleases. *The EMBO journal*, *17*(5), 1526-1534.
- Banerjee, S. (2007). MSH6 (mutS homolog 6 (E. Coli)). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*. 11(3):169-172
- Barnum, K. J., & O'Connell, M. J. (2014). Cell cycle regulation by checkpoints. In *Cell cycle control* (pp. 29-40). Humana Press, New York, NY.
- Behera, D. (2007). New approach to the treatment of lung cancer: the molecular targeted therapy. *Indian journal of chest diseases & allied sciences*, *49*(3), 149-158
- Behera, D. and Balamugesh, T. (2004), 'Lung cancer in india', *Indian Journal of Chest Diseases and Allied Sciences* 46, 269–282
- Beiner, M. E., Rosen, B., Fyles, A., Harley, I., Pal, T., Siminovitch, K., & Narod, S. A. Endometrial cancer risk is associated with variants of the mismatch repair genes MLH1 and MSH2. *Cancer Epidemiol. Biomark. Prev. 15*(9), 2006, 1636-1640.

- Beiner, M. E., Rosen, B., Fyles, A., Harley, I., Pal, T., Siminovitch, K., & Narod, S. A. Endometrial cancer risk is associated with variants of the mismatch repair genes MLH1 and MSH2. *Cancer Epidemiol. Biomark. Prev. 15*(9), 2006, 1636-1640.

- Belcher, S. M., Ma, X., & Le, H. H. (2009). Blockade of estrogen receptor signaling inhibits growth and migration of medulloblastoma. *Endocrinology*, *150*(3), 1112-1121.

- Belvederesi, L., Bianchi, F., Loretelli, C., Gagliardini, D., Galizia, E., Bracci, R., ... & Porfiri, E. (2006). Assessing the pathogenicity of MLH1 missense mutations in patients with suspected hereditary nonpolyposis colorectal cancer: correlation with clinical, genetic and functional features. *European journal of human genetics*, *14*(7), 853-859.
- Benjamini, Y. and Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the royal statistical society. Series B (Methodological)* pp. 289–300.
- Berndt SI, Platz EA, Fallin MD, *et al.*, Mismatch repair polymorphisms and the risk of colorectal cancer. *International Journal of Cancer* 2007 Apr 1;120(7):1548-54.
- Berndt SI, Platz EA, Fallin MD, *et al.*, Mismatch repair polymorphisms and the risk of colorectal cancer. *International Journal of Cancer* 2007 Apr 1;120(7):1548-54.
- Blackwell, L. J., Martik, D., Bjornson, K. P., Bjornson, E. S., & Modrich, P. (1998). Nucleotide-promoted release of hMutSα from heteroduplex DNA is consistent with an ATP-dependent translocation mechanism. *Journal of Biological Chemistry*, *273*(48), 32055-32062.
- Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L., & Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human mutation*, *30*(8), 1237-1244.
- Capriotti, E., & Fariselli, P. (2017). PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic acids research*, *45*(W1), W247-W252.
- Capriotti, E., Calabrese, R., & Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, *22*(22), 2729-2734.
- Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2. 0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic acids research*, *33*(suppl_2), W306-W310.
- Carrera-Lasfuentes, P., Lanas, A., Bujanda, L., Strunk, M., Quintero, E., Santolaria, S., ... & García-González, M. A. (2017). Relevance of DNA repair gene polymorphisms to gastric cancer risk and phenotype. *Oncotarget*, *8*(22), 35848.
- Cederquist, K., Emanuelsson, M., Göransson, I., Holinski-Feder, E., Müller-Koch, Y., Golovleva, I., & Grönberg, H. (2004). Mutation analysis of the MLH1, MSH2 and MSH6 genes in patients with double primary cancers of the colorectum and the endometrium: A population-based study in northern Sweden. *International journal of cancer*, *109*(3), 370-376.
- 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.
- Chang, A. (2011), 'Chemotherapy, chemoresistance and the changing treatment landscape or nsclc', *Lung cancer* 71(1), 3–10

- Chaudhry, R., & Bordoni, B. (2017). Anatomy, thorax, lungs.
- Chaudhry, R., & Bordoni, B. (2020). Anatomy, thorax, lungs
- Chen, Q., Wang, M., Xu, Z., Wang, M., Jin, S., Tian, S., & Xiao, S. (2020). Muir-Torre Syndrome with a Frame-shift Mutation in the MSH2 Gene: A Rare Case Report and Literature Review. *International Journal of Gynecological Pathology*, *39*(2), 136-140.
- Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, *62*(4), 1125-1132
- Chernapalli, V., Komanduri, M., & Sabbella, S. (2017). In silico analysis of single nucleotide polymorphisms in human proliferating cell nuclear antigen gene. *RJLBPCS*, *2*(5), 187.
- Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE 7(10): e46688. https://doi.org/10.1371/journal.pone.0046688
- Choi, Y., & Chan, AP (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* , *31* (16), 2745-2747.
- Conde, J., Silva, S. N., Azevedo, A. P., Teixeira, V., Pina, J. E., Rueff, J., & Gaspar, J. F. (2009). Association of common variants in mismatch repair genes and breast cancer susceptibility: a multigene study. *BMC cancer*, *9*(1), 1-9.
- Cooley, N., Elder, R. H., & Povey, A. C. (2010). The effect of Msh2 knockdown on methylating agent induced toxicity in DNA glycosylase deficient cells. *Toxicology*, *268*(1-2), 111-117.
- Curtin, K., Samowitz, W. S., Wolff, R. K., Caan, B. J., Ulrich, C. M., *et al.*, MSH6 G39E polymorphism and CpG island methylator phenotype in colon cancer. *Mol. Carcinog.: Published in cooperation with the University of Texas MD Anderson Cancer Center*, 2009; *48*(11), 989-994.
- Dao, V., & Modrich, P. (1998). Mismatch-, MutS-, MutL-, and helicase II-dependent unwinding from the single-strand break of an incised heteroduplex. *Journal of Biological Chemistry*, *273*(15), 9202-9207
- de Saro, F. J. L., Marinus, M. G., Modrich, P., & O'Donnell, M. (2006). The β sliding clamp binds to multiple sites within MutL and MutS. *Journal of Biological Chemistry*, *281*(20), 14340-14349.
- De Zio, D., Cianfanelli, V., & Cecconi, F. (2013). New insights into the link between DNA damage and apoptosis. *Antioxidants & redox signaling*, *19*(6), 559-571
- Detterbeck, F. C., Boffa, D. J., Kim, A. W., & Tanoue, L. T. (2017). The eighth edition lung cancer stage classification. *Chest*, *151*(1), 193-203.
- Doherty, J. A., Sakoda, L. C., Loomis, M. M., Barnett, M. J., Julianto, L., Thornquist, M. D., ... & Chen, C. (2013). DNA repair genotype and lung cancer risk in the beta-carotene and retinol efficacy trial. *International Journal of Molecular Epidemiology and Genetics*, *4*(1), 11.

- Domingo, E., & Schwartz Jr, S. (2005). MSH2 (human mutS homolog 2). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, 9(4):291-292.
- Domingo, E., & Schwartz Jr, S. (2006). MSH3 (mutS homolog 3 (E. coli)). *Atlas of Genetics and Cytogenetics in Oncology and Haematology*, 10(4):251-252
- Domingo, E; & Schwartz, Jr,S. (2005).MLH1 (human mutL homolog 1) *Atlas of Genetics and Cytogenetics in Oncology and Haematology* 9(2):120-122
- Dominguez-Brauer, C., Brauer, P. M., Chen, Y. J., Pimkina, J., & Raychaudhuri, P. (2010). Tumor suppression by ARF: gatekeeper and caretaker. *Cell Cycle*, *9*(1), 86-89.
- Dong, X., Jiao, L., Li, Y., Evans, D. B., Wang, H., Hess, K. R., ... & Li, D. (2009). Significant associations of mismatch repair gene polymorphisms with clinical outcome of pancreatic cancer. *Journal of clinical oncology*, *27*(10), 1592.
- Dong, X., Li, Y., Chang, P., Hess, K. R., Abbruzzese, J. L., & Li, D. (2012). DNA mismatch repair network gene polymorphism as a susceptibility factor for pancreatic cancer. *Molecular carcinogenesis*, *51*(6), 491-499.
- Donnadieu, N., Paesmans, M., & Sculier, J. P. (1991). Chimiothérapie des cancers bronchiques non à petites cellules: meta-analyse de la littérature en fonction de l'extension de la maladie. *Revue des maladies respiratoires*, *8*(2), 197-204.
- Drotschmann, K., Clark, A. B., Tran, H. T., Resnick, M. A., Gordenin, D. A., & Kunkel, T. A. (1999). Mutator phenotypes of yeast strains heterozygous for mutations in the MSH2 gene. *Proceedings of the National Academy of Sciences*, *96*(6), 2970-2975.
- Drummond, J. T., Genschel, J., Wolf, E., & Modrich, P. (1997). DHFR/MSH3 amplification in methotrexate-resistant cells alters the hMutSα/hMutSβ ratio and reduces the efficiency of base–base mismatch repair. *Proceedings of the National Academy of Sciences*, *94*(19), 10144-10149.
- Du, K., Sharma, M., & Lukacs, G. L. (2005). The ΔF508 cystic fibrosis mutation impairs domain-domain interactions and arrests post-translational folding of CFTR. *Nature structural & molecular biology*, *12*(1), 17-25.
- Dufner, P., Marra, G., Räschle, M., & Jiricny, J. (2000). Mismatch recognition and DNA-dependent stimulation of the ATPase activity of hMutSα is abolished by a single mutation in the hMSH6 subunit. *Journal of Biological Chemistry*, *275*(47), 36550-36555.
- Erie, D. A., & Weninger, K. R. (2014). Single molecule studies of DNA mismatch repair. *DNA repair*, *20*, 71-81
- Eso, Y., Takai, A., Matsumoto, T., Inuzuka, T., Horie, T., Ono, K., ... & Marusawa, H. (2016). MSH2 dysregulation is triggered by proinflammatory cytokine stimulation and is associated with liver cancer development. *Cancer research*, *76*(15), 4383-4393.
- Fan, Y., Liu, X., Zhang, H., Dai, J., Zhang, X., Zhu, M., Wang, Y, et al. (2006). Variations in exon 7 of the MSH2 gene and susceptibility to gastrointestinal cancer in a Chinese population. *Cancer genetics and cytogenetics*, *170*(2), 121-128.
- Fishel, R. (1998). Mismatch repair, molecular switches, and signal transduction. *Genes & development*, *12*(14), 2096-2101.

- Fredriksson, H., Ikonen, T., Autio, V., Matikainen, M. P., Helin, H. J., Tammela, T. L. J., Koivisto, P. A., & Schleutker, J. (2006). Identification of germline MLH1 alterations in familial prostate cancer. *European Journal of Cancer*, 42(16), 2802-2806.
- Frousios, K., Iliopoulos, C. S., Schlitt, T., & Simpson, M. A. (2013). Predicting the functional consequences of non-synonymous DNA sequence variants—evaluation of bioinformatics tools and development of a consensus strategy. *Genomics*, *102*(4), 223-228
- Fukui, K. (2010). DNA mismatch repair in eukaryotes and bacteria. *Journal of nucleic acids*, *2010*
- Gajalakshmi, V., Shanta, V., & Swaminathan, R. (2001). Cancer registration in India. *Asian Pacific Journal of Cancer Prevention*, *2*, 13-20
- Gazdar, A. F. (2007). DNA repair and survival in lung cancer—the two faces of Janus. *New England Journal of Medicine*, *356*(8), 771-773.
- Gradia, S., Subramanian, D., Wilson, T., Acharya, S., Makhov, A., Griffith, J., & Fishel, R. (1999). hMSH2–hMSH6 forms a hydrolysis-independent sliding clamp on mismatched DNA. *Molecular cell*, 3(2), 255-261.
- Gromiha, M. M., An, J., Kono, H., Oobatake, M., Uedaira, H., & Sarai, A. (1999). ProTherm: thermodynamic database for proteins and mutants. *Nucleic acids research*, *27*(1), 286-288
- Gromiha, M. M., Yugandhar, K. A., & Jemimah, S. (2017). Protein–protein interactions: scoring schemes and binding affinity. *Current opinion in structural biology*, *44*, 31-38.
- Guarné, A., Ramon-Maiques, S., Wolff, E. M., Ghirlando, R., Hu, X., Miller, J. H., & Yang, W. (2004). Structure of the MutL C-terminal domain: a model of intact MutL and its roles in mismatch repair. *The EMBO journal*, *23*(21), 4134-4145.
- Gupta, U., Mir, S. S., Garg, N., Agarwal, S. K., Pande, S. and Mittal, B. (2016), 'Association study of inflammatory genes with rheumatic heart disease in north Indian population: A multi-analytical approach', *Immunology letters* 174, 53–62
- Hardt, K., Heick, S. B., Betz, B., Goecke, T., Yazdanparast, H., Küppers, R., ... & Royer-Pokora, B. (2011). Missense variants in hMLH1 identified in patients from the German HNPCC consortium and functional studies. *Familial cancer*, *10*(2), 273-284.
- Hecht, S. S. (1999). Tobacco smoke carcinogens and lung cancer. *JNCI: Journal of the National Cancer Institute*, *91*(14), 1194-1210
- Heinen, C. D., Wilson, T., Mazurek, A., Berardini, M., Butz, C., & Fishel, R. (2002). HNPCC mutations in hMSH2 result in reduced hMSH2-hMSH6 molecular switch functions. *Cancer cell*, *1*(5), 469-478.
- Houtgraaf, J. H., Versmissen, J., & van der Giessen, W. J. (2006). A concise review of DNA damage checkpoints and repair in mammalian cells. *Cardiovascular Revascularization Medicine*, *7*(3), 165-172.
- Hsieh, Y. C., Cho, E. C., Tu, S. H., Wu, C. H., Hung, C. S., Hsieh, M. C., ... & Chiou, H. Y. (2017). Msh2 rs2303425 polymorphism is associated with early-onset breast cancer in taiwan. *Annals of surgical oncology*, *24*(2), 603-610.

- http://202.120.31.177/myAnalysis.php
- http://sdmproject.com/ utilities/?show=FDR
- http://www.nfdht.nl
- https://cutt.ly/SPgNXd7
- https://gco.iarc.fr/today/online-analysis
- https://www.disgenet.org/browser/1/1/2/2956/diseaseid__C1333990-source__ALL/
- https://www.genome.jp/pathway/ecd03430
- https://www.genome.jp/pathway/ecd03430
- https://www.genome.jp/pathway/ecd03430
- Iaccarino I, Marra G, Dufner P, Jiricny J (2000) Mutation in the magnesium binding site of hMSH6 disables the hMutS*α* sliding clamp from translocating along DNA. *J Biol Chem* **275**: 2080–2086
- Iaccarino, I., Marra, G., Palombo, F., & Jiricny, J. (1998). hMSH2 and hMSH6 play distinct roles in mismatch binding and contribute differently to the ATPase activity of hMutSα. *The EMBO journal*, *17*(9), 2677-2686.
- Irving, J. A., RT, B. D., RT, L. W., & Gilks, C. B. (2002). Cyclical expression of the DNA mismatch repair enzyme hMLH1 in normal endometrium. *Fertility and sterility*, *78*(1), 195-196.
- Isidro, G., Veiga, I., Matos, P., Almeida, S., Bizarro, S., Marshall, B., ... & Boavida, M. G. (2000). Four novel MSH2/MLH1 gene mutations in portuguese HNPCC families. *Human mutation*, *15*(1), 116-116
- Jung, C. Y., Choi, J. E., Park, J. M., Chae, M. H., Kang, H. G., Kim, K. M., ... & Park, J. Y. (2006). Polymorphisms in the hMSH2 gene and the risk of primary lung cancer. *Cancer Epidemiology and Prevention Biomarkers*, *15*(4), 762-768.
- Jung, C. Y., Choi, J. E., Park, J. M., Chae, M. H., Kang, H. G., Kim, K. M., ... & Park, J. Y. (2006). Polymorphisms in the hMSH2 gene and the risk of primary lung cancer. *Cancer Epidemiology and Prevention Biomarkers*, *15*(4), 762-768.
- Kadyrov, F. A., Dzantiev, L., Constantin, N., & Modrich, P. (2006). Endonucleolytic function of MutLα in human mismatch repair. *cell*, *126*(2), 297-308.
- Kadyrov, F. A., Dzantiev, L., Constantin, N., & Modrich, P. (2006). Endonucleolytic function of MutLα in human mismatch repair. *cell*, *126*(2), 297-308
- Kantelinen, J., Kansikas, M., Candelin, S., Hampel, H., Smith, B., Holm, L., ... & Nyström, M. (2012). Mismatch repair analysis of inherited MSH2 and/or MSH6 variation pairs found in cancer patients. *Human mutation*, *33*(8), 1294-1301.
- Kariola, R., Hampel, H., Frankel, W. L., Raevaara, T. E., de la Chapelle, A., & Nyström-Lahti, M. (2004). MSH6 missense mutations are often associated with no or low cancer susceptibility. *British journal of cancer*, *91*(7), 1287-1292.

- Khooshemehri, P., Jamaldini, S. H., Ziaee, S. A. M., Afshari, M., Sattari, M., Narouie, B., ... & Hasanzad, M. (2020). Genetic polymorphism of mismatch repair genes and susceptibility to prostate cancer. *Urology Journal*, *17*(3), 271-275.
- Kiyohara, C., Takayama, K., & Nakanishi, Y. (2010). Lung cancer risk and genetic polymorphisms in DNA repair pathways: a meta-analysis. *Journal of nucleic acids*, *2010*.
- Kolodner, R. D., Tytell, J. D., Schmeits, J. L., Kane, M. F., Gupta, R. D., Weger, J., ... & Li, F. P. (1999). Germ-line msh6 mutations in colorectal cancer families. *Cancer Research*, *59*(20), 5068-5074.
- Kumar, D. T., Emerald, L. J., Doss, C. G. P., Sneha, P., Siva, R., Jebaraj, W. C. E., & Zayed, H. (2018). Computational approach to unravel the impact of missense mutations of proteins (D2HGDH and IDH2) causing D-2-hydroxyglutaric aciduria 2. *Metabolic brain disease*, *33*(5), 1699-1710
- Lahue, R. S., Au, K. G., & Modrich, P. (1989). DNA mismatch correction in a defined system. *Science*, *245*(4914), 160-164
- Lawania, S., Singh, N., Behera, D., & Sharma, S. (2019). XPG polymorphisms and their association with lung cancer susceptibility, overall survival and response in North Indian patients treated with platinum-based doublet chemotherapy. *Future Oncology*, *15*(2), 151-165.
- Lee, E., Levine, E. A., Franco, V. I., Allen, G. O., Gong, F., Zhang, Y., & Hu, J. J. (2014). Combined genetic and nutritional risk models of triple negative breast cancer. *Nutrition and cancer*, *66*(6), 955-963
- Lee, J. B., Cho, W. K., Park, J., Jeon, Y., Kim, D., Lee, S. H., & Fishel, R. (2014). Single-molecule views of MutS on mismatched DNA. *DNA repair*, *20*, 82-93.
- Lee, J. Y., Chang, J., Joseph, N., Ghirlando, R., Rao, D. N., & Yang, W. (2005). MutH complexed with hemi-and unmethylated DNAs: coupling base recognition and DNA cleavage. *Molecular cell*, *20*(1), 155-166.
- Lengauer, C., Kinzler, K. W., & Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, *396*(6712), 643-649.
- Li, F., Liu, Q., Chen, Y. Y., Yu, Z. N., Zhang, Z. P., Zhou, Y. F., ... & Zhang, X. E. (2008). Escherichia coli mismatch repair protein MutL interacts with the clamp loader subunits of DNA polymerase III. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *637*(1-2), 101-110.
- Li, J., Li, Y., Ni, H., Yang, Z., Chen, J., Li, Y., ... & Deng, A. (2020). A novel splice-site mutation in MSH2 is associated with the development of lynch syndrome. *Frontiers in Oncology*, *10*, 983.
- Li, T., Suo, Q., He, D., Du, W., Yang, M., Fan, X., & Liu, J. (2012). Esophageal cancer risk is associated with polymorphisms of DNA repair genes MSH2 and WRN in Chinese population. *Journal Of Thoracic Oncology*, *7*(2), 448-452.

- Li, T., Suo, Q., He, D., Du, W., Yang, M., Fan, X., & Liu, J. (2012). Esophageal cancer risk is associated with polymorphisms of DNA repair genes MSH2 and WRN in Chinese population. *Journal Of Thoracic Oncology*, *7*(2), 448-452.
- Lim, W., Ridge, C. A., Nicholson, A. G., & Mirsadraee, S. (2018). The 8th lung cancer TNM classification and clinical staging system: review of the changes and clinical implications. *Quantitative imaging in medicine and surgery*, *8*(7), 709.
- Liu, B., Parsons, R., Papadopoulos, N., Nicolaides, N. C., Lynch, H. T., Watson, P., ... & Kinzler, K. W. (1996). Analysis of mismatch repair genes in hereditary non–polyposis colorectal cancer patients. *Nature medicine*, *2*(2), 169-174.
- Liu, J., Zheng, B., Li, Y., Yuan, Y., & Xing, C. (2019) Genetic polymorphisms of DNA repair pathways in sporadic colorectal carcinogenesis. *J. Cancer*, *10*(6), 1417.
- Loeb, K. R. and Loeb, L. A. (2000), 'Significance of multiple mutations in cancer', *Carcinogenesis* 21(3), 379–385
- *Lung Cancer* (2017). URL: *https://ghr.nlm.nih.gov/condition/lung-cancer*
- Lützen, A., de Wind, N., Georgijevic, D., Nielsen, F. C., & Rasmussen, L. J. (2008). Functional analysis of HNPCC-related missense mutations in MSH2. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *645*(1-2), 44-55.
- Mah, J. T., Low, E. S., & Lee, E. (2011). In silico SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery. *Drug discovery today*, *16*(17-18), 800-809
- Malik, P. S., & Raina, V. (2015). Lung cancer: prevalent trends & emerging concepts. *The Indian journal of medical research*, *141*(1), 5.
- Mankowski, N. L., & Bordoni, B. (2020). Anatomy, Head and Neck, Nasopharynx.
- Marafie, M. J., Al-Awadi, S., Al-Mosawi, F., Elshafey, A., Al-Ali, W., & Al-Mulla, F. (2009). Impact of 226C> T MSH2 gene mutation on cancer phenotypes in two HNPCC-associated highly-consanguineous families from Kuwait: emphasis on premarital genetic testing. *Familial cancer*, *8*(4), 289-298.
- Mathur, P., Sathishkumar, K., Chaturvedi, M., Das, P., Sudarshan, K. L., Santhappan, S., ... & ICMR-NCDIR-NCRP Investigator Group. (2020). Cancer statistics, 2020: report from national cancer registry programme, India. *JCO Global Oncology*, *6*, 1063-1075
- Mazur, D. J., Mendillo, M. L., & Kolodner, R. D. (2006). Inhibition of Msh6 ATPase activity by mispaired DNA induces a Msh2 (ATP)-Msh6 (ATP) state capable of hydrolysis-independent movement along DNA. *Molecular cell*, *22*(1), 39-49.
- Medina-Arana, V., Barrios, Y., Fernández-Peralta, A., Herrera, M., Chinea, N., Lorenzo, N., ... & González-Aguilera, J. J. (2006). New founding mutation in MSH2 associated with hereditary nonpolyposis colorectal cancer syndrome on the Island of Tenerife. *Cancer letters*, *244*(2), 268-273.
- Mendillo, M. L., Hargreaves, V. V., Jamison, J. W., Mo, A. O., Li, S., Putnam, C. D., Woods, V. L., & Kolodner, R. D. (2009). A conserved MutS homolog connector domain

interface interacts with MutL homologs. *Proceedings of the National Academy of Sciences*, *106*(52), 22223-22228.

- Metro, G., Chiari, R., Mare, M., Giannarelli, D., Tofanetti, F. R., Minotti, V., ... & Crinò, L. (2011). Carboplatin plus pemetrexed for platinum-pretreated, advanced non-small cell lung cancer: a retrospective study with pharmacogenetic evaluation. *Cancer chemotherapy and pharmacology*, *68*(6), 1405-1412.
- Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L. P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic acids research*, *49*(D1), D394-D403.
- Mik, M., Dziki, L., Malinowska, K., Trzcinski, R., Majsterek, I., & Dziki, A. (2017). Polymorphism of MSH2 Gly322Asp and MLH1–93G> A in non-familial colon cancer–a case-controlled study. *Archives of medical science: AMS*, *13*(6), 1295.
- Modrich, P., & Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. *Annual review of biochemistry*, *65*(1), 101-133.
- Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, *31*(13), 3812-3814
- Nogueira, G. A. S., Costa, E. F. D., Lopes-Aguiar, L., Lima, T. R. P., Visacri, M. B., Pincinato, E. C., ... & Lima, C. S. P. (2018). Polymorphisms in DNA mismatch repair pathway genes predict toxicity and response to cisplatin chemoradiation in head and neck squamous cell carcinoma patients. *Oncotarget*, *9*(51), 29538
- Nogueira, G. A. S., Costa, E. F. D., Lopes-Aguiar, L., Lima, T. R. P., Visacri, M. B., Pincinato, E. C., ... & Lima, C. S. P. (2018). Polymorphisms in DNA mismatch repair pathway genes predict toxicity and response to cisplatin chemoradiation in head and neck squamous cell carcinoma patients. *Oncotarget*, *9*(51), 29538.
- Nogueira, G. A. S., Lourenço, G. J., Oliveira, C. B. M., Marson, F. A. L., Lopes-Aguiar, L., Costa, E. F. D., & Lima, C. S. P. (2015). Association between genetic polymorphisms in DNA mismatch repair-related genes with risk and prognosis of head and neck squamous cell carcinoma. *International journal of cancer*, *137*(4), 810-818
- Obmolova, G., Ban, C., Hsieh, P., & Yang, W. (2000). Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature*, *407*(6805), 703-710.
- Ollila, S., Sarantaus, L., Kariola, R., Chan, P., Hampel, H., Holinski–Feder, E., ... & Nyström, M. (2006). Pathogenicity of MSH2 missense mutations is typically associated with impaired repair capability of the mutated protein. *Gastroenterology*, *131*(5), 1408-1417.
- Pal, T., Permuth-Wey, J., & Sellers, T. A. (2008). A review of the clinical relevance of mismatch-repair deficiency in ovarian cancer. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *113*(4), 733-742.

- Pardini, B., Corrado, A., Paolicchi, E., Cugliari, G., Berndt, S. I., Bezieau, S., Landi, S. (2020). DNA repair and cancer in colon and rectum: Novel players in genetic susceptibility. *International journal of cancer*, *146*(2), 363-372.
- Park, J. M., Huang, S., Tougeron, D., & Sinicrope, F. A. (2013). MSH3 mismatch repair protein regulates sensitivity to cytotoxic drugs and a histone deacetylase inhibitor in human colon carcinoma cells. *PLoS One*, *8*(5), e65369.
- Peltomäki, P. (2001). Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human molecular genetics*, *10*(7), 735-740.
- Pillot, G. A., Read, W. L., Hennenfent, K. L., Marsh, S., Gao, F., Viswanathan, A., ... & Govindan, R. (2006). A phase II study of irinotecan and carboplatin in advanced non-small cell lung cancer with pharmacogenomic analysis. *Journal of Thoracic Oncology*, *1*(9), 972-978.
- Plaschke, J., Engel, C., Krüger, S., Holinski-Feder, E., Pagenstecher, C., Mangold, E., ... & Schackert, H. K. (2004). Lower incidence of colorectal cancer and later age of disease onset in 27 families with pathogenic MSH6 germline mutations compared with families with MLH1 or MSH2 mutations: the German Hereditary Nonpolyposis Colorectal Cancer Consortium. *Journal of Clinical Oncology*, *22*(22), 4486-4494.
- Plociennik, A., Dzantiev, L., Iyer, R. R., Constantin, N., Kadyrov, F. A., & Modrich, P. (2010). PCNA function in the activation and strand direction of MutLα endonuclease in mismatch repair. *Proceedings of the National Academy of Sciences*, *107*(37), 16066-16071.
- Press, N. A. (1991), *Cells of Origin for Lung Cancer*, National Research Council (US) Panel on Dosimetric Assumptions Affecting the Application of Radon Risk Estimates. URL: *https://www.ncbi.nlm.nih.gov/books/NBK234231*
- RadiologyInfo.org (2017), 'Lung cancer treatment'
- Raevaara, T. E., Gerdes, A. M., Lönnqvist, K. E., Tybjærg-Hansen, A., Abdel-Rahman, W. M., Kariola, R., ... & Nyström-Lahti, M. (2004). HNPCC mutation MLH1 P648S makes the functional protein unstable, and homozygosity predisposes to mild neurofibromatosis type 1. *Genes, Chromosomes and Cancer*, *40*(3), 261-265.
- Ramilo, C., Gu, L., Guo, S., Zhang, X., Patrick, S. M., Turchi, J. J., & Li, G. M. (2002). Partial reconstitution of human DNA mismatch repair in vitro: characterization of the role of human replication protein A. *Molecular and cellular biology*, *22*(7), 2037-2046.
- Rami-Porta, R., Crowley, J. J., & Goldstraw, P. (2009). Review the revised TNM staging system for lung cancer. *Ann Thorac Cardiovasc Surg*, *15*(1), 5
- Refseq, Aug 2017 URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=MLH1
- Sancar, A., & Hearst, J. E. (1993). Molecular matchmakers. *Science*, *259*(5100), 1415-1420.
- Santos, L. S., Silva, S. N., Gil, O. M., Ferreira, T. C., Limbert, E.. (2018) Mismatch repair single nucleotide polymorphisms and thyroid cancer susceptibility. *Oncology Letter*, ; *15*(5), 6715-6726.

- Santos, P. A. R. D., Rocha, R. S. D., Pipkin, M., Silveira, M. L. D., Cypel, M., Rios, J. O., & Pinto, J. A. L. D. F. (2007). Concordance between clinical and pathological staging in patients with stages I or II non-small cell lung cancer subjected to surgical treatment. *Jornal Brasileiro de Pneumologia*, *33*, 647-654
- Sarasin A. An overview of the mechanisms of mutagenesis and carcinogenesis. (2003). *Mutation Research/Reviews in Mutation Research*, *544*(2-3), 99-106.
- Shankar, A., Dubey, A., Saini, D., Singh, M., Prasad, C. P., Roy, S., ... & Rath, G. K. (2019). Environmental and occupational determinants of lung cancer. *Translational lung cancer research*, *8*(Suppl 1), S31.
- Shih, C. M., Chen, C. Y., Lee, I., Kao, W. T., & Wang, Y. C. (2010). A polymorphism in the hMLH1 gene (-93G↷ A) associated with lung cancer susceptibility and prognosis. *International journal of molecular medicine*, *25*(1), 165-170.
- Shih, C., Chen, V. J., Gossett, L. S., Gates, S. B., MacKellar, W. C., Habeck, L. L., & Schultz, R. M. (1997). LY231514, a pyrrolo [2, 3-d] pyrimidine-based antifolate that inhibits multiple folate-requiring enzymes. *Cancer Research*, *57*(6), 1116-1123
- Singh, A., Singh, N., Behera, D., & Sharma, S. (2019). XRCC1 632 as a candidate for cancer predisposition via a complex interaction with genetic variants of base excision repair and double strand break repair genes. *Future Oncology*, *15*(33), 3845-3859.
- Singh, S. M., Kongari, N., Cabello-Villegas, J., & Mallela, K. M. (2010). Missense mutations in dystrophin that trigger muscular dystrophy decrease protein stability and lead to cross-β aggregates. *Proceedings of the National Academy of Sciences*, *107*(34), 15069-15074.
- Slováková, P., Majerová, L., Matakova, T., Škereňová, M., Kavcová, E., & Halašová, E. (2014). Mismatch repair gene polymorphisms and association with lung cancer development. In *Lung Cancer and Autoimmune Disorders* (pp. 15-22). Springer, Cham.
- Smith, T. R., Levine, E. A., Freimanis, R. I., Akman, S. A., Allen, G. O., Hoang, K. N., ... & Hu, J. J. (2008). Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis*, *29*(11), 2132-2138.
- Smith, T. R., Levine, E. A., Freimanis, R. I., Akman, S. A., Allen, G. O., Hoang, K. N., ... & Hu, J. J. (2008). Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis*, *29*(11), 2132-2138
- Smith, T. R., Levine, E. A., Freimanis, R. I., Akman, S. A., Allen, G. O., Hoang, K. N., ... & Hu, J. J. (2008). Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis*, *29*(11), 2132-2138.
- Smith, T. R., Levine, E. A., Freimanis, R. I., Akman, S. A., Allen, G. O., Hoang, K. N., ... & Hu, J. J. (2008). Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis*, *29*(11), 2132-2138
- Smolarz, B., & Romanowicz, H.(2018). Association between single nucleotide polymorphism of DNA repair genes and endometrial cancer: a case-control study. *International Journal of Clinical and Experimental Pathology*. *11*(3), 1732

- Smolarz, B., Makowska, M., Samulak, D., Michalska, M. M., & Romanowicz, H. (2015). Gly322Asp and Asn127Ser single nucleotide polymorphisms (SNPs) of hMSH2 mismatch repair gene and the risk of triple-negative breast cancer in Polish women. *Familial Cancer*, *14*(1), 81-88.
- Srivastava, K., Srivastava, A., & Mittal, B. (2010). Polymorphisms in ERCC2, MSH2, and OGG1 DNA repair genes and gallbladder cancer risk in a population of Northern India. *Cancer*, 116(13), 3160-3169.
- Sun, M. Z., Ju, H. X., Zhou, Z. W., Jin, H., & Zhu, R. (2014). Single nucleotide polymorphisms of DNA mismatch repair genes MSH2 and MLH1 confer susceptibility to esophageal cancer. *International journal of clinical and experimental medicine*, *7*(8), 2329.
- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, *71*(3), 209-249.

- Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians, 71*(3), 209-249.

- Surate Solaligue, D. E., Rodríguez-Castillo, J. A., Ahlbrecht, K., & Morty, R. E. (2017). Recent advances in our understanding of the mechanisms of late lung development and bronchopulmonary dysplasia. *American Journal of Physiology-Lung Cellular and Molecular Physiology*, *313*(6), L1101-L1153
- Tamura, T., Kurishima, K., Nakazawa, K., Kagohashi, K., Ishikawa, H., Satoh, H. and Hizawa, N. (2015), 'Specific organ metastases and survival in metastatic non-small-cell lung cancer', *Molecular and clinical oncology* 3(1), 217–221
- Tanaka, Y., Zaman, M. S., Majid, S., Liu, J., Kawakami, K., Shiina, H., Tokizane, T., Dhaiya, A.V., Sen, S.. & Nakajima, K. (2009). Polymorphisms of MLH1 in benign prostatic hyperplasia and sporadic prostate cancer. Biochemical and biophysical research communications, 383(4), 440-444.
- Tang, H., & Thomas, P. D. (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*, *203*(2), 635-647.
- Terrazzino, S., La Mattina, P., Masini, L., Caltavuturo, T., Gambaro, G., Canonico, P. L., ... & Krengli, M. (2012). Common variants of eNOS and XRCC1 genes may predict acute skin toxicity in breast cancer patients receiving radiotherapy after breast conserving surgery. *Radiotherapy and Oncology*, *103*(2), 199-205.
- Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Vandergriff, J. A, et al. (2003). PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research*, *31*(1), 334-341.

- Tomasini, P., Barlesi, F., Mascaux, C., & Greillier, L. (2016). Pemetrexed for advanced stage nonsquamous non-small cell lung cancer: latest evidence about its extended use and outcomes. *Therapeutic advances in medical oncology*, *8*(3), 198-208.
- Tulupova, E., Kumar, R., Hanova, M., Slyskova, J., Pardini, B., Polakova, V., ... & Vodicka, P. (2008). Do polymorphisms and haplotypes of mismatch repair genes modulate risk of sporadic colorectal cancer?. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *648*(1-2), 40-45.
- *Type of Tumors* (2016). URL: *http://pathology.jhu.edu/pc/BasicTypes1.php?area=ba*
- Vogelsang M, Wang Y, Veber N, *et al.*, The cumulative effects of polymorphisms in the DNA mismatch repair genes and tobacco smoking in oesophageal cancer risk. PloS one. 2012 May 18;7(5):e36962.
- Vogelsang, M., Paccez, J. D., Schäfer, G., Dzobo, K., Zerbini, L. F., & Parker, M. I. (2014). Aberrant methylation of the MSH3 promoter and distal enhancer in esophageal cancer patients exposed to first-hand tobacco smoke. *Journal of cancer research and clinical oncology*, *140*(11), 1825-1833.
- Vymetalkova, V., Pardini, B., Rosa, F., Di Gaetano, C., Novotny, J., Levy, M., ... & Vodicka, P. (2014). Variations in mismatch repair genes and colorectal cancer risk and clinical outcome. *Mutagenesis*, *29*(4), 259-265.
- Wang, D., Chen, D., Shen, B., Chen, X., Xiao, M., Li, X., & Shu, Y. (2019). A polymorphism within the mismatch repair gene predicts prognosis and adjuvant chemotherapy benefit in gastric cancer patients. *Gastric Cancer*, *22*(6), 1121-1129
- Whiffin, N., Broderick, P., Lubbe, S. J., Pittman, A. M., Penegar, S., Chandler, I., & Houlston, R. S. (2011). MLH1-93G>A is a risk factor for MSI colorectal cancer. *Carcinogenesis*, *32*(8), 1157-1161
- Wilson, T., Guerrette, S., & Fishel, R. (1999). Dissociation of mismatch recognition and ATPase activity by hMSH2-hMSH3. *Journal of Biological Chemistry*, *274*(31), 21659-21664.
- Winsey, S. L., Haldar, N. A., Marsh, H. P., Bunce, M., Marshall, S. E., Harris, A. L., ... & Welsh, K. I. (2000). A variant within the DNA repair gene XRCC3 is associated with the development of melanoma skin cancer. *Cancer Research*, *60*(20), 5612-5616.
- Worrillow, L. J., Travis, L. B., Smith, A. G., Rollinson, S., Smith, A. J., Wild, C. P., ... & Allan, J. M. (2003). An intron splice acceptor polymorphism in hMSH2 and risk of leukemia after treatment with chemotherapeutic alkylating agents. *Clinical cancer research*, *9*(8), 3012-3020.
- Wu, T. H., & Marinus, M. G. (1994). Dominant negative mutator mutations in the mutS gene of Escherichia coli. *Journal of bacteriology*, *176*(17), 5393-5400
- Xiao, X. Q., Gong, W. D., Wang, S. Z., Zhang, Z. D., Rui, X. P., Wu, G. Z., & Ren, F. (2012). Polymorphisms of mismatch repair gene hMLH1 and hMSH2 and risk of gastric cancer in a Chinese population. *Oncology letters*, *3*(3), 591-598.

- Xinarianos, G., Liloglou, T., Prime, W., Maloney, P., Callaghan, J., Fielding, P., ... & Field, J. K. (2000). hMLH1 and hMSH2 expression correlates with allelic imbalance on chromosome 3p in non-small cell lung carcinomas. *Cancer Research*, *60*(15), 4216-4221.
- Xu, X. L., Yao, Y. L., Xu, W. Z., Feng, J. G., & Mao, W. M. (2015). Correlation of MSH3 polymorphisms with response and survival in advanced non-small cell lung cancer patients treated with first-line platinum-based chemotherapy. *Genetics and Molecular Research*, *14*(2), 3525-33.
- Yadav, S. K., Sanyal, S., Gupta, S., Bhatt, M. L. B., Mishra, D. P., & Roy, D. (2018). Single nucleotide polymorphism of MSH3 gene alters head and neck squamous-cell carcinoma risk in North-India. *International Journal of Cancer Research*, *14*(1), 27-31.
- Yamamoto, A., Yamaguchi, T., Suzuki, O., Ito, T., Chika, N., Kamae, N., Ishida, H, et al. (2021). Prevalence and molecular characteristics of DNA mismatch repair deficient endometrial cancer in a Japanese hospital-based population. *Japanese Journal of Clinical Oncology*, *51*(1), 60-69
- Yong, Y. and Lin, H. (2005), 'Shesis, a powerful software platform for analyses of linkage disequilibrium, haplotype construction, and genetic association at polymorphism loci',*Cell research* 15(2), 97.
- Zanusso, C., Bortolus, R., Dreussi, E., Polesel, J., Montico, M., Cecchin, E., ... & Toffoli, G. (2017). Impact of DNA repair gene polymorphisms on the risk of biochemical recurrence after radiotherapy and overall survival in prostate cancer. *Oncotarget*, *8*(14), 22863.
- Zelga P, Przybyłowska-Sygut K, Zelga M, et al., Polymorphism of Gly39Glu (c. 116G> A) hMSH6 is associated with sporadic colorectal cancer development in the Polish population: Preliminary results. *Adv Clin Exp Med.* 2017;26(9):1425-9
- Zelga, P., Przybyłowska-Sygut, K., Zelga, M., Dziki, A., & Majsterek, I. (2017). Polymorphism of Gly39Glu (c.116G>A) hMSH6 is associated with sporadic colorectal cancer development in the Polish population: Preliminary results. *Advances in Clinical and Experimental Medicine*, *26*(9).
- Zelga, P., Przybyłowska-Sygut, K., Zelga, M., Dziki, A., & Majsterek, I. (2018). The 116G> A MSH6 and IVS1-1121C> T PMS2 Genes Polymorphisms Modulate the Risk of the Sporadic Colorectal Cancer Development in Polish Population. *Pathology & Oncology Research*, *24*(2), 231-235.
- Zhu, X., Wang, Z., Qiu, X., Wang, W., Bei, C., Tan, C., ... & Tan, S. (2018). Rs2303428 of MSH2 is associated with hepatocellular carcinoma prognosis in a Chinese population. *DNA and Cell Biology*, *37*(7), 634-641.

Appendix

Appendix-I Ethical Committee Report

Institutional Ethics Committee
Postgraduate Institute of Medical Education and Research, Chandigarh

Prof. K.N. Pathak
Chairman

Phone No. 0172-2755266

Dr. Samir Malhotra
Convener

No. PGI/IEC/2018/ 001310

Dated: - 08-09-2018

To: Dr. Navneet Singh, Dept. of Pulmonary Medicine, PGIMER, Chandigarh.
From: IEC Office
Date: 29.08.2018
IEC #: IEC-04/2018-884
Subject: Genetic polymorphism in phase I/II and ABC transporter gens in lung cancer pateint undergoing platinum bases chemotherapy and its association with toxcity and clinical outcome.

The Institutional Ethics Committee of PGIMER, Chandigarh reviewed and discussed the above mentioned Project in its meeting, which was held on 14/15.05.2018 and sent the observations for your clarification, after receiving your reply the same is APPROVED by Institutional Ethics committee. The following documents were reviewed:-

(a) Study Protocol (including amendments, if any)
(b) Patient Information Sheet and Informed Consent Form (including updates, if any)
(c) Principal Investigator's current CV.
(d) IEC form

The following members of the IEC were present at the meeting held on 29.08.2018 at 03:00 PM in the Board Room of Kairon Block, PGIMER, Chandigarh:-

Sr. No	Name of the member	Designation	Affiliation to PGI
1.	Prof. K. N. Pathak	Chairman	No
2.	Justice V. K. Jhanji	Member (Legal)	No
3.	Dr. Virender Kumar	Member (Lay Person)	No
4.	Dr. O. P. Katare	Member (Basic Scientist)	No
5.	Dr. Sukhbir Kaur,	Member (Basic Science)	No
6.	Dr. Anil Monga	Member (Social Scientist)	No
7.	Dr. Pallab Ray	Member (Basic Science)	Yes
8.	Dr. Ashish Bhalla	Member (Clinician)	Yes
9.	Dr. Neelam Aggarwal	Member (Clinician)	Yes
10.	Dr. Ritambra Nadda	Member (Basic Sciences)	Yes
11.	Dr. Kanya Mukhopadhyay	Member (Clinician)	Yes
12.	Dr. Manish Modi	Member (Clinician)	Yes
13.	Dr. Nusrat Shafiq	Member (Pharmacologist)	Yes
14.	Dr. P.V.M. Lakshmi	Member	Yes
15.	Dr. Samir Malhotra	Convener	Yes

The Institutional Ethics Committee expects to be informed about the progress of the study, please use the Annual Progress Report Form available on the PGIMER website. Please note that it is the responsibility of the PI to inform the IEC/DSMC regarding any SAE occurring in the course of the study, or any changes in the protocol and patient information sheet /informed consent form, if applicable. You are also requested to provide a copy of the final report in the Study Completion Form available on the PGIMER website. Please be advised that your study can be periodically monitored by the IEC.

Yours sincerely,

(Samir Malhotra)
Convener
Instt. Ethics Committee, PGIMER, Chandigarh.

INSTITUTIONAL ETHICS COMMITTEE (IEC)

Department of Biotechnology

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed-to-be-University u/s 3 of the UGC Act, 1956)
Thapar Technology Campus,
Bhadson Road, Patiala - 147 004 (Punjab) India
Phone : +91-175-2393021
Email : registrar@thapar.edu
URL : www.thapar.edu

Review Letter number IEC/TIET/DBT/09/03/18

Dated: 27.09.2018

The Institutional ethics committee of Thapar Institute of Engineering & Technology, Patiala, in its Sixth (06) meeting held on 27.09.2018, in the office of Head, Department of Biotechnology, reviewed and discussed the project submitted by Dr. Siddharth Sharma as PI from Department of Biotechnology.

TIET/DBT/IEC/09/03/18: Genetic polymorphism in Phase I/II and ABC transporter genes in Lung Cancer Patient undergoing Platinum based chemotherapy and its association with toxicity and clinical outcomes.

The members reviewed the following documents and raised queries which were satisfactorily answered by Dr. Siddharth Sharma.

a) Proposal & Protocol

b) Available Safety Information

c) Cell line/ Genetic material storage and discard procedure

It is understood that the study would be conducted under the PI's direction strictly as per the submitted protocol. In case of any deviation from the approved methods, the study documents will have to be resubmitted to ethics committee for re-approval.

Members	**Signature**
Dr. Manoj Baranwal (Member)	
Dr. Siddharth Sharma (Member)	
Dr. Vikas Handa (Member)	
Dr. Sanjai Saxena (Member Secretary)	

The present approval is for 01 year only; the investigators are advised to take a re-approval after one year. 06 Monthly compliance report is a must for seeking re-approval.

Signature

Prof. Moushumi Ghosh Moushumi Ghosh

Chairman, Institutional Ethics Committee

HEAD
Department of Biotechnology
Thapar Institute of Engineering & Technology
Patiala - 147004, Punjab

POSTGRADUATE INSTITUTE OF MEDICAL EDUCATION & RESEARCH, CHANDIGARH
MEDICAL EDUCATION AND RESEARCH CELL
PHONE: 0172-2755568

No: 79/179-Edu-18 /4913 Dated: 05/12/2018

To

Dr. Navneet Singh
Additional Professor
Department of Pulmonary Medicine
PGIMER, Chandigarh

Subject: Collaborative Research project titled **"Genetic Polymorphism in Phase I/II and ABC Transporter Genes in Lung Cancer patient undergoing platinum based chemotherapy and its association with toxicity and clinical outcomes"**

Please refer to your letter no. 2765 dated 02/11/2018 on the subject cited above.

Approval of the ICRC Committee is herewith conveyed for the collaborative research project titled "Genetic Polymorphism in Phase I/II and ABC Transporter Genes in Lung Cancer patients undergoing platinum based chemotherapy and its association with toxicity and clinical outcomes"

GSingh 5/12

Sub-Dean (Research)
Medical Education & Research Cell
PGIMER, Chandigarh

Appendix-II PATIENT CONSENT FORM

Patient consent form

Name of the participant: ____

Name of the Principal (Co-) Investigators: ______________________________

Name of the Institution: ____

Name and address of the sponsoring (funding) agency(ies): ________________

____

Documentation of the informed consent

I, … … … … … … … … … . have read the information in this form (or it has been read to me). I was free to ask any questions and they have been answered. I am over 18 years of age and, exercising my free power of choice, hereby give my consent to be included as a participant in the project "**Genetic Polymorphisms in Phase I/II and ABC Transporter Genes in Lung Cancer patient undergoing platinum based chemotherapy and its association with toxicity and clinical outcomes**"

(1) I have read and understood consent form and the information provided to me.

(2) I have had the consent document explained to me.

(3) I have been explained about the nature of the study.

(4) My rights and responsibilities have been explained to me by the investigator.

(5) I have the right to not be a subject in the above study and am not willing to give sample

(6) I have informed the investigator of all the treatments I am taking or have taken in the past … … . months including any *desi* (alternative) treatments.

(7) Patient data and confidentiality has been explained to me.

(8) I hereby give permission to the investigators to release the information obtained from me as result of participation in this study to the sponsors, regulatory authorities, Government agencies, and ethics committee. I understand that they may inspect my original records.

(9) My identity will be kept confidential if my data are publicly presented.

(10) I have had my questions answered to my satisfaction.

(11) I have decided to be in the research study.

I am aware, that if I have any questions during this study, I should contact the concerned authorities. By signing this consent from, I attest that the information given in this document. I will be given a copy of this consent document.

For adult participants

Name and signature / thumb impression of the participant (or legal representative if participant incompetent):

__________________ (Name) ____________ (Signature)

Date: ___________ Time: _________

Name and signature of impartial witness (required for illiterate patients):

___________________ (Name) ____________ (Signature)

Date:___________ Time: __________

Address and contact number of the impartial witness: _____________________

____

Name and signature of the Investigator or his representative obtaining consent:

_____________________ (Name) __________________ (Signature)

_____________________ (Date)

Investigator Certificate

I certify that all the elements including the nature of the research study and possible risks of the above study as described in this consent document have been fully explained to the subject. In my judgment, the participant possesses the legal capacity to give informed consent to participate in this research and is voluntarily and knowingly giving informed consent to participate,

Signature of the Investigator: ________________ Dated:__________

Name of the Investigator: ___________________

Appendix-III PATIENT INFORMATION SHEET

PATIENT INFORMATION SHEET

Department of Biotechnology, Thapar University &

Dept. of Pulmonary Medicine, PGIMER, Chandigarh

Investigator: Siddharth Sharma, Ph.D

Co-Investigator: DrNavneet Singh, M.D, D.M

Name of Patient ____

Address(Rural/Urban) ____

C.R Number ____

Date of Sampling ____

Type of Cancer

____

Stage of Cancer ____

Sex ____

Age ____

Education (Years) ____

Occupation ____

Marital Status

____

Smoking Status

1. Smoker or Non-Smoker ____

2. Type of Smoking

a) Cigarette b) Beedi c) Hookah d) Cigarette+Beedi e) Beedi+Hookah f) Any other form of smoke

3. No. of Cigarettes/ Beedis smoked

a) Per day____

b) Per Week ____

c) Amount of tobacco added to hookah ____

4. a) Age at which started Smoking ____

b) How many years have you smoked ________

5. Time taken to finish cigarette (in min) __________________

6. Brand of smoke preferred ________________

7. Do you chew tobacco/betelnut

a) Yes b) No

If yes amount chewed ______________________

8. Any familial case of Smoking related cancer in your family: a) Yes b) No

9. Are you only one to smoke in your family

Alcohol Consumption

1. Do you consume alcohol

a) Yes b) No

2. Intake of Alcohol

a) Daily b) 3-5 times a week c) 1-2 times a week d) once a week e) once a month

3. How much alcohol do you consume in ml_________________

4. What type of alcohol do you consume ____________________

Documentation of the informed consent

I, … … … … … … … … … .have read the information in this form (or it has been read to me). I was free to ask any questions and they have been answered. I am over 18 years of age and, exercising my free power of choice, hereby give my consent to be included as a participant in the project "**Association ofSingle Nucleotide Polymorphisms and Methylation status in the Wnt and AhR signaling pathways with risk for occurrence of Lung Cancer in a North Indian Population** "

(1) I have read and understood consent form and the information provided to me.

(2) I have had the consent document explained to me.

(3) I have been explained about the nature of the study.

(4) My rights and responsibilities have been explained to me by the investigator.

(5) I have the right to not be a subject in the above study and am not willing to give sample

(6) I have informed the investigator of all the treatments I am taking or have taken in the past … … . months including any *desi* (alternative) treatments.

(7) Patient data and confidentiality has been explained to me.

(8) I hereby give permission to the investigators to release the information obtained from me asresult of participation in this study to the sponsors, regulatory authorities, Government agencies, and ethics committee. I understand that they may inspect my original records.

(9) My identity will be kept confidential if my data are publicly presented.

(10) I have had my questions answered to my satisfaction.

(11) I have decided to be in the research study.

I am aware, that if I have any questions during this study, I should contact the concerned authorities. By signing this consent from, I attest that the information given in this document. I will be given a copy of this consent document.

Appendix-IV

- 0.5M EDTA: Dissolved 9.306g of disodium salt of EDTA in 20ml of deionised water, and then adjusted the pH to 8.0 by 1 M sodium hydroxide. Sterilized the solution by autoclaving.
- 10% SDS: Dissolved 1g of SDS in 10ml of deionised water.
- 3. 100mM Tris-Cl (pH 8.0): Dissolved 0.32g of Tris-Cl in 10 ml of deionised water, then adjusted the pH to 8.0 by 1M sodium hydroxide. Sterilized the solution by autoclaving.
- 10mg/ml Proteinase K: Dissolved 10mg Proteinase K in 1ml of double distilled water. Sterilized the solution by autoclaving.
- 1mg/ml BSA: Dissolved 100mg of BSA in 100ml of deionised sterile water and kept at 4 ˚C overnight.
- 5M Sodium chloride (NaCl): Dissolved 5.85g of sodium chloride in 20ml of deionised water. Sterilized the solution by autoclaving.
- 5X TBE buffer: Dissolved 54g of Tris base and 27.5g of boric acid in 980ml of double distilled water and then added 20ml of 0.5 EDTA. Sterilized the solution by autoclaving.
- Ethidium Bromide (10mg/ml): Dissolved 1g of ethidium bromide in 100ml of water. Mixed the solution properly.
- Magnesium chloride ($MgCl_2$) (100mM): Dissolved 0.41gms of $MgCl_2$ in 20ml of deionised water and sterilized by autoclaving.
- Sucrose (1M): Dissolved 3.41 g of sucrose in 10 ml of deionised water and sterilized by autoclaving.
- TE buffer (pH 8.0): Added 1ml of 100mM Tris-Cl (pH 8.0) and 200 µl of 0.5M EDTA solution to 8.8 ml of deioinsed water. Sterilized the solution.
- Triton X- 100 (10%): Took 100 µl of TritonX-100 and mixed with 900 µl of deionised water and mixed properly.

Original Article

Association of *MLH1-93G>A* polymorphisms toward lung cancer susceptibility and its association with clinical outcome in North Indian patients treated with platinum-based chemotherapy

ABSTRACT

Background: Lung cancer is one of the most prevalent and main causes of malignancy-related deaths worldwide, especially in developed countries. Epidemiological studies have demonstrated that individuals having alterations in a particular gene may have a high risk of developing certain types of cancer.

Materials and Methods: In the present study, 500 Indian lung cancer patients and 500 healthy controls were enrolled. Polymerase chain reaction-restriction fragment length polymorphism method was used to identify the genotype of enrolled individuals and MedCalc statistical package was used for carrying out statistical analysis.

Results: In this study, we found a reduced risk of developing adenocarcinoma in patients harboring variant ($P = 0.0007$) and combined type genotype ($P = 0.008$), whereas an increased risk for small-cell lung carcinoma (SCLC) development for those subject harboring GA genotypes ($P = 0.03$) was also observed. Further, heterozygous type and combined type genotype of heavy smokers for MLH1 polymorphism reported a 2-fold ($P = 0.001$) and 1.8-fold increased risk toward lung cancer development, respectively ($P = 0.007$). In case of females, the subjects harboring a variant allele have a significantly reduced risk for lung cancer development ($P = 0.0001$). For MLH1 polymorphism, reduced risk of developing tumor to T3 or T4 stage was observed ($P = 0.04$). Moreover, this is the first study reporting overall survival (OS) association for north Indian lung cancer patients with platinum-based doublet chemotherapy; for docetaxel, a three-fold increase in hazard ratio and corresponding low median standard survival time (8.4 months) for mutant and combined type genotype ($P = 0.04$) was observed.

Conclusions: These results suggest that MLH1-93G>A polymorphism is involved in modulating the risk toward lung cancer. Our study also concluded a negative association of OS in patients undergoing carboplatin/cisplatin and docetaxel chemotherapy.

KEY WORDS: Chemotherapy, lung cancer, MLH1, overall survival, pack-years, polymorphism

Sidhartha Singh, Navneet Singh[1], Manoj Baranwal, Siddharth Sharma

Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab, [1]Department of Pulmonary Medicine, Post Graduate Institute of Medical Education and Research, Chandigarh, India

For correspondence:
Dr. Siddharth Sharma, Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala - 147 004, Punjab, India.
E-mail: siddharthsharma.phd@thapar.edu

Submitted: 19-Mar-2021
Accepted in revised form: 09-Sep-2021
Published: 14-Jan-2022

INTRODUCTION

Lung cancer has become the single most common cause of death occurring due to cancer and hence has become a big health problem all over the world.[1] Although a lot of studies suggest that smoking is the leading factor for the occurrence of cancer, only about 15% of lung cancer patients are smokers proposing that there might be some others factors involved in the development of cancer.[2,3] Several Epidemiological studies have also reported that alteration in specific genes (DNA repair pathways) can increase the risk of developing different types of cancer.[4,5] MMR is a DNA repair pathway which is responsible for recognizing and repairing the mismatches occurring due to insertion, deletion, and misincorporation of bases during the process of replication and recombination.[6] MLH1, a DNA repair protein helps fix errors that are made when DNA is copied in preparation for cell division.

Access this article online
Website: www.cancerjournal.net
DOI: 10.4103/jcrt.jcrt_465_21
Quick Response Code:

Cite this article: Singh S, Singh N, Baranwal M, Sharma S. Association of *MLH1-93G>A* polymorphisms toward lung cancer susceptibility and its association with clinical outcome in North Indian patients treated with platinum-based chemotherapy. J Can Res Ther 0;0:0.

Research Article

For reprint orders, please contact: reprints@futuremedicine.com

Biomarkers in Medicine

Polymorphisms in the *MSH2* gene predict poor survival of North Indian lung cancer patients undergoing chemotherapy

Sidhartha Singh[1], Navneet Singh[2], Manoj Baranwal[1] & Siddharth Sharma*,[1]

[1]Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, 147001, India

[2]Department of Pulmonary Medicine, Post Graduate Institute of Medical Education & Research (PGIMER), Chandigarh, 160012, India

*Author for correspondence: Tel.: +91 950 168 8366; siddharthsharma.phd@thapar.edu

Aim: To estimate if *MSH2* polymorphisms, viz. rs63749993, rs2303425, rs2303426, rs4987188, rs2303428 and rs17217772, have any association with clinical outcomes in North Indian lung cancer patients. **Materials & methods:** PCR-RFLP was used for genotyping 500 cases. Logistic regression and survival analysis was performed by utilizing MedCalc software. **Results & conclusion:** Our study concluded, adenocarcinoma subjects having heterozygous genotype for rs2303425 have increased survival time (MST = 12.43, p = 0.03). In lung cancer patients undergoing paclitaxel therapy, heterozygous carriers for the rs17217772 polymorphism have reduced survival time (MST = 7.96 vs 2.63 months; HR = 2.09; p = 0.02). For rs63749993 polymorphism undergoing irinotecan therapy, subjects having mutant genotype showed poor survival (13.26 vs 6.06 months; HR = 5.37; p = 0.0004). The results suggest that *MSH2* polymorphisms are involved in decreasing overall survival for patients undergoing platinum-based chemotherapy.

First draft submitted: 15 July 2021; Accepted for publication: 15 November 2021; Published online: 27 January 2022

Keywords: chemotherapy • lung cancer • *MSH2* • polymorphism • survival

Lung cancer is a complex and multifactorial carcinoma that occurs because of the complex interaction between various environmental and genetic factors [1]. Thus, it implies that the lung carcinogenesis process is not under the control of a single gene or the only environmental factor responsible for the occurrence of this disease; somewhat, multiple interactions at the gene level affect the cancer susceptibility of a particular individual. Although environmental factors such as tobacco smoke, radiation and infectious agents increase the carcinogenic load to which humans are exposed, the exact mechanism underlying the aggressive nature of cancer to metastasize is mainly unknown [2]. Moreover, it is still unclear why certain people develop cancer even though they might not be exposed to any carcinogens. In contrast, some individuals, despite being exposed to carcinogens and various detrimental environmental factors, may not develop the disease [3]. Hence, uncovering the genes and pathways associated with lung cancer occurrence may help understand the mechanism of cancer metastasis, development of new diagnostic methods and molecules that have significant therapeutic potential [4]. Various studies have reported that genetic changes such as nucleic acid sequence mutations are involved in increasing the susceptibility of a particular individual toward lung cancer [5]. Previous GWAS studies have also revealed that rs7629386, rs969088 and rs3850370 polymorphisms were associated with poor NSCLC survival and rs41997, rs12000445 polymorphism were associated with better NSCLC survival in Chinese Han population [6]. Another study conducted by Tang and co-workers have also reported that rs10023113 polymorphism has also been associated with poor prognosis of early-stage NSCLC patients of Chinese origin [7]. Earlier studies conducted in Asian population on *CLPTM1L*, *TP63* and *MIPEP* genes have reported association of these polymorphisms with the advent of lung cancer [8–10].

Whenever any genomic injury such as single and double-strand breaks in DNA occurs due to different risk factors, the DNA repair system gets activated to repair those damages [11]. A plethora of research has delved into the role of genes involved in DNA repair pathways to find any association with susceptibility toward lung cancer [12,13].

Future Medicine

Received: 24 May 2021 | Accepted: 23 November 2021
DOI: 10.1002/bab.2295

WILEY

ORIGINAL ARTICLE

Identification of SNPs in hMSH3/MSH6 interaction domain affecting the structure and function of MSH2 protein

Sidhartha Singh | **Siddharth Sharma** | **Manoj Baranwal**

Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, Punjab, India

Correspondence
Manoj Baranwal, Department of Biotechnology, Thapar Institute of Engineering & Technology, Patiala, Punjab, 147004, India.
Email: manoj.baranwal@thapar.edu

Abstract

MutS homolog 2 (MSH2) is a mismatch repair gene that plays a critical role in DNA repair pathways, and its mutations are associated with different cancers. The present study aimed to find out the single nucleotide polymorphisms (SNPs) of MSH2 protein associated with causing structural and functional changes leading to the development of cancer with the help of computational tools. Four different tools for predicting deleterious SNPs (SIFT, PROVEAN, PANTHER, and PolyPhen), two tools each for identifying disease association (PhD-SNP and SNP&GO) and estimating stability (I-mutant and MUPro) were employed. Homology modeling, energy minimization, and root mean square deviation calculation were used to estimate structural variations. Twenty-seven SNPs and five SNPs (double amino acid change) were identified based on a consensus approach that might be associated with the structural and functional change in MSH2 protein. Molecular docking reveals that six SNPs affect the interaction of MSH2 and MSH6. Twelve identified SNPs were reported to be linked with hereditary nonpolyposis, colorectal cancer, and Lynch syndrome. Further, selected SNPs need to be validated in an in vitro system for their precise association with cancer predisposition.

KEYWORDS
cancer, DNA repair, in silico analysis, polymorphism, protein docking

1 | INTRODUCTION

MutS Homolog 2 (MSH2) protein plays an essential role in DNA repair and is associated with different cancer such as hereditary nonpolyposis colorectal cancer (HNPCC), Lynch syndrome, and lung cancer.[1–3] MSH2 comprises two interaction domains (MSH3/MSH6 and MSH2/PMS2) and one DNA binding domain, which are present in two different locations on the gene (Figure 1). MSH2 binds with MSH6/MSH3 to form MutSα/MutSβ complexes. DNA repair process is initiated by the binding of these complexes to the mismatches present in the DNA molecule.[4] MutSα binds to single base mismatches and dinucleotide indel, whereas MutSβ only binds to indel mismatches. After binding to mismatches, the MutS and MutL complex (comprising MSH2 and PMS2) interacts with each other and subsequently directs the downstream mismatch repair

Abbreviations: GWAS, Genome-wide association studies; HBA1, Hemoglobin Subunit Alpha 1; HNPCC, Hereditary nonpolyposis colorectal cancer; MMR, Mismatch Repair; MSH2, MutS Homolog 2; nsSNP, Non-synonymous single nucleotide polymorphisms; PANTHER, protein analysis through evolutionary relationships; PhD-SNP, Predictor of human Deleterious Single Nucleotide Polymorphisms; PolyPhen 2, Polymorphism Phenotyping v2; PROVEAN, Protein Variation Effect Analyzer; SIFT, Sorting Intolerant from Tolerant; SNP, Single nucleotide polymorphisms; SVM, Support vector machine.