

**COMPUTATIONAL STUDIES FOR THE  
IDENTIFICATION AND ANALYSIS OF CANDIDATE  
BIOMARKERS FOR HUMAN DISEASES IMPLICATED  
IN DNA REPAIR SYSTEM**

**A THESIS SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR  
THE DEGREE OF DOCTOR OF PHILOSOPHY**

**IN**

**BIOINFORMATICS**

**By**

**MANIKA SEHGAL**

**Enrollment No. 116501**

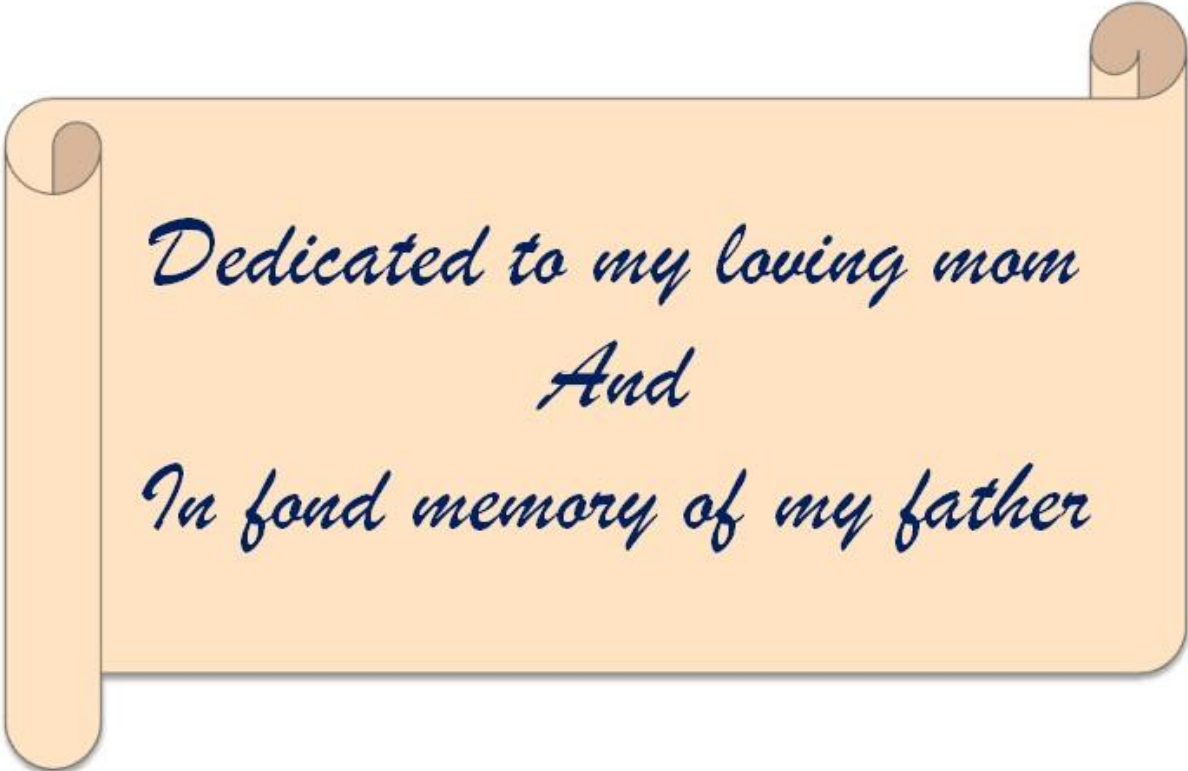


**JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY**

**WAKNAGHAT**

**FEBRUARY, 2015**

Copyright  
@  
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,  
WAKNAGHAT  
February, 2015  
ALL RIGHTS RESERVED



*Dedicated to my loving mom  
And  
In fond memory of my father*



## JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

(Established by H.P. State Legislative Vide Act no. 14 of 2002)  
Waknaghat, P.O. Dumehar Bani, Kandaghat, Distt. Solan – 173215 (H.P.) INDIA

Website: [www.juit.ac.in](http://www.juit.ac.in)

Phone No. (91) 01792-257999(30 Lines).

Fax: (91) 01792 245362

### DECLARATION

I certify that:

- a. The work contained in this thesis is original and has been done by me under the guidance of my supervisor.
- b. The work has not been submitted to any other organisation for any degree or diploma.
- c. Wherever, I have used materials (data, analysis, figures or text), I have given due credit by citing them in the text of the thesis.

**Manika Sehgal**

(Enrollment No. 116501)

Department of Biotechnology and Bioinformatics

Jaypee University of Information Technology, Waknaghat, India

Email: manika.sehgal22@gmail.com

**Date:** 09-02-2015



## JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

(Established by H.P. State Legislative Vide Act no. 14 of 2002)

Waknaghat, P.O. Dumehar Bani, Kandaghat, Distt. Solan – 173215 (H.P.) INDIA

Website: [www.juit.ac.in](http://www.juit.ac.in)

Phone No. (91) 01792-257999(30 Lines).

Fax: (91) 01792 245362

### CERTIFICATE

This is to certify that the thesis entitled, “**Computational studies for the identification and analysis of candidate biomarkers for human diseases implicated in DNA repair system**” which is being submitted by **Manika Sehgal (Enrollment No. 116501)** in fulfillment for the award of degree of **Doctor of Philosophy in Bioinformatics at Jaypee University of Information Technology, India** is the record of candidate’s own work carried out by her under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

**Dr. Tiratha Raj Singh**

Assistant Professor (Senior Grade)

Department of BT and BI

Jaypee University of Information Technology, India

**Email:** [tiratharaj@gmail.com](mailto:tiratharaj@gmail.com)

**Date:** 09-02-2015

## ACKNOWLEDGEMENT

***GOD is my ultimate strength and power who bring best out of me***

*I deem it my privilege and honour to consign my gratitude and indebtedness to the following without whose guidance, support and concern I would not have been able to complete my Ph. D thesis.*

*I feel fortunate to express my deep sense of reverence and gratitude to my revered mentor, **Dr. Tiratha Raj Singh**, for his support, immaculate guidance, constructive criticism, constant encouragement and providing requisite facilities to persist my research which otherwise would have remained incomplete. His nurturing and gentle concern has been a stimulus which I will always cherish. I also appreciate his untiring efforts during the entire tenure of my research work and patience during writing of this thesis. I have no words to express my gratitude for everything he has contributed in my Ph. D and without his blessings it was surely impossible for me to finish my work. I thank him from bottom of my heart.*

*I emphatically express my loyal and venerable thanks to **Prof. Shiban Kishen Kak** (Vice Chancellor, JUIT), **Brig. (Retd.) Balbir Singh** (Director, JUIT), **Prof. T. S.Lamba** (Dean, Academic & Research, JUIT) and **Dr. Y. Medury** (Ex-COO, Jaypee Education System) for providing opportunity to pursue a Doctorate Degree, teaching assistantship and advanced lab infrastructure to accomplish this scientific venture of my life. (Ability is of little account without opportunity).*

*I gratefully acknowledge the help rendered by **Prof. R. S. Chauhan** (Dean and HOD, Dept. of BT & BI) for his encouragement, timely help and cooperation throughout my research work. It gives me immense pleasure to express my gratitude to him for his ever smiling and positive disposition coming to my rescue in solving my problems and suggestions which helped me in maintaining my confidence.*

*I also feel indebted to my former advisor, **Dr. Nandita Bachhawat** (Former Head, Department of Bioinformatics, D.A.V. College, Panjab University, Chandigarh) for her valuable advice and unconditional support throughout since my Masters days. I would like to convey my special sincere gratitude to **Dr. Ahmed Moussa** (LabTIC Laboratory, ENSA, Abdelmalek Essaadi University, Tangier, Morocco) for his valuable guidance and sustained support during his visit to our Department as CV Raman Fellow.*

*This document would have remained an infant had it not received its necessary 'diet' in the form of comments, suggestions etc. by **Dr. Ragini Raj Singh** (Department of Physics and Material Science, JUIT), **Dr. Jayashree Ramanna**, **Dr. Dipankar Sengupta**, **Dr. Chittaranjan Rout**, **Dr. Chanderdeep Tandon** and **Dr. Hemant Sood**. I am short of words in expressing my thanks to them, for their innovative ideas that shaped this document. I wish to convey my sincere thanks to all the faculty members of Department of Biotechnology and Bioinformatics, for their help and guidance at the various stages of this study. I am also thankful to all the members of technical and non-technical staff of the department, especially **Mrs. Somlata Sharma** and **Mrs. Mamta** for their assistance and valuable contributions.*

*I am fortunate to have friends who have always stood beside me. I extend my heartfelt thanks to my sister **Kanika** and friends **Resham**, **Richa**, **Madhvi** and **Gargi** who have been my angels and always listened to my insane thoughts and ideas in spite of not understanding a bit of it. I am also grateful to my friends **Achla**, **Tamanna**, **Kritika**, **Rohit**, **Archit Sir**, **Nikhil Sir**, **Jibesh Sir**, **Tarun**, **Mani Mam**, **Shipra**, **Shifa** and **Amisha** for they have played a very important role by being there so as to make me sail through all the twists and turns of this journey. I also thank **Krishna Sir**, **Priya Mam**, **Ankita** and **Ashwani** for their sustained support and ever needed cooperation. It is my pleasure to express my gratitude to all research scholars of the Biotechnology & Bioinformatics Department for keeping me blessed with best wishes.*

*The purpose of this acknowledgement will be incomplete if I fail to appreciate the moral support and encouragement rendered by the most important person of my life i.e. my mom **Mrs. Promilla Sehgal**, who has always been my inspiration and never let me feel abandoned. I am also pleased for the support rendered by my entire family and fiancée, **Rohit Randhawa**. I express my love to my sweet niece and nephew **Tanishka** and **Tanveer** whose innocence refreshed me during my difficult times.*

*The successful culmination of this thesis could not have been possible without the blessings of my father and grandfather who always wished happiness and success for me in their lifetime. It was their blessings and love which constantly motivated me and helped in the successful completion of my research work. Without their inspiration and assistance throughout I would have been greatly incapacitated.*

*I would like to express my heartfelt gratitude to all those who have contributed directly or indirectly towards obtaining my doctorate degree and apologize if have missed out anyone.*

**Manika Sehgal**

---

**TABLE OF CONTENTS**

<b>CERTIFICATE</b>	<b>III</b>
<b>DECLARATION</b>	<b>IV</b>
<b>ACKNOWLEDGMENT</b>	<b>V-VI</b>
<b>LIST OF TABLES</b>	<b>XI</b>
<b>LIST OF FIGURES</b>	<b>XII-XIV</b>
<b>ABBREVIATIONS</b>	<b>XV-XVIII</b>
<b>ABSTRACT OF THE DISSERTATION</b>	<b>1-6</b>

<b>CHAPTER 1</b>
------------------

<b><i>INTRODUCTION</i></b>	<b>7-52</b>
----------------------------	-------------

<b>1.1 INTRODUCTION.....</b>	<b>9</b>
<b>1.2 DNA DAMAGE: GENOMIC INSTABILITY .....</b>	<b>11</b>
1.2.1 Damage classification based on factors causing the damage.....	12
1.2.1.1 Endogenous Damage.....	12
1.2.1.2 Exogenous Damage.....	13
1.2.2 Damage classification based on the consequences of damages .....	13
<b>1.3 DNA REPAIR SYSTEM .....</b>	<b>14</b>
1.3.1 Direct reversal of damage (DRD) .....	16
1.3.2 Single-strand damage .....	18
1.3.2.1 Base excision repair (BER).....	18
1.3.2.2 Nucleotide excision repair (NER).....	21
1.3.2.3 Mismatch Repair (MMR).....	24
1.3.3 Double-strand damage.....	27
1.3.3.1 Homologous recombination repair (HRR).....	27
1.3.3.2 Non-homologous end joining (NHEJ) .....	28
1.3.4 Translesion synthesis (TLS).....	30
<b>1.4 DISEASES SPECIFIC TO DNA REPAIR .....</b>	<b>33</b>



1.4.1 Xeroderma Pigmentosum (XP) .....	34
1.4.2 Cockayne syndrome (CS).....	34
1.4.3 Werner syndrome (WS) .....	34
1.4.4 Ataxia telangiectasia (A-T) .....	35
1.4.5 Hereditary nonpolyposis colorectal cancer (HNPCC) .....	35
1.4.6 Fanconi anemia (FA).....	36
<b>1.5 SYSTEMATIC PERSPECTIVE FOR DNA REPAIR MECHANISMS .....</b>	<b>36</b>
<b>REFERENCES .....</b>	<b>40</b>

## CHAPTER 2

### *Conception of DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems* 53-76

<b>ABSTRACT .....</b>	<b>55</b>
<b>2.1 INTRODUCTION.....</b>	<b>56</b>
<b>2.2 MATERIALS AND METHODS.....</b>	<b>58</b>
2.2.1 Database Design and Content .....	58
2.2.2 Collection of genotype data and quantitative genetic studies .....	60
2.2.3 Identification of nsSNPs and their functional effect on human repair system..	61
2.2.4 Detection of putative phosphorylation sites in DNA repair proteins .....	61
2.2.5 Computational classification of human DNA repair associated diseases and pathways.....	61
<b>2.3 RESULTS AND DISCUSSION .....</b>	<b>64</b>
<b>2.4 CONCLUSION.....</b>	<b>70</b>
<b>REFERENCES .....</b>	<b>71</b>

## CHAPTER 3

### *Systems biology approach for mutational and site-specific structural investigation of DNA repair genes for xeroderma pigmentosum* 77-102

<b>ABSTRACT .....</b>	<b>79</b>
<b>3.1 INTRODUCTION.....</b>	<b>80</b>

<b>3.2 MATERIALS AND METHODS.....</b>	<b>82</b>
3.2.1 Data Collection.....	82
3.2.2 Assessment of coding single nucleotide polymorphisms.....	82
3.2.3 Investigation of phenotypic impact of SNPs.....	83
3.2.4 Quantitative genetic association study and identification of phosphorylation sites as vital parameters implicated in XP.....	83
3.2.5 Detection of site-specific structural conservation for nsSNPs.....	84
3.2.6 Modeling nsSNPs on 3D protein structures and RMSD calculations.....	84
3.2.7 Effect of mutation on protein stability and secondary structures.....	85
3.2.8 Quantitative study to simulate XP gene regulatory pathway.....	86
<b>3.3 RESULTS AND DISCUSSION .....</b>	<b>86</b>
3.3.1 Identification and evaluation of nsSNPs and their phenotypic effects.....	86
3.3.2 Quantitative genetic analysis and putative phosphorylation sites.....	90
3.3.3 Inspection of nsSNP locations on the protein structures.....	91
3.3.4 Modeling of mutations onto the protein structures.....	91
3.3.5 Secondary structure and protein solvent accessibility.....	94
3.3.6 XP regulatory pathways simulation studies.....	94
<b>3.4 CONCLUSION.....</b>	<b>97</b>
<b>REFERENCES.....</b>	<b>97</b>

## CHAPTER 4

*An integrative approach for mapping differentially expressed genes and network components to elucidate key regulatory genes using novel parameters in colorectal cancer*      103-130

<b>ABSTRACT .....</b>	<b>105</b>
<b>4.1 INTRODUCTION.....</b>	<b>106</b>
<b>4.2 MATERIALS AND METHODS.....</b>	<b>108</b>
4.2.1 Biological data.....	109
4.2.2 Pre-processing of data.....	109
4.2.3 Identification of differentially expressed genes.....	109
4.2.4 Cluster analysis for co-expressed genes.....	110
4.2.5 Transcriptional regulation of CRC genes.....	110

4.2.6 Functional enrichment for differentially expressed genes .....	110
4.2.7 Detection of important regulatory patterns in CRC pathway.....	111
<b>4.3 RESULTS AND DISCUSSION .....</b>	<b>113</b>
<b>4.4 CONCLUSION.....</b>	<b>124</b>
<b>REFERENCES .....</b>	<b>124</b>

## CHAPTER 5

### *Decoding the intricate biological pathways in quest of candidate markers implicated in human DNA repair system* 131-158

<b>ABSTRACT .....</b>	<b>133</b>
<b>5.1 INTRODUCTION.....</b>	<b>134</b>
<b>5.2 MATERIALS AND METHODS.....</b>	<b>136</b>
5.2.1 Reconstruction and statistical estimation of DNA repair pathway .....	136
5.2.2 Biological pathways and detection of network motifs.....	138
5.2.3 Annotation of vital network components .....	138
5.2.4 Biological inference form available statistical constraints.....	139
5.2.5 Novel designed parameters for deducing crucial patterns .....	139
<b>5.3 RESULTS AND DISCUSSION .....</b>	<b>141</b>
<b>5.4 CONCLUSION.....</b>	<b>155</b>
<b>REFERENCES .....</b>	<b>156</b>

**CONCLUSION AND FUTURE DIRECTION** 159-165

**APPENDIX** 167-198

**Appendix A** 167-180

**Appendix B** 181-198

**LIST OF PUBLICATIONS** 199-204

## LIST OF TABLES

Tables	Title	Page No.
<b>Table 1.1</b>	DNA repair genes involved in direct reversal of damage mechanism.	18
<b>Table 1.2</b>	DNA repair proteins from base excision repair mechanism.	19-20
<b>Table 1.3</b>	A list of nucleotide excision repair genes and associated diseases.	23-24
<b>Table 1.4</b>	Mismatch repair genes and associated diseases.	26
<b>Table 1.5</b>	DNA repair genes implicated in homologous recombination repair.	29
<b>Table 1.6</b>	Genes drawn in non-homologous end joining repair mechanism.	30
<b>Table 1.7</b>	The DNA repair genes concerned with translesion synthesis mechanism.	32
<b>Table 1.8</b>	Specialized translesion polymerases in eukaryotes.	33
<b>Table 2.1</b>	No. of predicted nsSNPs that alter protein sequence and associated diseases with various DNA repair mechanisms.	62-63
<b>Table 2.2</b>	Identification and literature verification of phosphorylation sites in DNA repair proteins.	68
<b>Table 3.1</b>	Compiled list of XP genes, vital genetic markers and phenotypic consequences.	86
<b>Table 3.2</b>	The association of genetic mutations and phenotypic variations in XP genes.	87-88
<b>Table 3.3</b>	The risk association of various SNPs with their phenotypic effects.	89-90
<b>Table 3.4</b>	Impact of mutations on the protein solvent accessibility and its secondary structure.	94
<b>Table 4.1</b>	Major transcription factors identified in early colorectal cancer progression.	115
<b>Table 4.2</b>	Network motifs with their respective standard statistical parameters.	119
<b>Table 4.3</b>	Values of the designed parameters for each recurrent motif in the CRC pathway.	121
<b>Table 4.4</b>	Putative over-represented genes from CRC pathway as indicated by the most recurrent network motif.	122
<b>Table 5.1</b>	Network motifs from prostate cancer pathway with their respective standard statistical parameters.	147-150
<b>Table 5.2</b>	A list of DNA repair associated diseases subjected to pathway level analysis.	151
<b>Table 5.3</b>	Values of the designed parameters for each recurrent motif in prostate cancer pathway.	153
<b>Table 5.4</b>	List of identified key proteins in DNA repair diseases using newly designed parameters	154

## LIST OF FIGURES

Figures	Title	Page No.
<b>Figure 1.1</b>	Foundation for the field of genomics rests on The Human Genome Project. Human genome sequence and the identified genes provided valuable insights in understanding health and disease.	10
<b>Figure 1.2</b>	Schematic representation portraying complexity of a human cell where DNA serves as a crucial molecule consisting of 3.2 billion base pairs of A, T, G, C nucleotides. The human DNA comprise of approximately 19,000 genes encoding information for proteins which either independently or in the form of a complex carry out various cellular functions.	11
<b>Figure 1.3</b>	DNA damage, repair mechanisms and consequences. a. Common DNA damaging agents (top); examples of DNA lesions induced by these agents (middle); and most pertinent DNA repair mechanism responsible for the removal of lesions (bottom). b. Acute effects of DNA damage on cell-cycle progression, leading to transient arrest in G1, S, G2 and M phases (top), and on DNA metabolism (middle). Long-term consequences of DNA injury (bottom) include permanent changes in the DNA sequence and their biological effects. Abbreviations: <i>cis</i> -Pt and MMC, cisplatin and mitomycin C, respectively (both DNA cross-linking agents); (6-4)PP and CPD, 6-4 photoproduct and cyclobutane pyrimidine dimer, respectively.	15
<b>Figure 1.4</b>	The DNA repair system: various repair mechanisms for maintaining genome integrity.	16
<b>Figure 1.5</b>	The direct reversal DNA repair mechanism; a. Photoreactivation process for removal of thymine dimers. b. Direct removal of alkylated base by stoichiometric use of MGMT.	17
<b>Figure 1.6</b>	Base excision repair mechanism for dealing with single base changes.	21
<b>Figure 1.7</b>	Nucleotide excision repair mechanism for the removal of bulky and helix distorting lesions.	22
<b>Figure 1.8</b>	Mismatch repair system for the removal of mismatches in DNA.	25
<b>Figure 1.9</b>	Repair of double-strand breaks by homologous recombination repair mechanism.	28
<b>Figure 1.10</b>	The NHEJ mechanism repairing DSBs in the absence of homologous strand.	31
<b>Figure 1.11</b>	The bypassing of lesions via translesion synthesis mechanism.	32
<b>Figure 2.1</b>	DNA repair genes and associated mechanisms; the percentage (%) shows the number of genes present in each mechanism.	59
<b>Figure 2.2</b>	The pursued pipeline for DNA repair genetic association studies	59

	database.	
<b>Figure 2.3</b>	Genetic association details for 16 DNA repair mechanisms; the statistical distribution of number of genes in different mechanisms, important haplotype blocks and genetic markers associated with these genes.	64
<b>Figure 2.4</b>	The homepage for DNA Repair Genetic Association Studies (DR-GAS) database.	69
<b>Figure 2.5</b>	Demonstration and implementation of DR-GAS with various available search and advanced options. The illustrated output from the repository is represented in a combined image for all the generated results.	70
<b>Figure 3.1</b>	The methodology for mapping the mutations onto the protein structures.	85
<b>Figure 3.2</b>	(a) In DDB2 protein, ('A293T', 'K244E', 'D307Y') mutations were mapped on the structure of DDB2 (PDB ID: 4E54) and the site specific positions have been labeled where the three mutations have been observed at highly conserved sites of the protein due to bordeaux colour; (b) The POLH mutations ('M14V', 'G209V', 'K231N', 'C321W') were introduced in the structure (PDB ID: 3MR2) and been labeled where M14V and K231N mutations were found in the conserved region whereas C321W and G209V in less conserved as compared to earlier mutations due to light pink and very light pink colours; (c) In XPA protein, the C108F mutation was mapped to the protein structure (PDB ID: 1XPA) and was also found in the highly conserved region.	92
<b>Figure 3.3</b>	(a) For DDB2 protein, (I) is the native structure of DDB2 in cartoon representation and in (II-IV) the native structure (white) is superimposed with the mutant structure (orange) where (II) symbolizes the A293T mutation (III) corresponds to D307Y mutation (IV) represents K244E mutation where the native residues are highlighted in green colour and the mutants in red; (b) Likewise, in POLH, (I) symbolizes C321W mutation (II) corresponds to G209V mutation (III) represents K231N mutation (IV) showing M14V mutation where the native residues are highlighted in green colour and the mutants in red; (c) For XPA protein, (I) characterizes the native structure of XPA and (II) represents the native structure (white) superimposed to the mutant structure (orange) with C108F mutation in cartoon representation where the native residues are highlighted in green colour and the mutants in red.	93
<b>Figure 3.4</b>	A reconstructed pathway model for delineation of regulatory processes implicated in XP.	95
<b>Figure 3.5</b>	(a) A simulation performed through genes with respect to time and concentration where the genes corresponds to different coloured lines	96

	in the graph and are illustrated on right side of the image. <b>(b)</b> Includes the simulation studies executed with the DNA-protein complexes generated in the repair process.	
<b>Figure 4.1</b>	The workflow deliberated for recognizing candidate markers in colorectal cancer.	108
<b>Figure 4.2</b>	Pre-processing and normalization of DNA microarray data; <b>a.</b> shows the distribution of the microarray files before normalization; <b>b.</b> explains uniform distribution obtained after implementing normalization for removing noise from the data.	113
<b>Figure 4.3</b>	Identification of differential expression via significance analysis of microarray and volcano plot.	114
<b>Figure 4.4</b>	Functional enrichment and annotation analyses for differentially expressed genes.	116
<b>Figure 4.5</b>	Identified vital patterns (network motifs) from colorectal cancer pathway.	117
<b>Figure 4.6</b>	Bottom-up approach for classifying the network motifs.	118
<b>Figure 4.7</b>	Significance profile for all 4-8 node generated sub-graphs based on normalized z-scores; the motif significance profile evidently exemplifies that when the complexity in CRC pathway increases, the interactions among the nodes and intricacy in recognition of genes amplifies immensely. Lesser the node size, it becomes easy to annotate the nodes (genes) and their associations with stronger statistical significance (greater normalized z-scores).	120
<b>Figure 5.1</b>	A flow diagram for the applied methodology to detect key players from diseases involved in DNA repair.	140
<b>Figure 5.2</b>	The one to one, one to many and many to many interacting partners attained from clustering; revealed hub nodes as well as important connections.	142
<b>Figure 5.3</b>	The statistical inferences from DNA repair pathway comprising of shortest path length and neighborhood connectivity.	143
<b>Figure 5.4</b>	The Indegree distribution and Closeness centrality as observed in DNA repair pathway.	144
<b>Figure 5.5</b>	The average clustering coefficient, betweenness centrality and stress centrality generated from DNA repair proteins and elucidated pathway.	145-146
<b>Figure 5.6</b>	A few classified network motifs from the prostate cancer pathway.	144

---

**ABBREVIATIONS**

AP site	Apurinic site
AKT3	v-akt murine thymoma viral oncogene homolog 3
ARAF	v-raf murine sarcoma 3611 viral oncogene homolog
ASA	Accessible Surface Area
A-T	Ataxia telangiectasia
ATM	Ataxia Telangiectasia Mutated
BER	Base Excision Repair
BLM	Bloom syndrome
BRCA2	Breast Cancer 2
CIN	Chromosome Instability
CPDs	Cyclobutane Pyrimidine Dimers
CRC	Colorectal Cancer
CS	Cockayne Syndrome
CSB	Cockayne Syndrome complementation group type B
dbSNP	Single Nucleotide Polymorphism Database
DOR	Dense Overlapping Regulons
DiRE	Distant Regulatory Elements
DNA	Deoxyribonucleic acid
DNA-PKcs	DNA-dependent protein kinase catalytic subunit
DR1	Down-Regulator of transcription 1
DRD	Direct Reversal Of Damage
DSB	Double-Stranded Break
DSBR	Double-Strand Break Repair
dsDNA	double strand DNA
DSSP	Define Secondary Structure of Proteins
ELSI	Ethical, Legal, and Social Implications
ENCODE	Encyclopedia of DNA Elements
EXO1	Exonuclease 1
FA	Fanconi Anemia
FAP	Familial Adenomatous Polyposis
FastSNP	Function Analysis and Selection Tool for SNPs



---

FFL	Feedforward Loops
FOS	FBJ murine osteosarcoma viral oncogene homolog
GAD	Genetic Association Database
GEO	Gene Expression Omnibus
GG-NER	Global Genomic NER
HGD	Hyper-Geometric Distribution
HGP	Human Genome Project
HGVBBase	Human Genome Variation Database
HNF4	Hepatocyte Nuclear Factor 4
HNPCC	Hereditary Nonpolyposis Colorectal Cancer
HRR	Homologous Recombination Repair
HVP	Human Variome Project
H-W	Hardy-Weinberg
ICLs	Interstrand Crosslinks
JUN	jun proto-oncogene
KEGG	Kyoto Encyclopedia of Genes and Genomes
KRAS	kirsten rat sarcoma viral oncogene homolog
LD	Linkage Disequilibrium
LIG1	Ligase I
LOD	Likelihood Odds Ratio
MGMT	Methyl Guanine Methyl Transferase
mHG	minimum Hypergeometric
MIM	Multiple Input Module
MLH1	mutL Homolog 1
MMR	Mismatch Repair
MRN complex	Mre11-Rad50-Nbs1 complex
MSH2	mutS Homolog 2
MSI	Microsatellite Instability
NCBI	National Center for Biotechnology Information
NER	Nucleotide Excision Repair
NHEJ	Non-Homologous End Joining
NR2F1	Nuclear Receptor subfamily 2 group F member 1
nsSNPs	Non-Synonymous Single Nucleotide Polymorphisms

---

O6-MeG	O6-methylguanine
OMIM	Online Mendelian Inheritance In Man
PCA	Principal Component Analysis
PCNA	Proliferating Cell Nuclear Antigen
PDB	Protein Data Bank
PIK3R5	Phosphoinositide-3-kinase, regulatory subunit 5
Pol $\beta$	DNA polymerase $\beta$
pol $\lambda$	DNA polymerase $\lambda$
PolyPhen	Polymorphism Phenotyping
PPIs	Protein-Protein Interactions
PSIC	Position-Specific Independent Counts
QQ plot	Quantile-Quantile plot
RALGDS	Ral guanine nucleotide dissociation stimulator
RAND-ESU	Randomized Enumeration of Sub-graphs
REs	Regulatory Elements
RMA	Robust Multi Average Analysis
RMSD	Root Mean Square Deviation
RNS	Reacting Nitrogen Species
ROS	Reacting Oxygen Species
RPA1	Replication Protein A1
RSA	Relative Solvent Accessibility
S	Serine
SAM	Significance Analysis Of Microarrays
SAVES	Structural Analysis and Verification Server
SDSA	Synthesis-Dependent Strand Annealing
SIFT	Sorting Intolerant From Tolerant
SIM	Single Input Module
SLR	Signal Log Ratio
SNP	Single Nucleotide Polymorphism
SP	Significance Profile
SSBs	Single-Strand Breaks
T	Threonine
TC-NER	Transcription Coupled NER

TFIIH	Transcription Factor II Human
TFs	Transcription Factors
TLS	Translesion Synthesis
UV	Ultraviolet
VMD	Visual Molecular Dynamics
WebGestalt	WEB-based Gene SeT AnaLysis Toolkit
WS	Werner syndrome
XP	Xeroderma Pigmentosum
XPC	Xeroderma Pigmentosum, complementation group C
XRCC4	X-ray repair complementing defective repair in Chinese hamster cells 4
Y	Tyrosine
A	Alpha
B	Beta
Γ	Gamma
Z	Zeta
H	Eta
Θ	Theta
I	Iota
K	Kappa
Λ	Lambda
M	Mu
N	Nu

*“Imagination is more important than knowledge. For knowledge is limited to all we now know and understand, while imagination embraces the entire world, and all there ever will be to know and understand.” -Albert Einstein*

# ABSTRACT

**ABSTRACT**

In present era, the biomedical research has been revolutionized on the advent of human genome sequence and huge amount of genomic information is generated. This huge amalgamation of genomic data is rising exponentially due to employment of high-throughput techniques and application of modern biology. However, less emphasis is remunerated on the obligatory genomic analysis; subsequently there is a need to apply extensive computational approaches that could deal abreast with the remarkable growth in generated biological data. Since, the human system is very complex and the ultimate challenge is to decipher numerous multifaceted processes involved in the regulation of this system therefore systems biology has major relevance which assists in wide-ranging understanding of an entire system as a whole. The human genome is susceptible to numerous endogenous and exogenous damages resulting in genomic instability. These DNA lesions not only diversely impact genome but also enhance the predisposition to multi-system defects including premature aging and oncogenesis. On exposure to these damages, major cellular signaling processes are triggered which either result in cell-cycle arrest, damage removal and activation of repair pathways. There are a variety of DNA repair mechanisms which confiscates diverse form of lesions whether affecting single-stranded or double-stranded DNA. Genetic abnormalities or mutations in DNA repair genes have implications in multiple hereditary disorders, aging and cancers which have always been a crucial subject for analysis among researchers. Although some of the disease related mutations have been identified but there is a huge lack of knowledge underlying the associated mechanism. Hence, a comprehensive systems biology approach, integrating all biological components, genes/proteins, enzymes, genetic variations, complex networks is necessary for systematic understanding of the biological processes.

An extensive vision to thoroughly understand DNA repair associated diseases is complex and requires high-end computations and resources. Therefore, there is a prerequisite for a platform to store huge amalgamation of DNA repair data and tools for its analysis to comprehensively understand human diseases and thus bioinformatics grants this opportunity. Although DNA repair is a crucial process in human molecular systems but till date there is no such catalog available which provides appropriate classification of DNA repair genes in associated mechanisms. In order to accomplish the goal, we developed DR-GAS (DNA Repair Genetic Association Studies), a unique, consolidated and comprehensive DNA repair genetic association studies database of human DNA repair system. It presents information on 215 DNA repair genes, assorted mechanisms of repair, genetic markers, phosphorylation

sites, associated diseases and pathways drawn in human repair system. This database is integrated with a web interface and developed for DNA repair genes/proteins and is freely accessible for academic and research purposes at: <http://www.bioinfoindia.org/drgas>. Till date, there is no such catalog for DNA repair system available which provides appropriate classification of DNA repair genes in associated mechanisms as provided by DR-GAS (i.e. 16 major classes) along with genetic parameters and phosphorylation states in genomic and proteomic data. It is anticipated that this web-based resource would serve as a valuable accompaniment for in-depth analysis and future studies on human DNA repair systems and will also contribute scientific knowledge towards better understanding of other mammalian repair systems as well.

As discussed earlier, there are innumerable human diseases reported due to aberrations in repair processes and there is deficient available knowledge underlying the disease mechanisms. Consequently, to identify candidate markers for certain human diseases (xeroderma pigmentosum (XP) and colorectal cancer (CRC) implicated in DNA repair system; an integrative bioinformatics approach has been applied. A system level understanding for XP, a rare genetic skin disorder mainly caused due to extreme sensitivity towards ultraviolet (UV) radiations was executed. On exposure to UV radiations, DNA acquires damages and the repair process is not triggered leading to skin and often neurological abnormalities. Often mutations in nucleotide excision repair (NER) and translesion synthesis (TLS) pathway form the basis of XP. The analyses of such mutations in DNA repair genes are vital for understanding XP and the involved cancer genetics to facilitate identification of candidate markers and designing of anticancer therapeutics. We identified 8 XP associated DNA repair genes namely, DDB2, ERCC2, ERCC3, ERCC4, ERCC5, POLH, XPA and XPC. Further, we perceived 55 deleterious non-synonymous single nucleotide polymorphisms (nsSNPs) and examined them at structure-level (8 nsSNPs in DDB2, POLH and XPA) by altering structure, estimating secondary structure, solvent accessibility and performing site specific conservation analysis. There was no available report on the XP disease pathway although interactions between different entities in NER were known thus we designed a putative XP associated model and carried out several simulations. The perceptive of the effect of mutations at protein structure level and its diverse phenotypes will help identifying the reliance between genetic mutations and phenotypic variations in XP and once validated these mutations could serve as biomarker for the disease.

---

The next broadly studied DNA repair associated disease is colorectal cancer (CRC) for which an integrated *in silico* analyses was accomplished. Worldwide, CRC influences millions of people and exists as the most commonly diagnosed cancers along with lung and breast cancer. CRC is a polygenic disease caused due to different genetic factors and often due to aberrations in DNA repair genes. Although genes involved in mismatch repair (MMR) system like MLH1, MSH2, MSH6, PMS2 and other genes such as APC and MUTYH have already shown their influence on CRC but still the cause and progression of disease remains unrequited. Analyzing complex biological pathway of CRC is a convoluted process and requires a holistic approach for identifying biomarkers and chief network components. Thus, the several analyzed biological components for CRC in our study comprise of differentially expressed genes, their annotations, network patterns (motifs) that are functionally noteworthy. Understanding entire disease pathway is a drawn-out process hence, we identified crucial network components from the biological pathway. Various statistical parameters such as  $z$ -scores and significance profile (SP) were utilized for identifying key players from the network components. Also, we developed some novel parameters based on frequencies for estimating the effectiveness of each marker contributing towards disease. From the standard available and designed parameters, 5 key genes are reported i.e. KRAS, ARAF, PIK3R5, RALGDS and AKT3 that can be studied at length for its disease association. Further, investigating and targeting these proposed genes for experimental validations, instead being spellbound by the complicated pathway will certainly endow valuable insight in a well-timed systematic understanding of the disease.

Keeping in view the requisite for a systems biology approach for DNA repair related disorders and pathways; we proposed our final objective to fill the research gap. The biological network complexity is growing enormously and in order to reveal specific confined properties of these intricate networks, the breakdown of network into distinct clusters and components would be extremely valuable. The diverse networks comprise of different set of local structures as patterns which are referred to as network motifs that are functionally significant. Identification and annotation of these biological entities in DNA repair pathways may help in analyzing vital interactions involved in diseases associated to DNA repair and may further aid in resolving critical functions performed by these networks. For the exhaustive analysis, we switched from the systems level to component level understanding by annotating all components and their associated interactions in a biological

system and finally proposed a few major players implicated in DNA repair associated diseases.

The applied computational approach holds great potential in revealing candidate markers for diseases underlying DNA repair associated aberrations. Our study presented a few crucial mutations associated to XP and also provided a few key genes that may be analyzed further for their role in CRC instead being enthralled by intact pathway. The novel parameters designed for the pathway level analysis to identify vital network motifs in CRC and other DNA repair related diseases ultimately provides the major candidates contributing towards diseases. The established resource (DR-GAS) is a comprehensive compendium available for the scientific community that endows with vital information on DNA repair genes, associated diseases, pathways and interactions. These *in silico* techniques are not only time efficient but also cost effective and provides essential clues to experimental biologists and scientific community to design appropriate experiments and validate results.



*“It is possible to fall in many ways, while to succeed is possible only in one way.”*  
*-Aristotle*

# CHAPTER-1

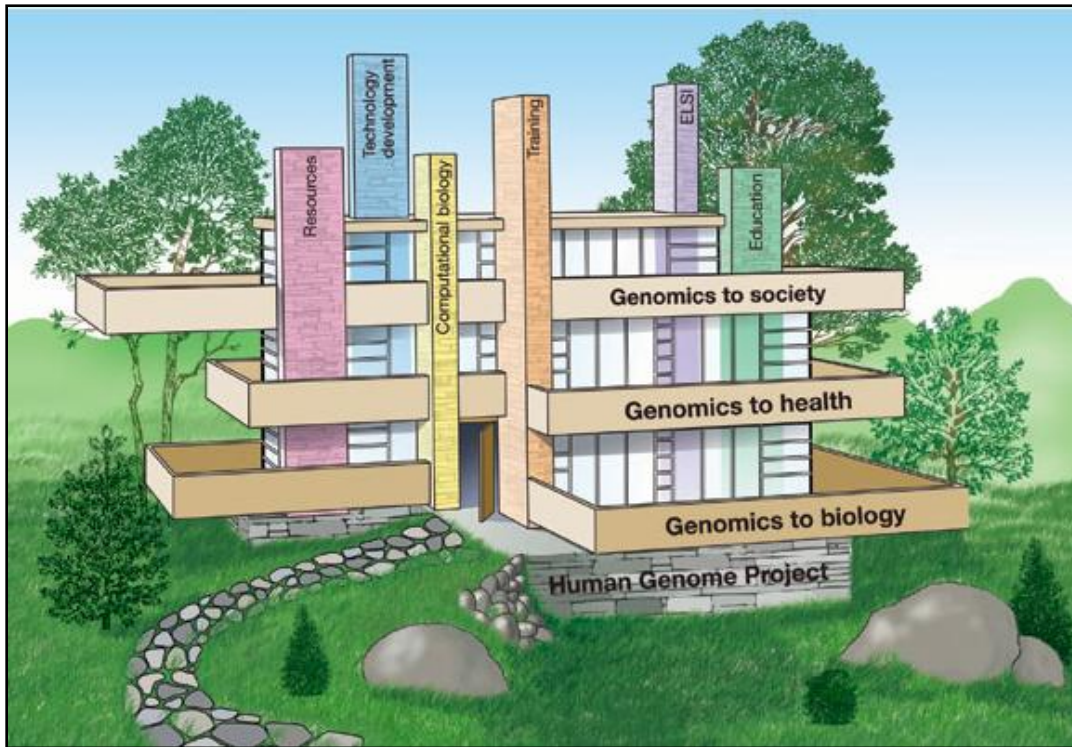
# INTRODUCTION

## 1.1 INTRODUCTION

A great revolution and landmark accomplishments were attained in the field of biomedical research in 2003 during the era of The Human Genome Project (HGP): when the entire human genome was sequenced and the genes were mapped under high-end collaborative efforts of many countries [1-3]. For the comprehensive understanding of human genome, this collective endeavor dealt with major goals such as providing an accurate sequence of 3.2 billion DNA base pairs and identification of the estimated 20,000-25,000 human genes [2]. The project also intended for establishing new tools and innovative technologies for analyzing the data and making it accessible worldwide. Moreover, it invoked Ethical, Legal, and Social Implications (ELSI) program for dealing with bioethics concerning the individual and society issues. Ultimately, the project has proved a promising platform offering information on structure, function and organization of complete set of human genes. Till date, The HGP is considered as a commendable joint initiative by the scientific community, pooling its skills and resources to achieve a common goal which served as a perfect example stating a quote by a Japanese poet, Ryunosuke Satoro, *“Individually, we are one drop. Together, we are an ocean”*. After this crucial epoch, the complete human genome sequence was ultimately accessible [2, 3] after 50 years of attaining the structure of DNA (by James Watson and Francis Crick in 1953) [4].

The HGP endows an extensive vision as depicted in **Figure 1.1** for the outlook of genomics research that lays a strong basis for health and disease. Since, the human genome sequence was now deciphered, the predisposition of a person to particular disease was easier to spot and hence steering a new direction for human genomics and comparative studies. The elucidation of multifaceted processes such as gene expression, effect of variations on phenotypic consequences, protein-protein interactions and other global analysis of human biology also became quite effortless. The achieved milestone via HGP facilitated the development of other vital projects and globally coordinated ventures like Human Variome Project (HVP) [5], Encyclopedia of DNA Elements (ENCODE) Project [6] and International HapMap Project [7]. These collaborative projects were involved in collection and curation of genetic variations impacting human health, identification of functional elements and genes associated to different diseases and their response to pharmaceuticals. On successful completion of these versatile international projects, the major challenge was the worthy analysis from this huge accumulation

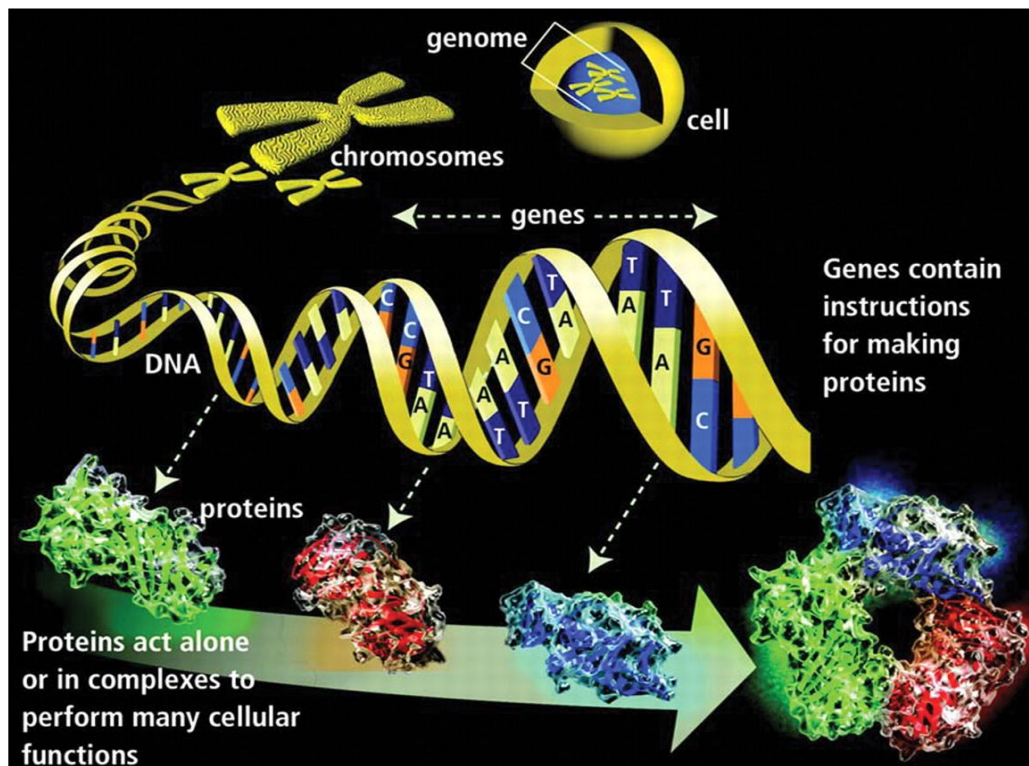
of generated data. This explosion of information needs computational approaches for storing data and tools for its analysis to comprehensively gain insights latent in human diseases [8].



**Figure 1.1** Foundation for the field of genomics rests on The Human Genome Project. Human genome sequence and the identified genes provided valuable insights in understanding health and disease [9].

The success of genomic projects has generated a vast amount of sequence data, and accelerated the pace of functional analysis of genes. Deoxyribonucleic acid (DNA) is the blueprint of life and a fundamental unit which determines the cellular behavior and functions in an organism (**Figure 1.2**). Subsequently, understanding this intricate molecule demonstrating genetic complexity and enclosing biological code for all genes and its obscure interactions, would certainly endow in-depth insight in resulting human diseases. The DNA is believed to be the basic hereditary material which when gets exposed to mutagens can acquire damages depending on the type of diverse mutagens. These damages if not repaired may cause genomic instability and are major factors governing the aging process where the accumulation of these damages take place [10, 11]. With this explosive growth of DNA data, major DNA transactions including damage recognition and removal, transcription, replication, and recombination thoroughly entwined DNA repair with rare human genetic disorders. In reciprocation, the study

of human diseases has been instrumental in the development of our understanding of human repair processes. The process of DNA damage, excision and repair was discovered six decades ago from studies on DNA processing in ultraviolet (UV)-irradiated bacteria [12]. Now it is evident that the ability to recognize and repair abnormal forms of DNA is common and necessary to all forms of life. It has also become apparent that DNA damage and repair processes are vital to understand the mechanisms of cancer, ageing and human genetic diseases [13].



**Figure 1.2** Schematic representation portraying complexity of a human cell where DNA serves as a crucial molecule consisting of 3.2 billion base pairs of A, T, G, C nucleotides. The human DNA comprise of approximately 19,000 genes [14] encoding information for proteins which either independently or in the form of a complex carry out various cellular functions. [Retrieved from: U.S. Department of Energy Office of Science].

## 1.2 DNA DAMAGE: GENOMIC INSTABILITY

The human genome is susceptible to numerous endogenous and exogenous damages resulting in genomic instability [15]. DNA damage occurs at the rate of 10,000 to 1 million molecular lesions per cell each day [10]. These damages are majorly responsible for altering the primary structure of the double helix by chemically modifying bases and often introduce unusual chemical bonds or bulky adducts that are not easily sustained in the standard double helix. Once, the cell

determines these diverse form of damages, a coordination of cellular responses comprising of transcriptional activation, cell cycle control, apoptosis, senescence and DNA repair processes are triggered [16]. These DNA lesions not only impact the genome but if not repaired or aberrantly repaired enhances the likelihood of developing diseases such as multiple form of cancers, neurological abnormalities, immunodeficiency and premature aging [17-19]. There are basically two kinds of classification for the damages; one depends on the type of factors or agents causing the DNA lesions and the other on the consequences of damages; discussed in the following sections.

### **1.2.1 Damage classification based on factors causing the damage**

There are numerous intrinsic as well external agents that contribute to the impairment in DNA where intrinsic factors include aberrations in metabolic and other cellular processes and external factors comprises of the environmental agents such as ionizing radiations and genotoxic compounds. Both the factors contributing towards damages which are mutagenic and toxic to the cell are explained in detail below:

#### **1.2.1.1 Endogenous Damage**

These damages are often introduced in DNA due to replication errors and other aberrant cellular metabolic processes like DNA replication, recombination and repair [10]. It also includes damages resulting from reacting oxygen and nitrogen species (ROS and RNS respectively) such as superoxide anions, hydrogen peroxide and hydroxyl radicals engendered as the byproducts from normal metabolic processes like lipid peroxidation and oxidative respiration [20]. The DNA suffering from endogenous damages consequently result in bulky adduct formation, hydrolysis (deamination, depurination, and depyrimidination of bases), oxidation (generation of 8-oxo-7,8-dihydroguanine and DNA strand interruptions), mismatches (resulting due to errors in replication) and alkylation (frequently due to 7-methylguanine, 1-methyladenine, 6-O-methylguanine) of bases [21, 22]. All these chemical modifications interfere with the normal cellular processes and destabilize the integrity of genome.

### 1.2.1.2 Exogenous Damage

There is a huge diversity in environmental agents causing damage to DNA such as UV radiations, X-rays, Gamma rays, plant toxins, thermal disruption, viruses, certain aromatic and genotoxic compounds [23-25]. All these external factors are alleged to alter the structure of DNA and produce aberrations in DNA such as formation of free radicals, cyclobutane pyrimidine dimers (CPDs) and pyrimidine–pyrimidone-(6-4)-photoproducts [23, 26]. Additionally, due to exogenous damage, the formed DNA adducts include oxidized bases, alkylated phosphotriesters and cross-linked DNA. Depurination and single-strand breaks (SSBs) in DNA are also perceived at the elevated temperatures. Since, there are a variety of DNA damaging agents, also the outcome of these damages are extremely versatile leading to oxidation, alkylation and hydrolysis of bases including deamination, depurination and depyrimidination. Other prominent DNA damages include mismatches, bulky adduct formation, pyrimidine dimers and cross-linking of DNA. Further, the formation of these adducts trigger processes such as cell-cycle arrest or cell death, transcriptional program activation, apoptosis and DNA repair. Thus, to counteract the deleterious effect of these damages and to maintain the integrity of genome, different mechanisms exist for repairing DNA in a precise manner [27]. A few lesions often escape repair process or remain unrepaired, leading to irreversible mutations in DNA that further alters the cellular phenotype and enhances the risk of oncogenesis [28] and other associated diseases [29].

### 1.2.2 Damage classification based on the consequences of damages

There are mainly four types of damages based on the effects of aberrations in DNA. The damages may either be at nucleotide level (affecting the bases) or may impact the phosphodiester backbone of DNA by inserting breaks or cross-linking of DNA. These damages include:

- i. *Covalent modification* at several positions in all the four bases in DNA (A, T, C, G). For instance, deamination of bases [30] where a loss of amino group from the cytosine leads to its conversion to a uracil residue [31].
- ii. *Mismatch* of bases due to aberrant proof reading during DNA replication [32]. For example: the most common mismatch found in DNA is for uracil instead of thymine as the former is normally present in RNA.

iii. *Insertion of breaks* in the backbone of DNA is frequently observed due to UV irradiations and certain chemicals. These breaks can either be limited to one of the two strands, i.e. SSB or can impact both the strands of DNA resulting in a double-stranded break (DSB) [33].

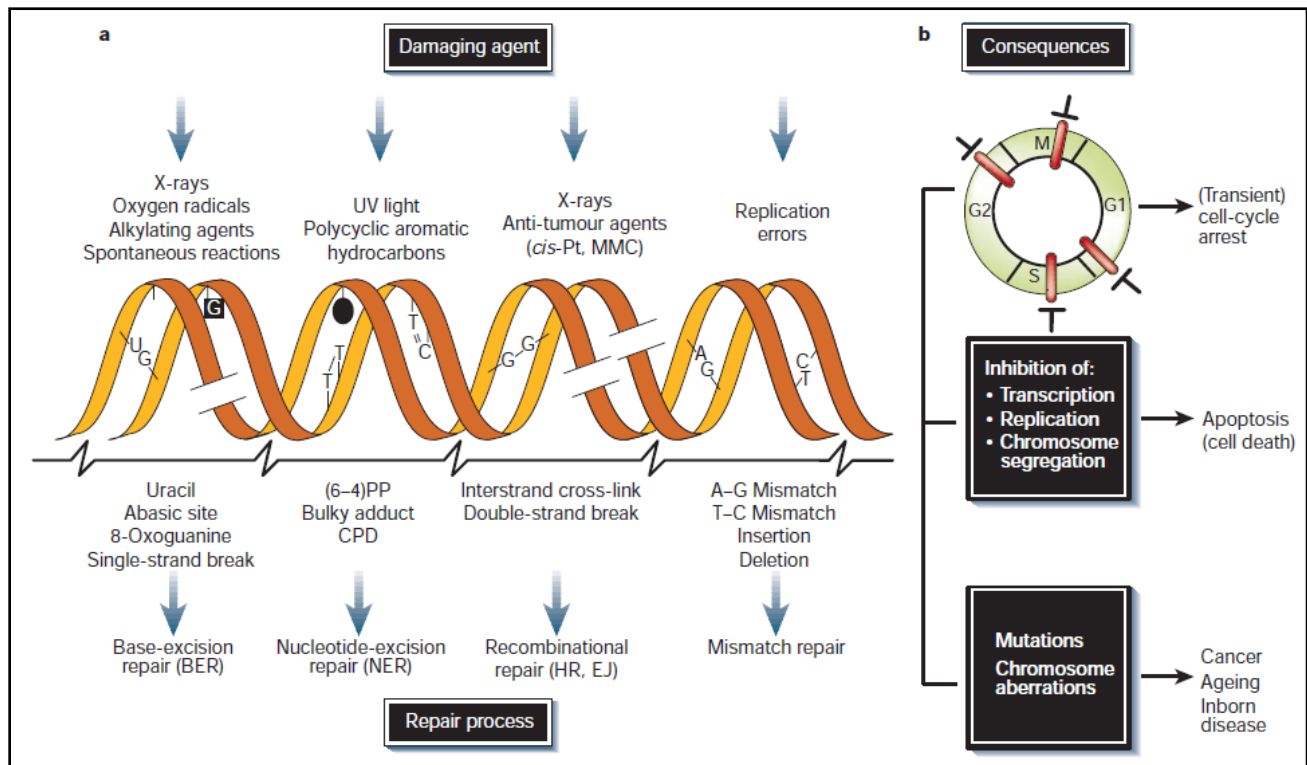
iv. *Cross-linking* of DNA due to several chemotherapeutic drugs used against cancers can cause damage to DNA by the covalent linkages between normal bases. These linkages can either be on the same strand of DNA i.e. intrastrand cross-link or on the opposite strand known as interstrand cross-link [34, 35].

### 1.3 DNA REPAIR SYSTEM

All the organisms ranging from prokaryotes to eukaryotes are equipped with DNA repair processes that deals with diverse form of lesions and prevent the genome from permanent mutations [36]. Each day approximately  $10^{16}$ – $10^{18}$  DNA repair events take place in a healthy adult containing  $10^{14}$  cells [37]. The process of DNA repair involves intricate interactions transpiring in a highly systematic behavior where the major steps comprise of recognizing the damage and initiating signaling process, recruitment of repair proteins, processing of lesions, resynthesis of double strand DNA (dsDNA) and finally ligation is achieved. A single DNA repair mechanism can never handle the plethora of lesions [33], therefore there exists several DNA repair mechanisms that individually tackle the specialized damages [17, 27]. DNA repair pathways are broadly classified into 4 major classes based on their ability to recognize and remove different damages namely, direct reversal of damage (DRD), single-strand damage, double-strand damage and translesion synthesis (TLS). Each of these mechanisms incorporates a wide range of proteins, enzymes and follows different approaches for repairing the damaged DNA. Thus, DNA repair is an intricate process which confiscates all these diverse form of lesions via different repair mechanisms and maintains the genome integrity.

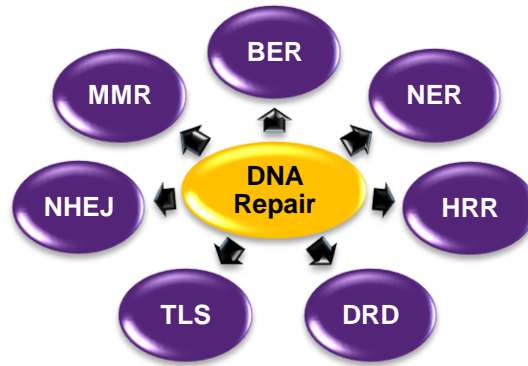
The ability to recognize and repair damaged DNA is common to all organisms and numerous DNA repair pathways have evolved in order to detect and repair almost all possible DNA lesions. **Figure 1.3** evidently portrays the different damaging agents, resulting aberrations, its consequences and the triggered repair processes in humans. The well-characterized DNA repair mechanisms include DRD, base excision repair (BER), nucleotide excision repair (NER),

mismatch repair (MMR), homologous recombination repair (HRR), non-homologous end joining (NHEJ) and TLS (**Figure 1.4**); these incorporates varied set of genes, enzymes and pathways for repairing the DNA [38]. Many human diseases are reported due to aberrations in these repair processes and there is a huge lack of acquaintance underlying the mechanisms. These DNA repair mechanisms are not only indispensable for eradicating damages but also prevent the genome from carcinogenesis and pre-mature aging. All the DNA repair mechanisms and associated diseases are explained in detail later in the chapter.



**Figure 1.3** DNA damage, repair mechanisms and consequences. **a.** Common DNA damaging agents (top); examples of DNA lesions induced by these agents (middle); and most pertinent DNA repair mechanism responsible for the removal of lesions (bottom). **b.** Acute effects of DNA damage on cell-cycle progression, leading to transient arrest in G1, S, G2 and M phases (top), and on DNA metabolism (middle). Long-term consequences of DNA injury (bottom) include permanent changes in the DNA sequence and their biological effects. Abbreviations: *cis*-Pt and MMC, cisplatin and mitomycin C, respectively (both DNA cross-linking agents); (6–4)PP and CPD, 6–4 photoproduct and cyclobutane pyrimidine dimer, respectively [38].





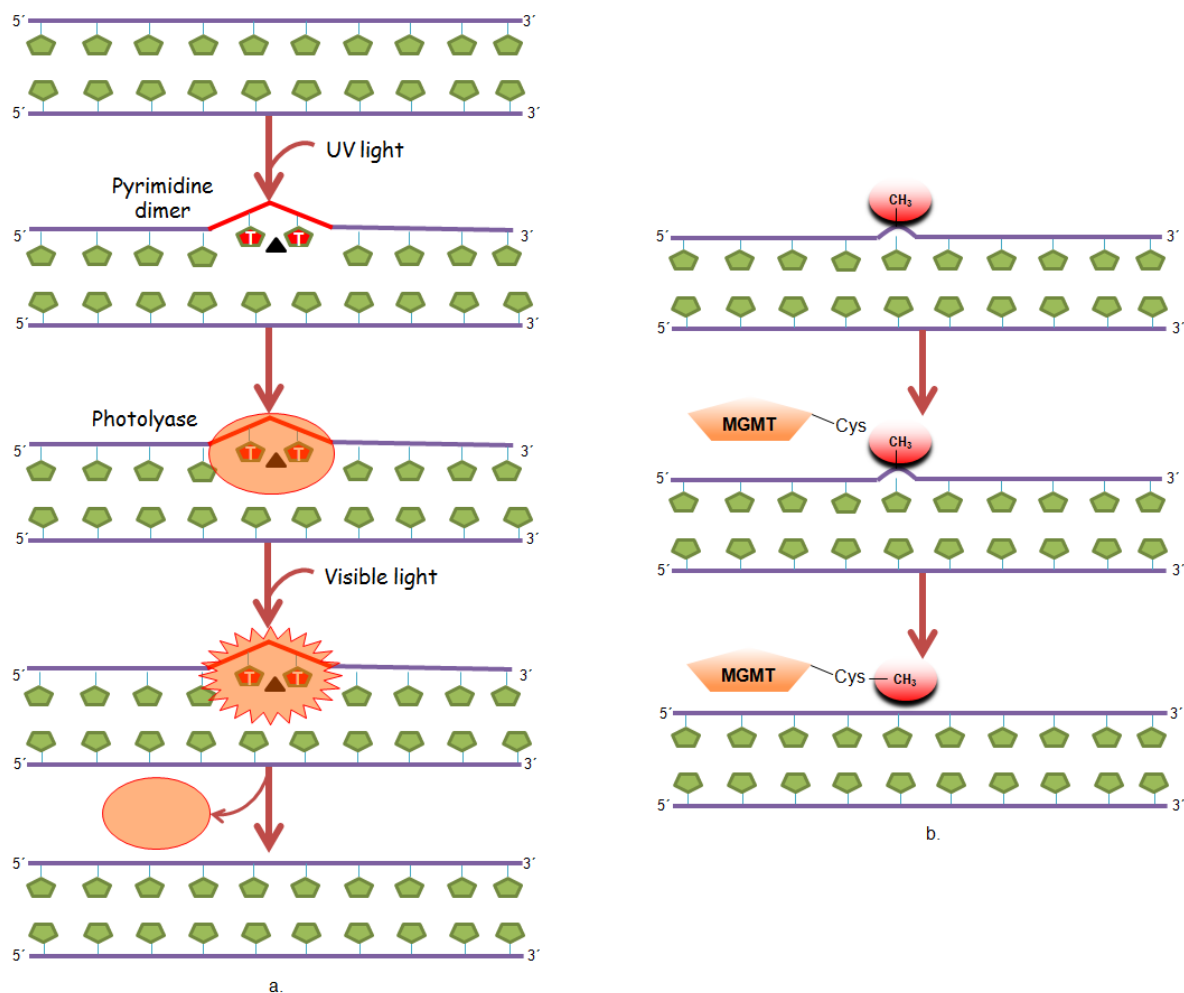
**Figure 1.4** The DNA repair system: various repair mechanisms for maintaining genome integrity.

### 1.3.1 Direct reversal of damage (DRD)

The DRD mechanism mainly focuses on lesions occurring in only one of the four bases and does not rupture the phosphodiester backbone of DNA for repairing damages. The mechanism doesn't require a DNA template to chemically reverse the damage. A wide range of damages such as CPDs [39] and alkylation of bases [39, 40] are directly removed via this mechanism. The CPDs resulting as a consequence of an abnormal covalent linkage between adjacent pyrimidine bases formed due to UV light irradiation are repaired directly by photoreactivation i.e. reversing the damage using photolyase enzyme. This photoreactivation reaction was the first DNA repair process to be discovered in the bacteriophage in 1949 [12, 41]. Cells are also capable of reversing the methylation of guanine residue, i.e. O6-methylguanine (O6-MeG) caused due to either environmental alkylating agents [42] or endogenously by S-adenosylmethionine which acts as a methyl donor in many cellular reactions [43]. The spontaneous addition of methyl group is one of the major factors responsible for point mutations in humans [44]. Most of the lesions formed by alkylation of DNA are repaired by O-6-methylguanine-DNA methyltransferase encoded by methyl guanine methyl transferase (MGMT) gene. This enzyme repairs the damage by stoichiometrically transferring the alkyl group from O-6 position to the cysteine residue of enzyme. The carried process is an expensive stoichiometric reaction since the enzyme is irreversibly inactivated.

DRD mechanism (**Figure 1.5**) depicts the removal of pyrimidine dimers and alkylated base from DNA. Due to UV exposure, the dimers formed in DNA by covalent binding of two consecutive thymine bases instead of normal base pairing are removed by photoreactivation

process in which the DNA photolyases recognize the bends in DNA. On excitation with blue light ( $>300$  nm), photolyases change the conformation and break the dimer apart; making the DNA free from damage as depicted in **(Figure 1.5a)**. The other damage reversal is the removal of methyl group from guanine residue via MGMT. In this process **(Figure 1.5b)**, the methyl group from guanine of DNA is transferred to the cytosine of MGMT molecule whereby the DNA restores its native form. One MGMT molecule can only remove one methyl group from DNA as mentioned earlier. **Table 1.1** represents the three important DNA repair proteins implicated in DRD mechanism and also presents the diseases resulting due to aberrations in these proteins. The literature reference corresponds to Pubmed IDs for the diseases mentioned in sequential order.



**Figure 1.5** The direct reversal DNA repair mechanism; a. Photoreactivation process for removal of thymine dimers. b. Direct removal of alkylated base by stoichiometric use of MGMT.

**Table 1.1 DNA repair genes involved in direct reversal of damage mechanism**

Genes	Full Name	Diseases	Literature References
ALKBH2	alkB, alkylation repair homolog 2	Lung cancer	21278781
ALKBH3	alkB, alkylation repair homolog 3	Prostate cancer	18077911
MGMT	O-6-methylguanine-DNA methyltransferase	Anaplastic astrocytoma, Brain tumors, NSCLC, Colorectal carcinoma, Lung cancer, Colon cancer, Gastric cancer, Gallbladder carcinoma, XP, HNSCC, Lymphoma, Hepatocellular carcinoma, Ovarian carcinoma, Endometrial cancer, Cervical cancer, Pancreatic tumor	9779706, 1518162, 12538473, 10811111, 17164358, 15800322, 19089477, 16309194, 20128036, 15184253, 16047061, 12619064, 10404091, 18973931, 15099958, 14506174

### 1.3.2 Single-strand damage

Damages occurring in any one of the DNA strands are often repaired by single-strand damage pathways where the repair process is guided by undamaged strand to remove and replace the lesion in damaged strand. In this process, the damaged nucleotide is excised from the damaged DNA strand and replaced by the correct residue complementary to undamaged strand thus giving rise to the excision repair process. The excision process is mainly divided into three categories:

#### 1.3.2.1 Base excision repair (BER)

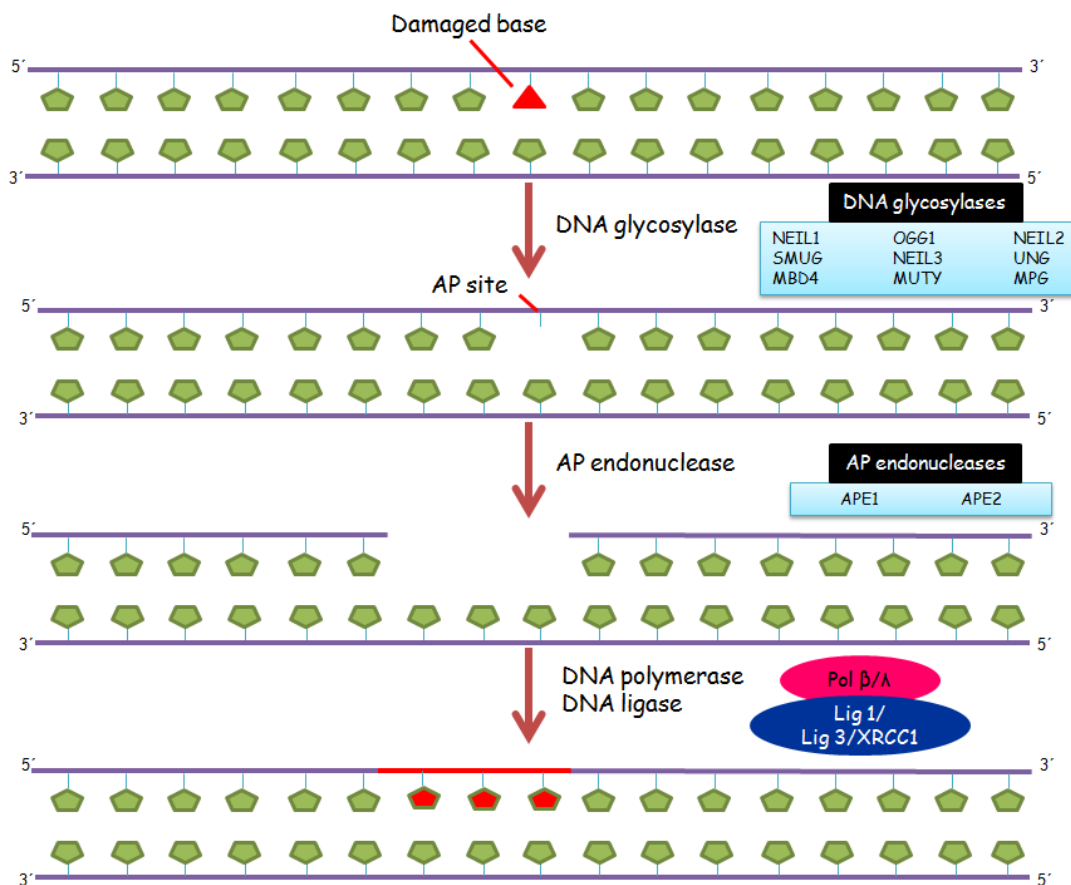
BER mechanism deals with single base changes introduced in genome due to factors such as oxidation, alkylation and hydrolysis of bases [45]. A few key enzymes implicated in this repair mechanism include DNA glycosylase, which identifies and removes the damage from DNA by cleaving N-glycosyl bond between base and the sugar thus creating a missing site in DNA. This abasic site is then detected by apurinic (AP) endonuclease, that further cuts the phosphodiester bond and cleaves the damage leaving a sugar attached to the 5' side of nick. The consequential 3' hydroxyl acts as a substrate for the repair polymerase, i.e., DNA polymerase  $\beta$  (Pol  $\beta$ ), which also has a lyase activity that removes the sugar attached to 5' phosphate. The DNA polymerase then resynthesizes the omitted portion and finally nicks are sealed by DNA ligase enzyme. This process of damage removal is referred to as short patch BER and in some conditions it can be redirected to a longer patch process, often because the sugar is inefficiently removed from the 5' end of the nick. In this case, the damage-containing strand is displaced and removed by FEN1 and a variety of polymerases such as DNA polymerase  $\lambda$  (pol  $\lambda$ ) can take over. The nick sealed in

long patch BER is via Ligase 1 whereas the nick in short patch BER is sealed either by Ligase 1 or Ligase 3/XRCC1 [45]. Thus, the single base lesions in DNA (**Figure 1.6**) are removed in a systematic and precise manner via different DNA repair proteins (**Table 1.2**). The Table also represents the major associated diseases with literature references along with the KEGG IDs of the genes (if any) implicated in diverse pathways. A recent study conducted in The United States revealed that genetic variations in 19 genes of BER pathway are directly associated with the risk of bladder cancer [46] so a comprehensive list for various associated diseases has been prepared.

**Table 1.2 DNA repair proteins from base excision repair mechanism**

Proteins/ Genes	Full Name	Diseases	Literature References	KEGG ID
MBD4	methyl-CpG binding domain protein 4	Rett Syndrome, Angelman Syndrome, Colorectal cancer, Encephalopathy, Neurological disorders, Lung cancer	15857580, 11283202, 19127118, 2180070, 17965612, 18495292	hsa03410
MPG	N-methylpurine-DNA glycosylase	Lung cancer, Breast cancer, Ovarian cancer, NSCLC	10477488, 9684856, 17200364, 16613673	hsa03410
MUTYH	mutY homolog	FAP, Colorectal cancer, Microsatellite Instability, Rectal cancer, Colon cancer, Gastric cancer, Lung cancer, Hepatocellular carcinoma	19279422, 20223032, 17031395, 17703316, 19998059, 16929514, 14579148, 12658805	hsa03410
NEIL1	nei endonuclease VIII-like 1	Cancer	19443904	hsa03410
NEIL2	nei endonuclease VIII-like 2	Bladder cancer, Lung cancer	22701660, 22497777	hsa03410
NEIL3	nei endonuclease VIII-like 3	Prostate cancer	21810555	hsa03410
NTHL1	nth endonuclease III-like 1	Tuberous sclerosis, Cancer	9831664, 19443904	hsa03410
OGG1	8-oxoguanine DNA glycosylase	XP, Colorectal cancer, Gastric cancer, Neurodegenerative diseases, NSCLC, Lung cancer, Hepatocellular carcinoma, Breast cancer, Endometrial cancer	12187202, 19561388, 9765618, 15841414, 15551330, 17951408, 12658805, 19391486, 11465542	hsa03410
APEX1	APEX nuclease (multifunctional DNA repair enzyme) 1	XP, Colon cancer, Cervical cancer, Lung cancer, Breast cancer, Neurodegenerative diseases, Ovarian cancer, NSCLC, Colorectal cancer, Pancreatic cancer, Alzheimers disease, Prostate cancer	7775413, 18535621, 19292061, 10547596, 18669164, 20473298, 17974506, 19324449, 14767913, 18645011, 18672023, 16406883	hsa03410

APEX2	APEX nuclease 2	Growth retardation	15319281	hsa03410
LIG3	ligase III, DNA, ATP-dependent	Bloom syndrome, Cancer	8532526, 15257099	hsa03410
XRCC1	X-ray repair complementing defective repair in Chinese hamster cells 1	XP, NSCLC, Breast cancer, Colorectal cancer, Gastric cancer, Ataxia telangiectasia, Prostate cancer, Neurodegenerative diseases, HNSCC, Lung cancer	11016934, 19958624, 16457697, 18806752, 17593927, 7513822, 19428062, 18682529, 19230024, 16865671	hsa03410
SMUG1	single-strand-selective monofunctional uracil-DNA glycosylase 1	Cancer	22447450	hsa03410
TDG	thymine-DNA glycosylase	Tumors, Cancer, Lung cancer	9230216, 19402749, 15225156	hsa03410
UNG	uracil-DNA glycosylase	Bloom syndrome, Colon adenocarcinoma, Lung cancer, Neurodegenerative diseases, Glioblastoma, Colorectal cancer	2106500, 11554295, 15182505, 17207936, 9852992, 11172609	hsa03410, hsa05340
UNG2	uracil-DNA glycosylase	Lymphoma b-cell, Immunodeficiency	17062624, 17272283	hsa03410
PCNA	proliferating cell nuclear antigen	Gastric carcinoma, Breast cancer, XP, Colorectal carcinoma, NSCLC, Endometrial carcinoma, Bladder cancer, Prostate cancer, Lung cancer, Ovarian carcinoma	8101481, 8101708, 1354671, 7905362, 18279620, 15642194, 10072865, 18312749, 7923002, 10951706	hsa03030, hsa03410, hsa03420, hsa03430, hsa04110
POLB	polymerase (DNA directed), beta	Werner syndrome, Breast cancer, Gastric cancer, Colon carcinoma, Brain tumors, Colorectal cancer	8168825, 15764500, 10556592, 15280658, 8692923, 1511447	hsa03410
POLD1	polymerase (DNA directed), delta 1, catalytic subunit	Breast cancer	21455670	hsa00230, hsa00240, hsa03030
POLG	polymerase (DNA directed), gamma	Neurodegenerative diseases, Liver diseases, Infertility, Parkinson disease	15477547, 15122711, 15650046, 20399836	hsa01100
FEN1	flap structure-specific endonuclease 1	Werner syndrome, Bloom syndrome, XP, Colorectal cancer	11598021, 14688284, 9305916, 19218431	hsa03030, hsa03410
LIG1	ligase I, DNA, ATP-dependent	Bloom syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, XP, Pancreatic carcinoma, Lung cancer	1900268, 15966765, 15966765, 15966765, 12408985, 12009232	hsa03030, hsa03410, hsa03420
POLD3	polymerase (DNA-directed), delta 3, accessory subunit	Lung cancer	1849244	hsa00230, hsa00240, hsa03030

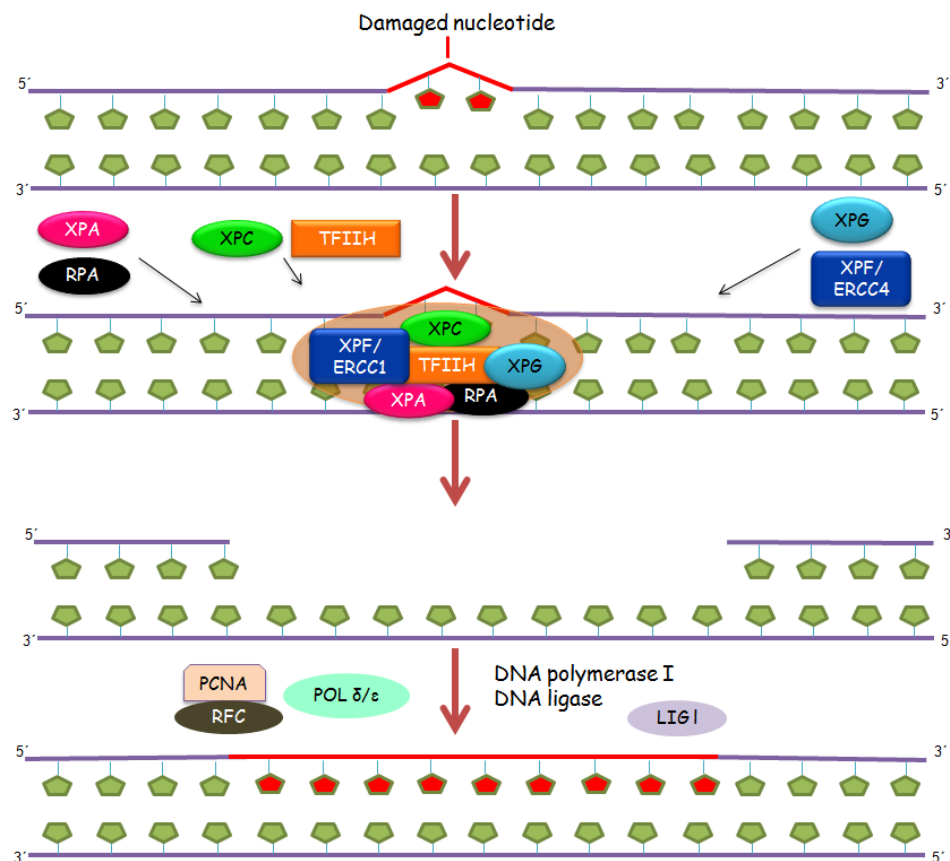


**Figure 1.6** Base excision repair mechanism for dealing with single base changes.

### 1.3.2.2 Nucleotide excision repair (NER)

Due to exposure of UV radiations, bulky and helix-distorting adducts such as CPDs and 6,4 photoproducts are introduced in the genome. These bulky aberrations are removed by NER mechanism which eliminates the lesion (~12-24 nucleotides long) enclosing single-stranded DNA segment. The damaged strand is then resynthesized on the basis of complementarity with the undamaged DNA strand by DNA polymerase and thus the final ligation is performed by DNA ligase. NER mainly deals with helix distorting lesions arising from exogenous sources that primarily impedes transcription and normal replication. The NER pathway is further classified into two sub-pathways i.e. transcription coupled NER (TC-NER) which is especially dedicated to the actively transcribed genes i.e. which swiftly repair regions of DNA which are “active” and are transcribed into RNA and the other is global genomic NER (GG-NER) that acts slowly on the entire genome [38, 47].

In humans, NER involves many proteins for the recognition, excision and resynthesis of native DNA (**Figure 1.7**). Xeroderma pigmentosum, complementation group C (XPC) and XPE (DDB2) are the DNA binding proteins that recognizes damage to DNA [48, 49]. Specifically in GGR, the unwinding of damaged site is carried out by XPB (ERCC3) and XPD (ERCC2) helicases [50, 51], which are a component of TFIIH (Transcription factor II Human) complex [52]. XPA protein is required for stabilizing the unwound region and then XPG (ERCC5) and XPF (ERCC4) cuts the DNA on either side of damage so as the intact DNA can replace the damaged portion [53, 54]. The XPC and XPE proteins are the major components of GGR whereas other NER-related XP groups (XP-A, B, D, F and G) are found in both the sub-pathways (GGR and TCR). If mutations in any of the above genes are not rectified, the damage by UV may also cause mutations in the cell's DNA [55]. The deficiency in NER system is associated with a disorder of dry and pigmented skin recognized as xeroderma pigmentosum (XP) [38, 56] and numerous other disorders as listed in **Table 1.3**.



**Figure 1.7** Nucleotide excision repair mechanism for the removal of bulky and helix distorting lesions.

**Table 1.3 A list of nucleotide excision repair genes and associated diseases**

Genes	Full Name	Diseases	Literature References	KEGG ID
PCNA	proliferating cell nuclear antigen	Hepatocellular carcinoma, Breast cancer, Colon cancer, XP, Colorectal carcinoma, NSCLC, Bladder cancer, Cervical carcinoma, Prostate cancer, Ovarian carcinoma	7911731, 8101708, 9149129, 1354671, 7905362, 18279620, 10072865, 15617346, 18312749, 10951706	hsa03030, hsa03410, hsa03420, hsa03430, hsa04110
POLD1	polymerase (DNA directed), delta 1, catalytic subunit	Breast cancer	21455670	hsa00230, hsa00240, hsa03030, hsa03410
POLE	polymerase (DNA directed), epsilon, catalytic subunit	None	None	hsa00230, hsa00240, hsa03410
DDB1	damage-specific DNA binding protein 1, 127kDa	XP, Cancer, Tumor	9632823, 16260596, 2333286	hsa03420, hsa04120
DDB2	damage-specific DNA binding protein 2, 48kDa	XP, Breast cancer, Lung cancer, Melanoma	10771487, 19339246, 16522664, 16888633	hsa03420, hsa04115, hsa04120
ERCC1	excision repair cross-complementing rodent repair deficiency, complementation group 1	XP, NSCLC, Cockayne syndrome, Lung cancer, Ovarian cancer, Breast cancer, Colon cancer, Testicular cancer, Brain tumors, Prostate carcinoma	8972858, 19538866, 10910954, 17502833, 18756932, 18495602, 18497992, 15885892, 9703867, 12643788	hsa03420
ERCC2	excision repair cross-complementing rodent repair deficiency, complementation group 2	XP, Cockayne syndrome, Lung cancer, NSCLC, Fanconi anemia, Bladder cancer, Colorectal cancer, Ovarian cancer, Leukemia, Breast carcinoma	11709541, 11710928, 15182505, 16061005, 16973432, 20514470, 19561388, 10738106, 16537383, 18752184	hsa03022, hsa03420
ERCC3	excision repair cross-complementing rodent repair deficiency, complementation group 3	Trichothiodystrophy, Cockayne syndrome, XP, Leukemia, Lung cancer, Ovarian cancer	11062469, 16947863, 1454518, 16537383, 16550608, 10948350	hsa03022, hsa03420
ERCC4	excision repair cross-complementing rodent repair deficiency, complementation group 4	XP, Fanconi anemia, Lung cancer, Bladder cancer, Prostate cancer, Breast cancer	20062074, 12571280, 18068852, 20062074, 19902366, 15886521	hsa03420
ERCC5	excision repair cross-complementing rodent repair deficiency, complementation group 5	XP, Cockayne syndrome, Trichothiodystrophy, Lung cancer, Growth retardation, Prostate cancer	11479630, 9864391, 11104904, 12869423, 15082767, 10517877	hsa03420
ERCC6	excision repair cross-complementing rodent repair deficiency, complementation group 6	Cockayne syndrome, XP, Trichothiodystrophy, Neurodegenerative diseases	9312053, 11782547, 11104904, 17055654	hsa03420
ERCC8	excision repair cross-complementing rodent repair deficiency, complementation group 8	Cockayne syndrome, XP, Genetic disorder, Neurodegenerative diseases	8811173, 11782547, 8811173, 17055654	hsa03420, hsa04120



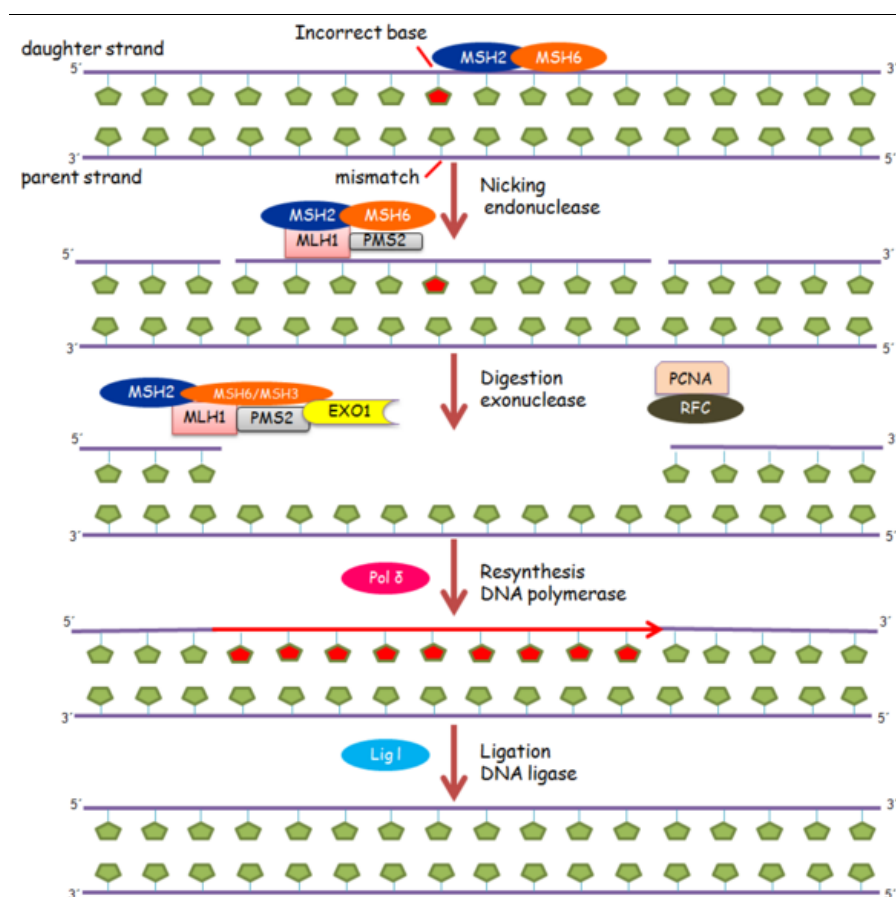
GTF2H2	general transcription factor IIH, polypeptide 2, 44kDa	Muscular atrophy spinal, Werdnig-hoffmann disease, Cockayne syndrome, XP	17903057, 8981949, 8652557, 8652557	hsa03022, hsa03420
LIG1	ligase I, DNA, ATP-dependent	Bloom syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, XP, Pancreatic carcinoma, Lung cancer	1900268, 15966765, 15966765, 15966765, 12408985, 12009232	hsa03030, hsa03410, hsa03420, hsa03430
RPA1	replication protein A1, 70kDa	XP, Werner syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, Breast cancer	8972858, 15965237, 18003706, 15929725, 15102447	hsa03030, hsa03420, hsa03430
RPA2	replication protein A2, 32kDa	Ataxia telangiectasia, XP, Colon cancer, Breast cancer	16483312, 10340474, 17361204, 11895905	hsa03030, hsa03430
TFIIH	general transcription factor IIH, polypeptide 3	Trichothiodystrophy, Cockayne syndrome, XP, Lung cancer, Ovarian cancer	11062469, 16947863, 1454518, 16550608, 10948350	hsa03022, hsa03420
XPA	xeroderma pigmentosum, complementation group A	XP, Trichothiodystrophy, Cockayne syndrome, Ataxia telangiectasia, Lung cancer, Ovarian cancer, Neurological disorders	1918083, 11104904, 9415314, 16862173, 15333465, 8973600, 8765158	hsa03420
XPC	xeroderma pigmentosum, complementation group C	XP, Cockayne syndrome, Bladder cancer, Lung cancer, Breast cancer, NSCLC, Colorectal cancer, Prostate cancer	15964821, 9415314, 17052994, 17705814, 18196582, 17508409, 12242345, 19902366	hsa03420
RFC1	replication factor C (activator 1) 1, 145kDa	Retinoblastoma, Leukemia, Tumors	15897598, 9279361, 12509469	hsa03030, hsa03430
POLD3	polymerase (DNA-directed), delta 3, accessory subunit	Lung cancer	1849244	hsa01100, hsa03030, hsa03410

### 1.3.2.3 Mismatch Repair (MMR)

During replication, mismatches are inserted in DNA due to aberrant proof reading by the DNA polymerase. The errors or mismatch of bases are then removed from DNA via mismatch repair system by first detecting the mismatch and then recruiting an endonuclease that cleaves the newly synthesized DNA strand close to the region of damage followed by resynthesis and nick sealing of native DNA. The first evidence for MMR was obtained from *Streptococcus pneumonia* and then a number of mutational inactivated genes that caused hypermutable strains were identified in *Escherichia coli* [57, 58]. After replication, MMR enzymes travel down the new DNA molecules to identify mistakes in the form of a bulge resulting from a mismatched pair. When an error is sensed, the MMR enzymes further activates other enzymes (discussed below) that complete the process of repair. MMR maintains DNA homeostasis by facilitating post-replication repair, policing homologous recombination events and protecting against genetic exchange between species. Loss of MMR function results in the accumulation of potential

mutations and there are various disorders reported due to mutations in MMR genes which effect the genomic stability and result in microsatellite instability (MSI) [59].

The process of damage removal is initiated by MutS $\alpha$  (MSH2-MSH6) or MutS $\beta$  (MSH2-MSH3) which recognizes lesion by binding to the mismatch, then MutL $\alpha$  (MLH1-PMS2) is recruited to the heteroduplex [60]. The assembly of MutL-MutS-heteroduplex ternary complex in presence of RFC and PCNA is sufficient to activate endonuclease activity of PMS2 [61, 62]. It introduces SSBs near the mismatch and thus generates new entry points for EXO1 to degrade the strand containing mismatch. DNA methylation ensures that only the newly mutated DNA strand is to be corrected thus preventing any incorrect cleavage. The MutL $\alpha$  (MLH1-PMS2) is known to interact physically with the clamp loader subunits of DNA polymerase III suggesting its role in recruiting the polymerase at the site of mismatch. The entire mechanism of damage removal via MMR is demonstrated in **Figure 1.8** and the implicated genes in **Table 1.4**.



**Figure 1.8** Mismatch repair system for the removal of mismatches in DNA.

**Table 1.4 Mismatch repair genes and associated diseases**

Genes	Full Name	Diseases	Literature References	KEGG ID
MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	HNPCC, Endometrial carcinoma, Ovarian carcinoma, Gallbladder carcinoma, Breast cancer, Werner syndrome, HNSCC, NSCLC	8895729, 11391585, 16879751, 16309194, 18329696, 17412712, 19786217, 15583832	hsa03430, hsa05200, hsa05210, hsa05213
MLH3	mutL homolog 3 (E. coli)	Colorectal cancer, Somatic mutations, Cancer	11317354, 10934138, 18521850	hsa03430
MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	HNPCC, Endometrial cancer, Breast cancer, Brain tumors, Ovarian carcinoma, Cervical cancer, Neurodegenerative diseases	18999873, 16803540, 17922223, 16372347, 16879751, 10999751, 9215683	hsa03430, hsa05200, hsa05210
MSH3	mutS homolog 3 (E. coli)	Endometrial carcinoma, Colorectal carcinoma, Ovarian cancer, Bladder cancer	9699180, 14871813, 16774946, 15541380	hsa03430, hsa05200, hsa05210
MSH4	mutS homolog 4 (E. coli)	None	None	None
MSH5	mutS homolog 5 (E. coli)	Colorectal cancer	9740671	None
MSH6	mutS homolog 6 (E. coli)	Colorectal cancer, Endometrial cancer, HNPCC, Ovarian cancer, Breast cancer, Prostate cancer	19492230, 15805151, 16237223, 12376742, 15805151, 18355840	hsa03430, hsa05200, hsa05210
PMS1	PMS1 postmeiotic segregation increased 1 (S. cerevisiae)	HNPCC, Gastric cancer, Ovarian cancer, Neck cancer, Breast cancer, Prostate cancer	8895729, 15133479, 16774946, 9568786, 12851690, 11444857	hsa03430
PMS2	PMS2 postmeiotic segregation increased 2 (S. cerevisiae)	HNPCC, Gastrointestinal tumor, Ovarian cancer, Breast cancer	10671064, 17258725, 18723338, 16227397	hsa03430
PCNA	proliferating cell nuclear antigen	Hepatocellular carcinoma, Breast cancer, XP, Colorectal carcinoma, NSCLC, Bladder cancer, Cervical carcinoma, Prostate cancer	7911731, 8101708, 1354671, 7905362, 18279620, 10072865, 15617346, 18312749	hsa03030, hsa03410, hsa03420, hsa03430
EXO1	exonuclease 1	HNPCC, Werner syndrome, Colorectal cancer, Lung cancer	14623461, 12704184, 11375940, 18079015	hsa03430
RPA4	replication protein A4, 30kDa	None	None	hsa03030, hsa03420
LIG1	ligase I, DNA, ATP-dependent	Bloom syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, XP, Pancreatic carcinoma, Lung cancer	1900268, 15966765, 15966765, 15966765, 12408985, 12009232	hsa03030, hsa03410, hsa03420, hsa03430
RPA1	replication protein A1, 70kDa	XP, Werner syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, Breast cancer	8972858, 15965237, 18003706, 15929725, 15102447	hsa03030, hsa03420, hsa03430, hsa03440
RPA2	replication protein A2, 32kDa	Ataxia telangiectasia, XP, Colon cancer, Breast cancer	16483312, 10340474, 17361204, 11895905	hsa03030, hsa03420, hsa03430, hsa03440
RFC1	replication factor C (activator 1) 1, 145kDa	Retinoblastoma, Leukemia, Tumors	15897598, 9279361, 12509469	hsa03030, hsa03420, hsa03430
POLD3	polymerase delta 3, accessory subunit	Lung cancer	1849244	hsa00230, hsa03030

### 1.3.3 Double-strand damage

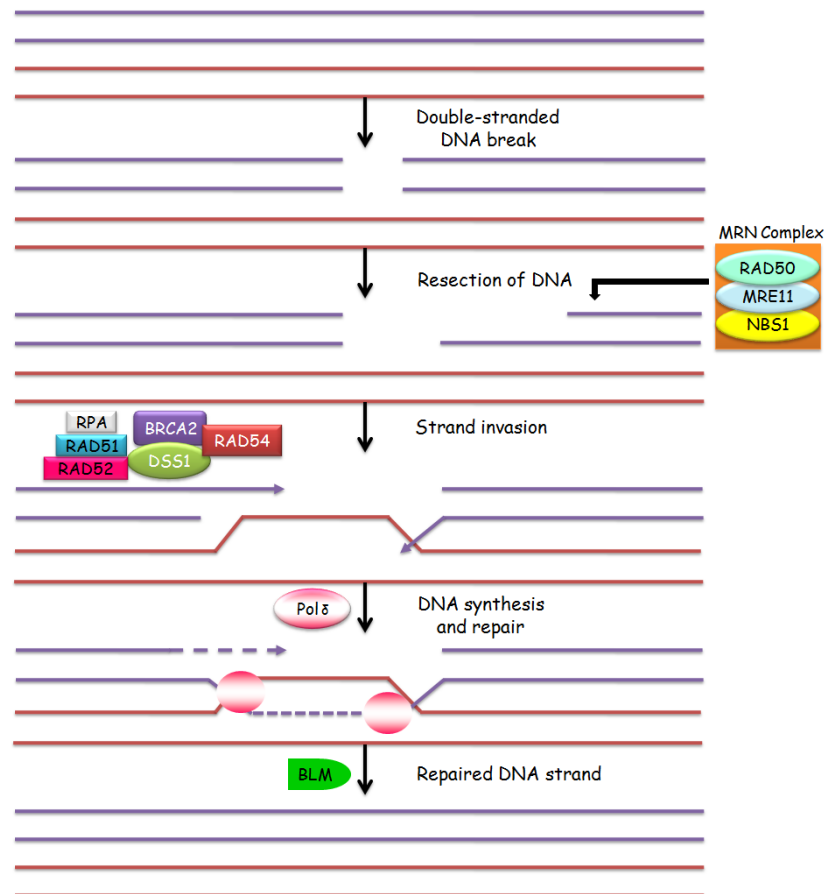
The double-strand damages in DNA lead to DSBs which are quite challenging to repair since both the strands are affected. For effective guiding of repair, the main task for the cell is to know which ends belong together and the problem magnifies many folds depending on the size of mammalian genome [38]. There are major two pathways that deal with DSBs, i.e. HRR and NHEJ which direct the repair by different means and are described in detail below:

#### 1.3.3.1 Homologous recombination repair (HRR)

HRR is a specialized DNA repair mechanism for repairing or tolerating complex damages such as DSBs and DNA interstrand crosslinks (ICLs) which arise in the genome due to ionizing radiations, drugs used for cancer treatment and errors in DNA replication [63]. HRR mechanism not only performs template dependent repair to maintain genome integrity, but also has implications in authentic duplication of genome and telomere maintenance. When a DSB is detected in DNA, region around the 5' ends of the break is removed by resection followed by invasion where an overhanging 3' end of the broken DNA molecule invades a similar or identical DNA molecule that is not broken. After the strand invasion step, either of the two sub-pathways are activated i.e. DSBR (double-strand break repair) pathway or the SDSA (synthesis-dependent strand annealing) pathway. The DSBR pathway mostly results in crossing over whereas in SDSA pathway, non-crossover recombinants are produced. The HRR occurring during the repair process favors SDSA pathway rather than DSBR pathway as it results in non-crossover products and restores the native DNA molecule as it was before the damage [64].

**Figure 1.9** illustrates the mechanism underlying HRR to repair DSBs which initiates by sensing damage [65] through Mre11-Rad50-Nbs1 (MRN) complex that resects 5' ends around the break leaving 3'-ssDNA overhangs. RPA then binds with high affinity to the ssDNA which further is displaced by Rad51 and Rad52. The nucleation of Rad51 on ssDNA promotes the formation of a nucleoprotein filament that catalyzes the homology search and subsequent strand exchange [66]. Brca2 is also recruited along with other accessory factors such as Rad54 [67] to the damaged sites to initiate HRR through a sister chromatid exchange [68-70]. Finally, the resolvases restore the holiday junctions and error-free DNA molecules are thus retrieved. There

are a huge number of mutations reported in HRR genes (**Table 1.5**) that are responsible for majority of cancers.



**Figure 1.9** Repair of double-strand breaks by homologous recombination repair mechanism.

### 1.3.3.2 Non-homologous end joining (NHEJ)

Besides HRR, DSBs are repaired by another mechanism i.e., NHEJ which functions mainly in G<sub>0</sub>, G<sub>1</sub> and early S phases of the cell cycle [71] since HRR is down regulated during these phases in multi-cellular eukaryotes [72]. NHEJ is considered less accurate as compared to HRR that predominantly operates at S and G<sub>2</sub> phases of cell cycle (DNA is replicated) due to the fact that HRR uses homologous sequences for repair whereas NHEJ simply ligates the broken ends, often its error prone. **Table 1.6** presents NHEJ DNA repair genes and their implication in different diseases. The NHEJ mechanism functions by two subsequent repair steps, initially, Ku protein and the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) bind to DNA ends and the process is referred as synapsis where the end alignment may either have a terminal microhomology of ~1–4 nucleotides or non-homology between the two ends.

**Table 1.5 DNA repair genes implicated in homologous recombination repair**

Genes	Full Name	Diseases	Literature References	KEGG ID
H2AFX	H2A histone family, member X	Ataxia telangiectasia, Nijmegen breakage syndrome, Fanconi anemia, Lung cancer, Ovarian cancer, Breast cancer, Colorectal cancer	15059890, 14712078, 17471025, 16820894, 15975956, 18644834, 18499365	hsa05322
POLD1	polymerase (DNA directed), delta 1, catalytic subunit	Breast cancer	21455670	hsa00230, hsa03030, hsa03410
BRCA2	breast cancer 2, early onset	Breast cancer, Fanconi anemia, Ovarian cancer, Ataxia telangiectasia, Prostate cancer, Colorectal cancer, Pancreatic carcinoma, HNPCC	9145678, 19530235, 10560359, 12393516, 19476645, 10458128, 8968085, 17601911	hsa03440, hsa05200, hsa05212
BRIP1	BRCA1 interacting protein C-terminal helicase 1	Fanconi anemia, Breast cancer, Ovarian cancer, Cancer, Anemia, Tumors	16153896, 16430786, 12565990, 20173781, 17768402, 18345034	hsa03460
BRCA1	breast cancer 1, early onset	Breast cancer, Ovarian cancer, Fanconi anemia, Prostate cancer, HNPCC, Endometrial cancer	15564800, 10560359, 12483114, 19223505, 17601911, 16962648	hsa04120
MRE11A	MRE11 meiotic recombination 11 homolog A ( <i>S. cerevisiae</i> )	Nijmegen breakage syndrome, Ataxia telangiectasia, Fanconi anemia, Werner syndrome, Colorectal cancer, Breast cancer	10391882, 16905549, 11733219, 11733219, 11850399, 19383352	hsa03440, hsa03450
RAD50	RAD50 homolog ( <i>S. cerevisiae</i> )	Ataxia telangiectasia, Colorectal cancer, Breast cancer, Gastric carcinoma	16971555, 18830935, 16385572, 12007281	hsa03440, hsa03450
RAD51	RAD51 homolog ( <i>S. cerevisiae</i> )	Fanconi anemia, Ataxia telangiectasia, Ovarian cancer, NSCLC, Breast carcinoma, Colon cancer, Pancreatic cancer, Prostate cancer	12239151, 15135073, 11535547, 19799875, 10962436, 16516153, 19139112, 20002770	hsa03440, hsa05200, hsa05212
RAD52	RAD52 homolog ( <i>S. cerevisiae</i> )	Germ-line mutation, Breast cancer, Cancer, Ovarian cancer	17405295, 10463575, 18449888, 12883740	hsa03440
RAD54L	RAD54-like ( <i>S. cerevisiae</i> )	Bloom syndrome, Werner syndrome, XP	10640146, 10640146, 10640146	hsa03440
BLM	Bloom syndrome, RecQ helicase-like	Bloom syndrome, Werner syndrome, Ataxia telangiectasia, Fanconi anemia, Nijmegen breakage syndrome	15137905, 15702347, 16199871, 19738377, 11733219	hsa03440
RAD51B	RAD51 homolog B ( <i>S. cerevisiae</i> )	Uterine leiomyoma, Chromosomal aberrations, Tumors	11135437, 16778173, 16778173	hsa03440
LIG1	ligase I, DNA, ATP-dependent	Bloom syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, XP, Pancreatic carcinoma, Lung cancer	1900268, 15966765, 15966765, 15966765, 12408985, 12009232	hsa03030, hsa03410, hsa03420, hsa03430
RPA1	replication protein A1, 70kDa	XP, Werner syndrome, Nijmegen breakage syndrome, Ataxia telangiectasia, Breast cancer	8972858, 15965237, 18003706, 15929725, 15102447	hsa03030, hsa03430, hsa03440
POLD3	polymerase (DNA-directed), delta 3, accessory subunit	Lung cancer	1849244	hsa00230, hsa03030, hsa03410
ATM	ataxia telangiectasia mutated	Ataxia telangiectasia, Breast cancer, Ovarian cancer, Lung cancer, Colorectal cancer, Neurodegenerative diseases	9622061, 11830610, 9949303, 18164969, 10203610, 17100756	hsa04110, hsa04115, hsa04210

The repair mechanism as depicted in **Figure 1.10** involves binding of Ku heterodimer with two subunits- Ku70 and Ku80 to damaged DNA in the form of ring shaped structure whereby the damaged end threading through middle of the ring [73]. Further to bridge the two strands, DNA-PKcs interacts with Ku forming a complete DNA-dependent protein kinase [74]. The damaged end is then removed and processed by Artemis–DNA-PKcs complex and gaps are filled by polymerases. The final step is the ligation which is performed by the XRCC4–DNA-ligase-IV (LIG4) complex [75, 76].

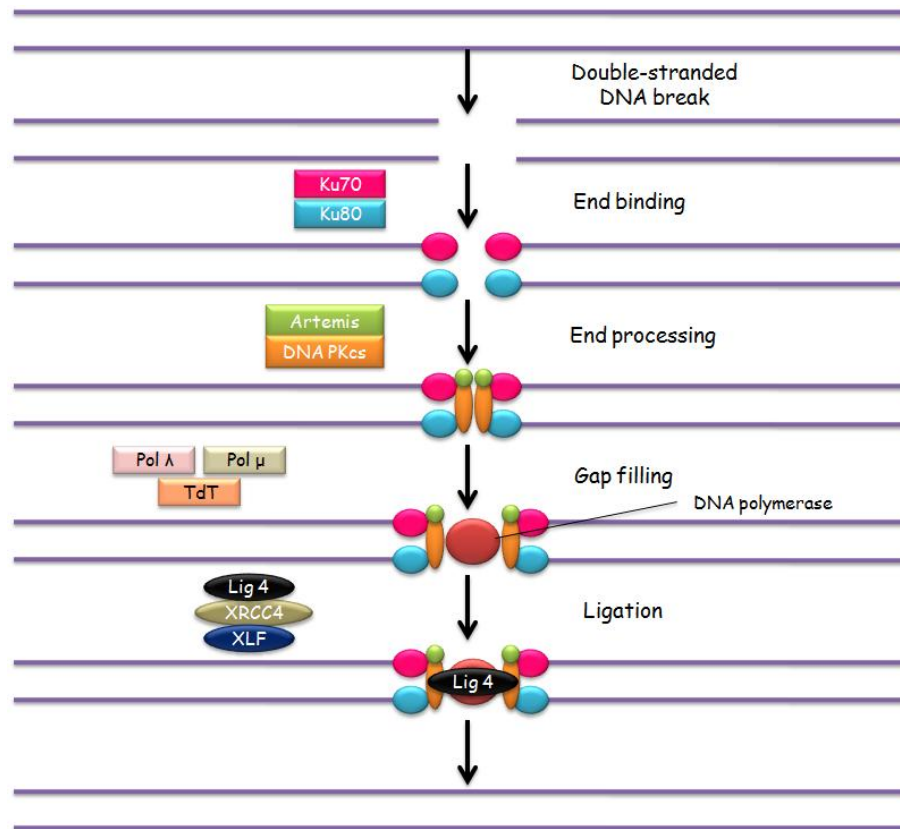
**Table 1.6 Genes drawn in non-homologous end joining repair mechanism**

Genes	Full Name	Diseases	Literature References	KEGG ID
RAD50	RAD50 homolog (S. cerevisiae)	Nijmegen breakage syndrome, Fanconi anemia, Colorectal cancer, Breast cancer	10888888, 17524422, 18830935, 16385572	hsa03440, hsa03450
POLL	polymerase (DNA directed), lambda	None	None	hsa03410, hsa03450
LIG4	ligase IV, DNA, ATP-dependent	Lig4 syndrome, XP, Lung cancer, Leukemia, Breast cancer	15333585, 15966765, 15609317, 1349135, 12023982	hsa03450
XRCC4	X-ray repair complementing defective repair in Chinese hamster cells 4	Autoimmune response, Leukemia, Breast cancer	12218164, 20332465, 12750264	hsa03450
XRCC5	X-ray repair complementing defective repair in Chinese hamster cells 5	Ataxia telangiectasia, Werner syndrome, Cervical cancer, Leukemia, Lung cancer , Breast cancer, Colorectal cancer	9636207, 9636207, 12902903, 15215044, 17289874, 17982634, 15254745	hsa03450
XRCC6	X-ray repair complementing defective repair in Chinese hamster cells 6	Bloom syndrome, Ataxia telangiectasia, NSCLC, Colorectal cancer, Cervical cancer, Breast cancer, Prostate cancer	15184650, 12210072, 15273727, 15254745, 19672258, 20496270, 17545612	hsa03450
WRN	Werner syndrome, RecQ helicase-like	Werner syndrome, Bloom syndrome, Ataxia telangiectasia	12927431, 10319867, 18209099	None
TDP1	tyrosyl-DNA phosphodiesterase 1	Spinocerebellar degenerations, Neurodegenerative diseases	16793421, 15744309	None
POLM	polymerase (DNA directed), mu	Burkitt lymphoma	15520469	hsa03450

### 1.3.4 Translesion synthesis (TLS)

Although cells can deal with almost all forms of lesions with the aid of highly complex DNA repair systems but circumstantially when the lesion still exists before initiation of replication, it can block the replication machinery and ultimately lead to cell death. In such situations, there is a need for dedicated mechanisms for tolerating the DNA damage without mediating repair of a lesion. These DNA damage tolerance processes still maintain the integrity and survive damages

to genome but often lead to mutations. In normal repair systems, the DNA polymerases cannot bypass these lesions and the replication halts (if not repaired) but in the damage tolerance process or TLS, there are specialized polymerases [77] recruited on the damage called translesion polymerases.

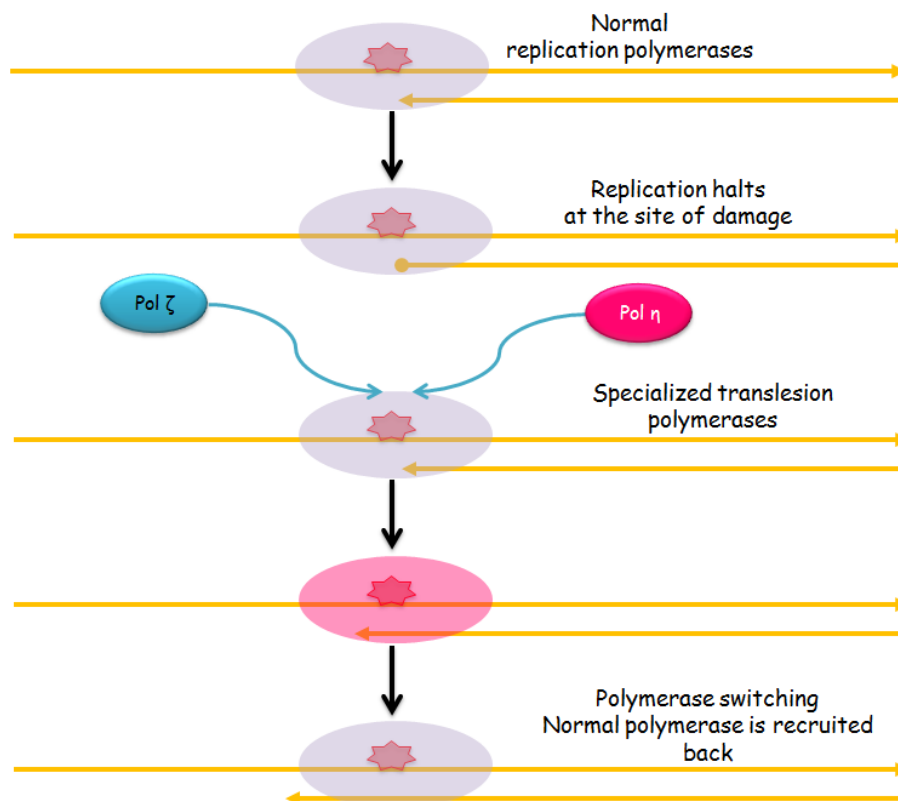


**Figure 1.10** The NHEJ mechanism repairing DSBs in the absence of homologous strand.

These translesion polymerases replace the stalled replicative polymerases at the 3'-OH end of a primed DNA template in a process called polymerase switching [78]. After following synthesis over the stretch contacting lesion, the process is reversed and replicative polymerases again resume the synthesis (**Figure 1.11**). These translesion polymerases are usually error-prone and have implications in a variety of cancers as depicted in **Table 1.7**. There are a few specialized polymerases that accurately bypass specific classes of DNA lesions whereas others bypass the same lesion with high error rates [79-81]. Thus, recruitment of inappropriate specialized polymerase during TLS could also result in mutations or genome rearrangements. A list of specialized DNA polymerases in eukaryotes is portrayed in **Table 1.8**. In order to



comprehensively understand various cellular processes such as cell stress responses, DNA damage following cell death, genomic integrity, generation of mutations and disease development and progression, gaining insights on these potentially mutagenic yet highly conserved polymerases is vital [82].



**Figure 1.11** The bypassing of lesions via translesion synthesis mechanism.

**Table 1.7** The DNA repair genes concerned with translesion synthesis mechanism

Gene Name	Full Name	Diseases	Literature References	KEGG ID
REV3L	REV3-like, polymerase (DNA directed), zeta, catalytic subunit	Colon carcinoma, Tumors, Cancer	16428501, 10660610, 16428501	hsa01100
POLH	polymerase (DNA directed), eta	XP, Skin cancer, Somatic mutations, Cancer	12244178, 12967656, 16823845, 12546696	None
POLI	polymerase (DNA directed) iota	XP, Burkitt lymphoma, Tumors	17409408, 12410315, 17056006	None
POLK	polymerase (DNA directed) kappa	Lung cancer	15202001	None
POLM	polymerase (DNA directed), mu	Burkitt lymphoma	15520469	hsa03450
POLN	polymerase (DNA directed) nu	None	None	None
POLQ	polymerase (DNA directed), theta	Breast cancer	21301045	None
REV1L	REV1, polymerase (DNA directed)	Lung cancer	19176310	None

**Table 1.8 Specialized translesion polymerases in eukaryotes**

Polymerase	Gene	Family	Putative functions
η (eta)	POLH	Y	Bypass UV lesions
ι (iota)	POLI	Y	Bypass synthesis
κ (kappa)	POLK	Y	Bypass synthesis
λ (lambda)	POLL	X	Base excision repair, NHEJ
μ (mu)	POLM	X	NHEJ
θ (theta)	POLQ	A	DNA repair
ζ (zeta)	POLZ	B	Bypass synthesis
Rev 1	REV1	Y	Incorporation of dC opposite abasic sites
ν (nu)	POLN	A	Unknown but Pol ν has unique error signature, G-dTMP mismatches

These are the major DNA repair pathways known that confiscates diverse forms of lesions in different ways and involve varied set of enzymes and mechanisms for repairing DNA to sustain the integrity of genome. In addition, these pathways also trigger other cellular responses to deal with damage removal via DNA damage response and signaling. The key mediators for the signaling are ATM and ATR kinases that facilitate repair process via their downstream targets and induce cell cycle arrest [83]. These processes have gained a lot of attention for cancer therapy since resistance to genotoxic therapies has been associated to damage response signaling and targeting the process will prove vital. Over recent years, the role of DNA repair pathways and abnormalities in them have provided insights in understanding the development of numerous diseases such as aging [84], cancer [85], neurological aberrations [86] and also their therapeutic relevance [87]. In view of the significance of DNA repair in human diseases, we have compiled a list of diseases occurring due to mutations or other abnormalities in DNA repair system in the upcoming section. The objectives designed for my research work also focus on some of the diseases and the role of associated DNA repair mechanisms and pathways.

#### 1.4 DISEASES SPECIFIC TO DNA REPAIR

Genetic instability leading to carcinogenesis is known to stimulate by DNA damages and errors created by the DNA machinery [88]. These damages unless repaired lead to mutations in DNA repair genes, impacting the phenotypic consequences and hence give rise to numerous human genetic diseases [55, 59, 89] by escalating the predisposition to a variety of cancers [10, 28]. Defects in all the above mentioned (**Section 1.3**) repair mechanisms have been reported to associate with several disorders therefore appropriate understanding of these intricate

mechanisms is essential to comprehend the underlying human genetic diseases. In the following section, we provide a brief description for a few DNA repair associated disorders along with the respective biomarkers.

#### **1.4.1 Xeroderma Pigmentosum (XP)**

XP is a rare autosomal genetic disorder in which extreme sensitivity to UV radiations is observed [90]. It results due to aberrations or mutations in NER associated genes [91] (Ddb2, Ercc2, Ercc3, Ercc4, Ercc5, Xpa and Xpc) and Polh, which corresponds to TLS mechanism [92]. The normal ability of cell to repair UV induced damages (thymine dimers) is affected in XP due to aberrations in these genes [93]. The disorder is mainly manifested by symptoms such as freckles, dark patches, corneal ulcerations and often neurological abnormalities. XP increases the susceptibility to develop skin cancer and other form of cancers i.e. cancer on eyes, lips and ears. XP is predominantly found in the Japanese population [94] and prevalent in both males and females with equal probabilities [90]. The two DNA repair mechanisms thus contributing towards XP are NER, specifically GG-NER and TLS repair mechanisms.

#### **1.4.2 Cockayne syndrome (CS)**

CS is a genetic autosomal recessive disorder characterized by growth retardation (short stature), neurological abnormalities, premature aging and sensitivity to sunlight. There are four major variations of CS i.e. CS Type I, II, III and XP-CS where CS type I to III are classified on the basis of severity and age of onset of disease whereas in XP-CS, person suffers from both XP and CS. Mutations in two DNA repair genes, i.e. ERCC6, also known as Cockayne syndrome complementation group type B (CSB) [95, 96] and ERCC8, also referred as Cockayne syndrome complementation group type A (CSA) [97] are responsible for CS. Both the genes are implicated in TC-NER, the mechanism activated in highly transcribed region. Mutations in ERCC6 gene alone contributes to ~70% of CS cases [98]. If either of the genes is mutated or altered, DNA is not repaired and the damage is accumulated hence stimulating cell death.

#### **1.4.3 Werner syndrome (WS)**

WS is exemplified by premature aging and is also known as adult progeria [99]. The clinical manifestations of the disorder includes juvenile bilateral cataracts, mask-like face, bird-like nose,

diabetes mellitus, atherosclerosis and osteoporosis [100, 101]. It has been reported that mutations in WRN gene forms the basis for WS since the mutation leads to production of an abnormally short nonfunctional werner protein [102]. The growth failure observed in WS is due to the altered werner protein which enforces the cells to either divide slowly or stop dividing as compared to normal. This mutation also causes damage to accumulate in the genome thus impairing the normal cellular activities. The normal WRN gene encodes a protein whose central domain resembles members of the RecQ helicases that play an important role in repairing DSBs i.e. mainly implicated in HRR and NHEJ mechanisms. Thus, mutations in the gene may impact genetic stability and lead to development of WS [103].

#### **1.4.4 Ataxia telangiectasia (A-T)**

A-T, also known as Louis–Bar syndrome is a rare genetic neurodegenerative disease wherein ataxia refers to poor coordination ability and telangiectasia to the small dilated blood vessels [104]. The symptoms of disease include neurological abnormalities since cerebellum, the region of brain functioning in movement and coordination is affected, weakening of immune system and increased susceptibility of developing multiple cancers. A-T results due to aberrations in ATM gene [105], which normally assists in repairing DSBs in DNA and plays crucial role in normal development of nervous and immune systems. Due to mutation in the gene, normal repair doesn't take place resulting in damage accumulation and subsequent cell death in certain areas, for instance, brain. In addition, the A-T patients are more prone (~25% lifetime risk) to cancers specifically lymphomas and leukemia [106]. A-T is inherited in an autosomal recessive pattern, i.e. mutations in both the copies of ATM gene contribute to the disease.

#### **1.4.5 Hereditary nonpolyposis colorectal cancer (HNPCC)**

HNPCC, also referred as lynch syndrome is an autosomal dominant disorder which increases the susceptibility to develop colon cancer [107]. In HNPCC patients, other cancers such as endometrial, skin, ovary, brain, stomach also have high degree of prevalence [108, 109]. Mutations in MMR genes [110] are mainly responsible for HNPCC which lead to MSI [111]. The five aberrant MMR genes implicated in developing HNPCC are MLH1, MSH2, MSH6, PMS2 and EPCAM. MLH1 and MSH2 account for about 70 to 80 % cases whereas mutations in MSH6, PMS2 and EPCAM gene also contribute towards the severity of disease. The mutations

in these MMR genes and altered microsatellites in the coding genes for tumor initiation and progression are the major factors concerning the disease.

#### **1.4.6 Fanconi anemia (FA)**

FA is another DNA repair associated autosomal recessive disorder characterized by congenital defects such as short stature, growth failure and abnormalities in skin, ears, kidneys and eyes [112]. Leukemia and bone marrow failure are common traits of the disease i.e. the person's capability to produce blood cells is distorted [112]. FA is an outcome of mutations in 15 DNA repair genes i.e. BRCA2, BRIP1, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, PALB2, RAD51C and SLX4 involved in FA pathway [113, 114]. When the DNA replication halts due to DNA damage, the FA pathway is activated which recruits certain proteins at the site of damage which repair damages and replication can then continue [115]. Thus, these genes play vital roles in repairing DNA and maintaining integrity of genome. Mutations in these genes have adverse impact on the damage accumulation and finally contribute to cell death and malignancies [114, 115].

### **1.5 SYSTEMATIC PERSPECTIVE FOR DNA REPAIR MECHANISMS**

A high fidelity of transmission of genetic information is absolutely necessary in ensuring normal life development. Genomic stability is essential for the maintenance of life. Eukaryotic cells exposed to DNA damaging agents also activate important defensive pathways by inducing multiple proteins involved in DNA repair. Specifically, the p53 (or TP53) protein which is involved in many anti-cancer and apoptosis mechanisms is only present in animals while it is absent in plants and fungi. This protein plays an important role in human as it is an activator of components of the NER pathway [116, 117]. In addition, extensive studies over the last two decades have generated a wealth of information on the DNA repair systems and pathways of cell. We now have an improved understanding of DNA damaging factors and the myriad of mechanisms by which cells protect their genomic integrity. DNA damage response can be described as a collaborative effect of DNA repair systems and cell-cycle regulation. Proteins involved in DNA damage checkpoints are found to perform similar functions in regular cell-cycle control processes. Some proteins such as p53 and Rad51 have regulatory functional effects

over multiple pathways, establishing the linkages between different pathways in the network [118, 119].

Evolution in the field of human genomics and continuous innovations and developments in the areas such as genetics, high-throughput techniques, comparative genomics, systems biology and bioinformatics has added a new dimension to the biomedical research. There are evidences of decreased lifespan and increased cancer incidence in experimental animals with genetic deficiency in DNA repair. Inherited mutations that affect DNA repair genes are strongly associated with high cancer risk in humans as already discussed in section 1.4. HNPCC is strongly associated with specific mutations in genes involved in MMR pathway. BRCA1 and BRCA2, two well known mutations conferring a hugely increased risk of breast cancer on carrier and both are associated with a large number of DNA repair pathways [120, 121]. Recently, nine mutations (3 nonsense, 5 missense and 1 affecting mRNA splicing) in RECQL gene implicated in double-strand repair are reported in early breast cancer progression [122] which now serves as a marker for its screening. Other severe human genetic disorders associated with DNA damage and repair pathways (especially NER) include XP and CS [119, 123]. It is widely accepted that the genetic basis of important traits in human diseases can be best understood by assessing the association between the occurrence of particular haplotype and particular trait. The most abundant type of variation among haplotypes possessed by individuals in a population is single nucleotide polymorphism (SNP), in which different alleles are present at a given locus. SNPs are invaluable tool for genome mapping whereas the low mutation rate of SNPs makes them excellent markers for studying complex genetic traits and allow comprehensive understanding of genome evolution. Identification and analysis of SNPs that contribute to susceptibility to common diseases can provide highly accurate diagnostic information and facilitate early diagnosis, prevention and treatment of human disease [124-128].

Identification of genetic variants underlying human diseases have already gained a lot of attention in recent decade but still most of inherited risks remains inexplicable. A comprehensive study for these diseases entail genome-wide analysis that completely scrutinize less common alleles in populations with extensive range of ancestry [7]. Thus, to comprehend the influence of genetic aberrations in DNA repair genes complementing to diverse hereditary disorders,

association-based study comprising of contributing variations, SNPs, linkage disequilibrium (LD), haplotype blocks and Tag SNPs was executed. Although analyzing SNPs provide significant insights but methods based on multiple SNPs on same chromosome i.e. haplotypes endows with added power for mapping the complex diseased genes. There is a need to perform detailed analysis on repair pathways and associated human genetic disorders which can provide insights for the prologue diagnosis of human diseases. Recently, role of hydroxymethylation and DNA repair genes has been crucial area under investigation [129] in genome imprinting.

Although, our understanding of DNA repair mechanisms and its implications in major malignancies and multi-system defects has amplified to a great extent but still dealing with harmful consequences of damages is not adequately feasible and also there is no absolute therapy for cancer. There is an imperative need to understand these repair mechanisms and triggered signaling cascades so that damages could be sensed well in advance. Upon sensing the damage, factors involved in the regulation of pathways should be clearly elucidated along with their biological roles and positions where they exert their functions. Thorough understanding of these mechanisms can delineate that how these repair pathways are regulated and can connect the interplay with other cellular processes. Thus, to unravel the complexity hidden in intricate biological repair processes, a systems biology approach is essential for the interaction of genes/proteins/networks for understanding myriad of cellular processes. This integrative approach relies on a group of interacting genes or a whole sum of networks; analyzing which is vital for an extensive perspective on cellular processes and their outcomes which may serve as novel targets for therapeutic interventions.

The genome sequencing and the latest research strategies have generated huge and complex genomic data which is available in repertoire of online resources and databases. Enormous research tools are also available for the extensive analysis of available genomics data. Similarly, efficient detection of vital components drawn in DNA repair requires novel *in silico* approaches and robust computational infrastructure for specialized housing, accessing and analyzing ever-growing and amplifying complex data. In order to address this vast topic, knowledge regarding diverse biological aspects such as DNA damage response, signal transduction, replication, involved components and damage removal needs to be integrated.

Nevertheless, the regulating factors and their biological functions in diverse repair pathways have not been scrutinized in great detail and remains a critical question for future research. The proposed research involves integrated structural, functional and evolutionary analysis of genes and associated proteins involved in DNA repair and linked human genetic disorders, their correlated association with SNPs, and consequently development of a comprehensive human disease-oriented database related to repair pathways. Since, the growth of knowledge on functional analysis of genes and cellular processes demands wide-ranging databases that can represent proteins of interests as well as their functional interactions. An integrated approach of bioinformatics and systems biology has been applied to put into focus gene/protein interactions and molecular pathways to elucidate the required biological information.

Based on the worldwide limitations in current research area, the following objectives have been designed accordingly to fulfill the state-of-the-art requirements and the overall thesis is divided into five chapters which includes the current introduction chapter.

1. Conception of DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems.
2. Systems biology approach for mutational and site-specific structural investigation of DNA repair genes for xeroderma pigmentosum.
3. An integrative approach for mapping differentially expressed genes and network components to elucidate key regulatory genes using novel parameters in colorectal cancer.
4. Decoding the intricate biological pathways in quest of candidate markers implicated in human DNA repair system.

Various bioinformatics and statistical approaches have been used to understand the structural, functional and evolutionary clues latent in biological sequences of interest. A few important genetic disorders were explored at the biological level through computational analyses to provide clues about putative but predictive biomarkers for these diseases. Additionally, top-down approach was applied for the annotation of complex biological networks. The annotation and functional enrichment was made through various pathways and network motifs to generate



biological knowledge. The expectation from the carried research is that it would certainly assist in learning additional factors and regulatory modes concerning DNA repair. Eventually, the exhaustive mechanistic perspective on these DNA repair mechanisms will help in enhanced realization of the disposition to many human hereditary diseases such as cancer and aging. The *in silico* study on DNA repair process and its implicated mechanisms will not only uncover vital clues regarding oncogenesis and age-related diseases but will also provide innovative paradigms for genetic susceptibility, prevention, diagnosis and the rational therapy.

## REFERENCES

- [1] C. Human Genome Sequencing, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, pp. 860-921, 2001.
- [2] C. Human Genome Sequencing, "Finishing the euchromatic sequence of the human genome," *Nature*, vol. 431, pp. 931-945, 2004.
- [3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, and e. al., "The Sequence of the Human Genome," *Science*, vol. 291, pp. 1304-1351, 2001.
- [4] J. D. WATSON and F. H. C. CRICK, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature*, vol. 171, pp. 737 - 738, 1953.
- [5] H. Z. Ring, P. Y. Kwok, and R. G. H. Cotton, "Human Variome Project: an international collaboration to catalogue human genetic variation," *Pharmacogenomics*, vol. 7, pp. 969-972, 2006.
- [6] K. R. Rosenbloom, T. R. Dreszer, J. C. Long, V. S. Malladi, C. A. Sloan, B. J. Raney, M. S. Cline, D. Karolchik, G. P. Barber, H. Clawson, and e. al., "ENCODE whole-genome data in the UCSC Genome Browser: update 2012," *Nucleic Acids Research*, vol. 40, pp. D912-D917, 2012.
- [7] C. The International HapMap, "Integrating common and rare genetic variation in diverse human populations," *Nature*, vol. 467, pp. 52-58, 2010.
- [8] "What is the Human Variome Project?," *Nat Genet*, vol. 39, pp. 423-423, 2007.
- [9] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer, "A vision for the future of genomics research," *Nature*, vol. 422, pp. 835-847, 2003.

- 
- [10] J. H. J. Hoeijmakers, "DNA Damage, Aging, and Cancer," *New England Journal of Medicine*, vol. 361, pp. 1475-1485, 2009.
- [11] A. A. Freitas and J. P. de Magalhães, "A review and appraisal of the DNA damage theory of ageing," *Mutation Research/Reviews in Mutation Research*, vol. 728, pp. 12-22, 2011.
- [12] R. Dulbecco, "EXPERIMENTS ON PHOTOREACTIVATION OF BACTERIOPHAGES INACTIVATED WITH ULTRAVIOLET RADIATION," *Journal of Bacteriology*, vol. 59, pp. 329-347, 1950.
- [13] P. C. Hanawalt, "Four decades of DNA repair: from early insights to current perspectives," *BIOCHIMIE*, vol. 85, pp. 1043-1052, 2003.
- [14] I. Ezkurdia, D. Juan, J. M. Rodriguez, A. Frankish, M. Diekhans, J. Harrow, J. Vazquez, A. Valencia, and M. L. Tress, "Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes," *Human Molecular Genetics*, vol. 23 pp. 5866-5878, 2014.
- [15] T. Lindahl, "Instability and decay of the primary structure of DNA," *Nature*, vol. 362, pp. 709-715, 1993.
- [16] B. B. S. Zhou and S. J. Elledge, "The DNA damage response: putting checkpoints in perspective," *Nature*, vol. 408, pp. 433-439, 2000.
- [17] M. C. Moraes, J. B. Neto, and C. F. Menck, "DNA repair mechanisms protect our genome from carcinogenesis," *Front Biosci*, vol. 17, pp. 1362-1388, 2012.
- [18] J. Knoch, Y. Kamenisch, C. Kubisch, and M. Berneburg, "Rare hereditary diseases with defects in DNA-repair," *European Journal of Dermatology*, vol. 22, pp. 443-455, 2012.
- [19] B. P. Best, "Nuclear DNA damage as a direct cause of aging," *Rejuvenation research*, vol. 12, pp. 199-208, 2009.
- [20] B. Halliwell, "Oxidative stress and cancer: have we moved forward?," *Biochemical Journal*, vol. 401, pp. 1-11, 2007.
- [21] F. Drabløs, E. Feyzi, P. A. Aas, C. B. Vaagbø, B. Kavli, M. S. Bratlie, J. Peña Diaz, M. Otterlei, G. Slupphaug, and H. E. Krokan, "Alkylation damage in DNA and RNA—repair mechanisms and medical significance," *DNA Repair*, vol. 3, pp. 1389-1407, 2004.
- [22] R. De Bont and N. van Larebeke, "Endogenous DNA damage in humans: a review of quantitative data," *Mutagenesis*, vol. 19, pp. 169-185, 2004.

- 
- [23] K. Heil, D. Pearson, and T. Carell, "Chemical investigation of light induced DNA bipyrimidine damage and repair," *Chemical Society Reviews*, vol. 40, pp. 4271-4278, 2011.
- [24] G. Slupphaug, B. Kavli, and H. E. Krokan, "The interacting pathways for prevention and repair of oxidative DNA damage," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 531, pp. 231-251, 2003.
- [25] A. Roulston, R. C. Marcellus, and P. E. Branton, "VIRUSES AND APOPTOSIS," *Annual Review of Microbiology*, vol. 53, pp. 577-628, 1999.
- [26] E. C. Friedberg, L. D. McDaniel, and R. A. Schultz, "The role of endogenous and exogenous DNA damage and mutagenesis," *Current Opinion in Genetics & Development*, vol. 14, pp. 5-10, 2004.
- [27] T. Lindahl, "DNA Repair Enzymes," *Annual Review of Biochemistry*, vol. 51, pp. 61-87, 1982.
- [28] J. M. Furgason and E. M. Bahassi, "Targeting DNA repair mechanisms in cancer," *Pharmacology & Therapeutics*, vol. 137, pp. 298-308, 2013.
- [29] U. Rass, I. Ahel, and S. C. West, "Defective DNA Repair and Neurodegenerative Disease," *Cell*, vol. 130, pp. 991-1004, 2007.
- [30] M. R. Baldwin and P. J. O'Brien, "Defining the functional footprint for recognition and repair of deaminated DNA," *Nucleic Acids Research*, vol. 40, pp. 11638-11647, 2012.
- [31] J. Dabney, M. Meyer, and S. Pääbo, "Ancient DNA damage," *Cold Spring Harbor perspectives in biology*, vol. 5, p. a012567, 2013.
- [32] F. Grosse, G. Krauss, J. W. Knill-Jones, and A. R. Fersht, "Accuracy of DNA polymerase-alpha in copying natural DNA," *The EMBO Journal*, vol. 2, pp. 1515-1519, 1983.
- [33] R. B. Painter, "DNA DAMAGE AND REPAIR IN EUKARYOTIC CELLS," *Genetics*, vol. 78, pp. 139-148, 1974.
- [34] P. Shukla, A. Solanki, K. Ghosh, and B. R. Vundinti, "DNA interstrand cross-link repair: understanding role of Fanconi anemia pathway and therapeutic implications," *European Journal of Haematology*, vol. 91, pp. 381-393, 2013.

- 
- [35] C. Clauson, O. D. Schärer, and L. Niedernhofer, "Advances in understanding the complex mechanisms of DNA interstrand crosslink repair," *Cold Spring Harbor perspectives in medicine*, vol. 3, pp. a012732-a012732, 2013.
- [36] G. A. Cromie, J. C. Connelly, and D. R. F. Leach, "Recombination at Double-Strand Breaks and DNA Ends," *Molecular Cell*, vol. 8, pp. 1163-1174, 2001.
- [37] B. A. Lodish H, Matsudaira P, Kaiser CA, Krieger M, Scott MP, Zipursky SL, Darnell J., in *Molecular Biology of the Cell*, 5th ed New York, NY.: WH Freeman, 2004, p. 963.
- [38] J. H. J. Hoeijmakers, "Genome maintenance mechanisms for preventing cancer," *Nature*, vol. 411, pp. 366-374, 2001.
- [39] C. Yi and C. He, "DNA repair by reversal of DNA damage," *Cold Spring Harbor perspectives in biology*, vol. 5, p. a012575, 2013.
- [40] D. Fu, J. A. Calvo, and L. D. Samson, "Balancing repair and tolerance of DNA damage caused by alkylating agents," *Nat Rev Cancer*, vol. 12, pp. 104-120, 2012.
- [41] A. Hollaender and B. M. Duggar, "The Effects of Sublethal Doses of Monochromatic Ultraviolet Radiation on the Growth Properties of Bacteria," *Journal of Bacteriology*, vol. 36, pp. 17-37, 1938.
- [42] P. Vaughan, T. Lindahl, and B. Sedgwick, "Induction of the adaptive response of *Escherichia coli* to alkylation damage by the environmental mutagen, methyl chloride," *Mutation Research/DNA Repair*, vol. 293, pp. 249-257, 1993.
- [43] L. R. Barrows and P. N. Magee, "Nonenzymatic methylation of DNA by S-adenosylmethionine in vitro," *Carcinogenesis*, vol. 3, pp. 349-351, 1982.
- [44] A. T. Natarajan, J. W. I. M. Simons, E. W. Vogel, and A. A. van Zeeland, "Relationship between cell killing, chromosomal aberrations, sister-chromatid exchanges and point mutations induced by monofunctional alkylating agents in Chinese hamster cells a correlation with different ethylation products in DNA," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 128, pp. 31-40, 1984.
- [45] S. S. Wallace, "Base excision repair: A critical player in many games," *DNA Repair*, vol. 19, pp. 14-26, 2014.
- [46] H. Xie, Y. Gong, J. Dai, X. Wu, and J. Gu, "Genetic variations in base excision repair pathway and risk of bladder cancer: A case-control study in the United States," *Molecular Carcinogenesis*, vol. 54, pp. 50-57, 2015.

- [47] E. C. Friedberg, "How nucleotide excision repair protects against cancer," *Nat Rev Cancer*, vol. 1, pp. 22-33, 2001.
- [48] M. Volker, M. J. Moné, P. Karmakar, A. van Hoffen, W. Schul, W. Vermeulen, J. H. J. Hoeijmakers, R. van Driel, A. A. van Zeeland, and L. H. F. Mullenders, "Sequential Assembly of the Nucleotide Excision Repair Factors In Vivo," *Molecular Cell*, vol. 8, pp. 213-224, 2001.
- [49] K. Sugasawa, Y. Okuda, M. Saijo, R. Nishi, N. Matsuda, G. Chu, T. Mori, S. Iwai, K. Tanaka, K. Tanaka, and F. Hanaoka, "UV-Induced Ubiquitylation of XPC Protein Mediated by UV-DDB-Ubiquitin Ligase Complex," *Cell*, vol. 121, pp. 387-400, 2005.
- [50] J. R. Hwang, V. Moncollin, W. Vermeulen, T. Seroz, H. van Vuuren, J. H. J. Hoeijmakers, and J. M. Egly, "A 3' → 5' XPB Helicase Defect in Repair/Transcription Factor TFIIH of Xeroderma Pigmentosum Group B Affects Both DNA Repair and Transcription," *Journal of Biological Chemistry*, vol. 271, pp. 15898-15904, 1996.
- [51] G. S. Winkler, S. J. Araújo, U. Fiedler, W. Vermeulen, F. Coin, J.-M. Egly, J. H. J. Hoeijmakers, R. D. Wood, H. T. M. Timmers, and G. Weeda, "TFIIH with Inactive XPD Helicase Functions in Transcription Initiation but Is Defective in DNA Repair," *Journal of Biological Chemistry*, vol. 275, pp. 4258-4266, 2000.
- [52] L. Schaeffer, V. Moncollin, R. Roy, A. Staub, M. Mezzina, A. Sarasin, G. Weeda, J. H. Hoeijmakers, and J. M. Egly, "The ERCC2/DNA repair protein is associated with the class II BTF2/TFIIH transcription factor," *The EMBO Journal*, vol. 13, pp. 2388-2392, 1994.
- [53] A. O'Donovan, A. A. Davies, J. G. Moggs, S. C. West, and R. D. Wood, "XPG endonuclease makes the 3' incision in human DNA nucleotide excision repair," *Nature*, vol. 371, pp. 432-435, 1994.
- [54] T. Matsunaga, D. Mu, C. H. Park, J. T. Reardon, and A. Sancar, "Human DNA Repair Excision Nuclease: ANALYSIS OF THE ROLES OF THE SUBUNITS INVOLVED IN DUAL INCISIONS BY USING ANTI-XPG AND ANTI-ERCC1 ANTIBODIES," *Journal of Biological Chemistry*, vol. 270, pp. 20862-20869, 1995.
- [55] K. Sugasawa, "Xeroderma pigmentosum genes: functions inside and outside DNA repair," *Carcinogenesis*, vol. 29, pp. 455-465, 2008.

- 
- [56] W. N. Feller L, Motswaledi MH, Khammissa RA, Meyer M, Lemmer J, "Xeroderma pigmentosum: a case report and review of the literature," *Journal of Preventive Medicine and Hygiene*, vol. 51, pp. 87-91, 2010.
- [57] S. D. Priebe, S. M. Hadi, B. Greenberg, and S. A. Lacks, "Nucleotide sequence of the hexA gene for DNA mismatch repair in *Streptococcus pneumoniae* and homology of hexA to mutS of *Escherichia coli* and *Salmonella typhimurium*," *Journal of Bacteriology*, vol. 170, pp. 190-196, 1988.
- [58] M. Prudhomme, B. Martin, V. Mejean, and J. P. Claverys, "Nucleotide sequence of the *Streptococcus pneumoniae* hexB mismatch repair gene: homology of HexB to MutL of *Salmonella typhimurium* and to PMS1 of *Saccharomyces cerevisiae*," *Journal of Bacteriology*, vol. 171, pp. 5332-5338, 1989.
- [59] Kobayashi K, Matsushima M, Koi S, Saito H, Sagae S, Kudo R, and N. Y, "Mutational analysis of mismatch repair genes, hMLH1 and hMSH2, in sporadic endometrial carcinomas with microsatellite instability," *Japanese Journal of Cancer Research*, vol. 87, pp. 141-145, 1996.
- [60] J. Jiricny, "MutL $\alpha$ : At the Cutting Edge of Mismatch Repair," *Cell*, vol. 126, pp. 239-241, 2006.
- [61] R. R. Iyer, T. J. Pohlhaus, S. Chen, G. L. Hura, L. Dzantiev, L. S. Beese, and P. Modrich, "The MutS $\alpha$ -Proliferating Cell Nuclear Antigen Interaction in Human DNA Mismatch Repair," *The Journal of Biological Chemistry*, vol. 283, pp. 13310-13319, 2008.
- [62] H. Flores-Rozas, D. Clark, and R. D. Kolodner, "Proliferating cell nuclear antigen and Msh2p-Msh6p interact to form an active mismatch recognition complex," *Nat Genet*, vol. 26, pp. 375-378, 2000.
- [63] X. Li and W. D. Heyer, "Homologous recombination in DNA repair and DNA damage tolerance," *Cell research*, vol. 18, pp. 99-113, 2008.
- [64] S. L. Andersen and J. Sekelsky, "Meiotic versus Mitotic Recombination: Two Different Routes for Double-Strand Break Repair: The different functions of meiotic versus mitotic DSB repair are reflected in different pathway usage and different outcomes," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 32, pp. 1058-1066, 2010.

- [65] A. Ratanaphan, "A DNA repair BRCA1 estrogen receptor and targeted therapy in breast cancer," *International journal of molecular sciences*, vol. 13, pp. 14898-14916, 2012.
- [66] M. E. Stauffer and W. J. Chazin, "Structural Mechanisms of DNA Replication, Repair, and Recombination," *Journal of Biological Chemistry*, vol. 279, pp. 30915-30918, 2004.
- [67] W. D. Heyer, X. Li, M. Rolfmeier, and X. P. Zhang, "Rad54: the Swiss Army knife of homologous recombination?," *Nucleic Acids Research*, vol. 34, pp. 4115-4125, 2006.
- [68] R. A. Greenberg, B. Sobhian, S. Pathania, S. B. Cantor, Y. Nakatani, and D. M. Livingston, "Multifactorial contributions to an acute DNA damage response by BRCA1/BARD1-containing complexes," *Genes & development*, vol. 20, pp. 34-46, 2006.
- [69] D. J. R. Ransburgh, N. Chiba, C. Ishioka, A. E. Toland, and J. D. Parvin, "Identification of breast tumor mutations in BRCA1 that abolish its function in homologous DNA recombination," *Cancer research*, vol. 70, pp. 988-995, 2010.
- [70] D. S. Yu, E. Sonoda, S. Takeda, C. L. H. Huang, L. Pellegrini, T. L. Blundell, and A. R. Venkitaraman, "Dynamic Control of Rad51 Recombinase by Self-Association and Interaction with BRCA2," *Molecular Cell*, vol. 12, pp. 1029-1041, 2003.
- [71] M. R. Lieber, Y. Ma, U. Pannicke, and K. Schwarz, "Mechanism and regulation of human non-homologous DNA end-joining," *Nat Rev Mol Cell Biol*, vol. 4, pp. 712-720, 2003.
- [72] M. Takata, M. S. Sasaki, E. Sonoda, C. Morrison, M. Hashimoto, H. Utsumi, Y. Yamaguchi-Iwai, A. Shinohara, and S. Takeda, "Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells," *The EMBO Journal*, vol. 17, pp. 5497-5508, 1998.
- [73] J. R. Walker, R. A. Corpina, and J. Goldberg, "Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair," *Nature*, vol. 412, pp. 607-614, 2001.
- [74] L. G. DeFazio, R. M. Stansel, J. D. Griffith, and G. Chu, "Synapsis of DNA ends by DNA-dependent protein kinase," *The EMBO Journal*, vol. 21, pp. 3192-3200, 2002.
- [75] S. H. Teo and S. P. Jackson, "Lif1p targets the DNA ligase Lig4p to sites of DNA double-strand breaks," *Current Biology*, vol. 10, pp. 165-168, 2000.

- 
- [76] T. E. Wilson, L. M. Topper, and P. L. Palmbo, "Non-homologous end-joining: bacteria join the chromosome breakdance," *Trends in Biochemical Sciences*, vol. 28, pp. 62-66, 2003.
- [77] H. Ohmori, E. C. Friedberg, R. P. P. Fuchs, M. F. Goodman, F. Hanaoka, D. Hinkle, T. A. Kunkel, C. W. Lawrence, Z. Livneh, T. Nohmi, and e. al., "The Y-Family of DNA Polymerases," *Molecular Cell*, vol. 8, pp. 7-8, 2001.
- [78] M. D. Sutton, "COORDINATING DNA POLYMERASE TRAFFIC DURING HIGH AND LOW FIDELITY SYNTHESIS," *Biochimica et biophysica acta*, vol. 1804, pp. 1167-1179, 2009.
- [79] S. D. McCulloch, R. J. Kokoska, C. Masutani, S. Iwai, F. Hanaoka, and T. A. Kunkel, "Preferential cis-syn thymine dimer bypass by DNA polymerase  $\eta$  occurs with biased fidelity," *Nature*, vol. 428, pp. 97-100, 2004.
- [80] D. F. Jarosz, V. G. Godoy, J. C. Delaney, J. M. Essigmann, and G. C. Walker, "A single amino acid governs enhanced activity of DinB DNA polymerases on damaged templates," *Nature*, vol. 439, pp. 225-228, 2006.
- [81] W. L. Neeley, S. Delaney, Y. O. Alekseyev, D. F. Jarosz, J. C. Delaney, G. C. Walker, and J. M. Essigmann, "DNA Polymerase V Allows Bypass of Toxic Guanine Oxidation Products in Vivo," *Journal of Biological Chemistry*, vol. 282, pp. 12741-12748, 2007.
- [82] L. S. Waters, B. K. Minesinger, M. E. Wilttrout, S. D'Souza, R. V. Woodruff, and G. C. Walker, "Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance," *Microbiology and Molecular Biology Reviews : MMBR*, vol. 73, pp. 134-154, 2009.
- [83] A. M. Weber and A. J. Ryan, "ATM and ATR as therapeutic targets in cancer," *Pharmacology & Therapeutics*, vol. 149, pp. 124-138, 2015.
- [84] M. Akbari and H. E. Krokan, "Cytotoxicity and mutagenicity of endogenous DNA base lesions as potential cause of human aging," *Mechanisms of Ageing and Development*, vol. 129, pp. 353-365, 2008.
- [85] S. Negri, V. G. Gorgoulis, and T. D. Halazonetis, "Genomic instability—an evolving hallmark of cancer," *Nature reviews Molecular cell biology*, vol. 11, pp. 220-228, 2010.
- [86] R. L. Rolig and P. J. McKinnon, "Linking DNA damage and neurodegeneration," *Trends in Neurosciences*, vol. 23, pp. 417-424, 2000.



- 
- [87] S. Jalal, J. N. Earley, and J. J. Turchi, "DNA repair: From genome maintenance to biomarker and therapeutic target," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 17, pp. 6973-6984, 2011.
- [88] J. M. Ceruti, M. E. Scassa, J. M. Flo, C. L. Varone, and E. T. Canepa, "Induction of p19INK4d in response to ultraviolet light improves DNA repair and confers resistance to apoptosis in neuroblastoma cells," *Oncogene*, vol. 24, pp. 4065-4080, 2005.
- [89] J. M. D. WHEELER, W. F. BODMER, and N. J. M. MORTENSEN, "DNA mismatch repair genes and colorectal cancer," *Gut*, vol. 47, pp. 148-153, 2000.
- [90] A. R. Lehmann, D. McGibbon, and M. Stefanini, "Xeroderma pigmentosum," *Orphanet Journal of Rare Diseases*, vol. 6, p. 70, 2011.
- [91] K. K. Stefanini M, "Xeroderma pigmentosum," in *Neurocutaneous Diseases*, Ruggieri M, Pascual-Castroviejo I, and D. R. C, Eds., ed, 2008, pp. 771-792.
- [92] A. R. Lehman, S. Kirk Bell, C. F. Arlett, M. C. Paterson, P. H. Lohman, E. A. de Weerd Kastelein, and D. Bootsma, "Xeroderma pigmentosum cells with normal levels of excision repair have a defect in DNA synthesis after UV-irradiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 72, pp. 219-223, 1975.
- [93] P. T. Bradford, A. M. Goldstein, D. Tamura, S. G. Khan, T. Ueda, J. Boyle, K.-S. Oh, K. Imoto, H. Inui, S. I. Moriwaki, and e. al., "CANCER AND NEUROLOGIC DEGENERATION IN XERODERMA PIGMENTOSUM: LONG TERM FOLLOW-UP CHARACTERIZES THE ROLE OF DNA REPAIR," *Journal of medical genetics*, vol. 48, pp. 168-176, 2011.
- [94] Y. Hirai, Y. Kodama, S. I. Moriwaki, A. Noda, H. M. Cullings, D. G. MacPhee, K. Kodama, K. Mabuchi, K. H. Kraemer, C. E. Land, and N. Nakamura, "Heterozygous individuals bearing a founder mutation in the XPA DNA repair gene comprise nearly 1% of the Japanese population," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 601, pp. 171-178, 2006.
- [95] C. Troelstra, W. Heslen, D. Bootsma, and J. H. Hoeijmakers, "Structure and expression of the excision repair gene ERCC6, involved in the human disorder Cockayne's syndrome group B," *Nucleic Acids Research*, vol. 21, pp. 419-426, 1993.

- 
- [96] C. Troelstra, A. van Gool, J. de Wit, W. Vermeulen, D. Bootsma, and J. H. J. Hoeijmakers, "ERCC6, a member of a subfamily of putative helicases, is involved in Cockayne's syndrome and preferential repair of active genes," *Cell*, vol. 71, pp. 939-953, 1992.
- [97] K. A. Henning, L. Li, N. Iyer, L. D. McDaniel, M. S. Reagan, R. Legerski, R. A. Schultz, M. Stefanini, A. R. Lehmann, L. V. Mayne, and E. C. Friedberg, "The Cockayne syndrome group A gene encodes a WD repeat protein that interacts with CSB protein and a subunit of RNA polymerase II TFIIF," *Cell*, vol. 82, pp. 555-564, 1995.
- [98] M. Stefanini, H. Fawcett, E. Botta, T. Nardo, and A. R. Lehmann, "Genetic analysis of twenty-two patients with Cockayne syndrome," *Human genetics*, vol. 97, pp. 418-423, 1996.
- [99] T. Davis, F. S. Wyllie, M. J. Rokicki, M. C. Bagley, and D. Kipling, "The Role of Cellular Senescence in Werner Syndrome," *Annals of the New York Academy of Sciences*, vol. 1100, pp. 455-469, 2007.
- [100] M. D. Gray, J. C. Shen, A. S. Kamath Loeb, A. Blank, B. L. Sopher, G. M. Martin, J. Oshima, and L. A. Loeb, "The Werner syndrome protein is a DNA helicase," *Nature genetics*, vol. 17, pp. 100-103, 1997.
- [101] M. V. Masala, C. Olivieri, C. Pirodda, M. A. Montesu, M. A. Cuccuru, S. Pruneddu, C. Danesino, and D. Cerimele, "Epidemiology and clinical aspects of Werner's syndrome in North Sardinia: description of a cluster," *European Journal of Dermatology*, vol. 17, pp. 213-216, 2007.
- [102] L. Chen, S. Huang, L. Lee, A. Davalos, R. H. Schiestl, J. Campisi, and J. Oshima, "WRN, the protein deficient in Werner syndrome, plays a critical structural role in optimizing DNA repair," *Aging Cell*, vol. 2, pp. 191-199, 2003.
- [103] K. A. Bernstein, S. Gangloff, and R. Rothstein, "The RecQ DNA helicases in DNA Repair," *Annual review of genetics*, vol. 44, pp. 393-417, 2010.
- [104] E. Boder, "Ataxia-telangiectasia: an overview," *Kroc Foundation Series*, vol. 19, pp. 1-63, 1984.
- [105] K. Savitsky, A. Bar-Shira, S. Gilad, G. Rotman, Y. Ziv, L. Vanagaite, D. A. Tagle, S. Smith, T. Uziel, and S. Sfez, "A single ataxia telangiectasia gene with a product similar to PI-3 kinase," *Science*, vol. 268, pp. 1749-1753, 1995.

- 
- [106] A. Reiman, V. Srinivasan, G. Barone, J. I. Last, L. L. Wootton, E. G. Davies, M. M. Verhagen, M. A. Willemsen, C. M. Weemaes, P. J. Byrd, and e. al., "Lymphoid tumours and breast cancer in ataxia telangiectasia; substantial protective effect of residual ATM kinase activity against childhood tumours," *British Journal of Cancer*, vol. 105, pp. 586-591, 2011.
- [107] K. Ushio, "Genetic and familial factors in colorectal cancer," *Japanese journal of clinical oncology*, vol. 15, pp. 281-298, 1985.
- [108] M. C. Lim, S. S. Seo, S. Kang, M. W. Seong, B. Y. Lee, and S. Y. Park, "Hereditary Non-polyposis Colorectal Cancer/Lynch Syndrome in Korean Patients with Endometrial Cancer," *Japanese Journal of Clinical Oncology*, vol. 40, pp. 1121-1127, 2010.
- [109] W. D. Foulkes, "Inherited Susceptibility to Common Cancers," *New England Journal of Medicine*, vol. 359, pp. 2143-2153, 2008.
- [110] H. T. Lynch and A. de la Chapelle, "Hereditary Colorectal Cancer," *New England Journal of Medicine*, vol. 348, pp. 919-932, 2003.
- [111] M. Kloor, L. Staffa, A. Ahadova, and M. von Knebel Doeberitz, "Clinical significance of microsatellite instability in colorectal cancer," *Langenbeck's Archives of Surgery*, vol. 399, pp. 23-31, 2014.
- [112] B. P. Alter, "Diagnosis, Genetics, and Management of Inherited Bone Marrow Failure Syndromes," *ASH Education Program Book*, vol. 2007, pp. 29-39, 2007.
- [113] L. Chang, W. Yuan, H. Zeng, Q. Zhou, W. Wei, J. Zhou, M. Li, X. Wang, M. Xu, F. Yang, and e. al., "Whole exome sequencing reveals concomitant mutations of multiple FA genes in individual Fanconi anemia patients," *BMC Medical Genomics*, vol. 7, pp. 24-24, 2014.
- [114] J. P. de Winter and H. Joenje, "The genetic and molecular basis of Fanconi anemia," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 668, pp. 11-19, 2009.
- [115] J. Soulier, "Fanconi Anemia," *ASH Education Program Book*, vol. 2011, pp. 492-497, 2011.
- [116] J. A. Eisen and P. C. Hanawalt, "A Phylogenomic Study of DNA Repair Genes, Proteins, and Processes," *Mutation research*, vol. 435, pp. 171-213, 1999.

- 
- [117] S. S. Pintus, E. S. Fomin, I. S. Oshurkov, and V. A. Ivanisenko, "Phylogenetic Analysis of the p53 and p63/p73 Gene Families," *In Silico Biology*, vol. 7, pp. 319-332, 2007.
- [118] R. M. A. Costa, V. ChiganÃ§as, R. da Silva Galhardo, H. Carvalho, and C. F. M. Menck, "The eukaryotic nucleotide excision repair pathway," *Biochimie*, vol. 85, pp. 1083-1099, 2003.
- [119] A. R. Lehmann, "DNA repair-deficient diseases, xeroderma pigmentosum, Cockayne syndrome and trichothiodystrophy," *Biochimie*, vol. 85, pp. 1101-1111, 2003.
- [120] P. Karran, J. Offman, and M. Bignami, "Human mismatch repair, drug-induced DNA damage, and secondary cancer," *Biochimie*, vol. 85, pp. 1149-1160, 2003.
- [121] S. P. Lees-Miller and K. Meek, "Repair of DNA double strand breaks by non-homologous end joining," *Biochimie*, vol. 85, pp. 1161-1173, 2003.
- [122] J. Sun, Y. Wang, Y. Xia, Y. Xu, T. Ouyang, J. Li, T. Wang, Z. Fan, T. Fan, B. Lin, H. Lou and Y. Xie, "Mutations in RECQL Gene Are Associated with Predisposition to Breast Cancer," *PLoS Genetics*, vol. 11, pp. e1005228, 2015. [In Press]
- [123] V. A. Bohr, M. Sander, and K. H. Kraemer, "Rare diseases provide rare insights into DNA repair pathways, TFIIH, aging and cancer," *DNA repair*, vol. 4, pp. 293-302, 2005.
- [124] K. Garg, P. Green, and D. A. Nickerson, "Identification of Candidate Coding Region Single Nucleotide Polymorphisms in 165 Human Genes Using Assembled Expressed Sequence Tags," *Genome Research*, vol. 9, pp. 1087-1092, 1999.
- [125] A. C. Syvanen, "Assessing genetic variation: genotyping single nucleotide polymorphisms," *Nat Rev Genet*, vol. 2, pp. 930-942, 2001.
- [126] A. Cavallo and A. C. R. Martin, "Mapping SNPs to protein sequence and structure data," *Bioinformatics*, vol. 21, pp. 1443-1450, 2005.
- [127] L. Wang and Y. Xu, "Haplotype inference by maximum parsimony," *Bioinformatics*, vol. 19, pp. 1773-1780, 2003.
- [128] D. F. Conrad, T. D. Andrews, N. P. Carter, M. E. Hurles, and J. K. Pritchard, "A high-resolution survey of deletion polymorphism in the human genome," *Nat Genet*, vol. 38, pp. 75-81, 2006.
- [129] A. Shukla, M. Sehgal, and T. R. Singh, "Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives," *Gene*, vol. 564, pp. 109-118, 2015.

# CHAPTER-2

Conception of DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems



*“Real strength comes from your ability to walk in your own shoes and not the shoes of others.”  
-Dianne Adams*

## **ABSTRACT**

We present DNA Repair Genetic Association Studies (DR-GAS), a unique, consolidated and comprehensive DNA repair genetic association studies database of human DNA repair system. It provides information on repair genes, assorted mechanisms of DNA repair, linkage disequilibrium (LD), haplotype blocks, non-synonymous single nucleotide polymorphisms (nsSNPs), phosphorylation sites, associated diseases and pathways implicated in repair systems. DNA repair is an intricate process which is critical in maintaining the integrity of genome by eradicating the damaging effect of internal and external changes. Hence, it is crucial to extensively understand the intact process of DNA repair, genes involved, nsSNPs which perhaps affect the function, phosphorylated residues and other related genetic parameters. All the corresponding entries for DNA repair genes, such as proteins, OMIM IDs, literature references and pathways are cross-referenced to their respective primary databases. DNA repair genes and their associated parameters are either represented in tabular or in graphical form through images elucidated by computational and statistical analyses. It is believed that the database will assist molecular biologists, biotechnologists, therapeutic developers and other scientific community to encounter biologically meaningful information, and meticulous contribution of genetic level information towards treacherous diseases in human DNA repair systems. DR-GAS is freely available for academic and research purposes at: <http://www.bioinfoindia.org/drgas>.

## 2.1 INTRODUCTION

DNA repair is a very complex and vital process through which a cell recognizes damage to the DNA caused by endogenous or environmental factors. The cells strive to repair these damages to retain the integrity of its genome. DNA repair is present in both prokaryotes and eukaryotes, whereas in later the genome and repair mechanisms are much more complex [1]. Various factors such as reactive oxygen species [2], replication errors, ultraviolet radiations, X-rays, gamma rays, thermal disruption and viruses involved for incorporating changes or aberrations in DNA can all result in the damaged DNA [3]. There is a high rate of recurrence for endogenous DNA damage as compared to exogenous damage and the type of damages produced due to both factors is roughly indistinguishable [4]. The damage to DNA can further lead to oxidation of bases, generation of DNA strand interruptions, alkylation of bases [5], bulky adduct formation, mismatches and pyrimidine dimers that often trigger viral interactions [6]. The process of eliminating damaged DNA from the genome involves a number of repair proteins like DDB2, MLH1, XPA and different associated mechanisms for repairing diverse type of lesions. The numerous known mechanisms by which the damaged DNA is repaired includes BER, NER, MMR, HRR, NHEJ, DDS and TLS which incorporates different set of genes, enzymes and pathways for repairing DNA. These mechanisms not only maintain the genetic stability but also prevent the genome from carcinogenesis, pre-mature aging and other genetic abnormalities [7-20], among which we are intended to analyze a few of them for better understanding of the entire process.

Innumerable genetic sequence patterns comprising of common haplotype blocks, genetic markers and LD plots are associated with DNA repair related disorders, where some of the DNA repair genes have already been analyzed for its strong association with genetic diseases. The *XRCC1* DNA repair gene is said to be involved in pancreatic cancer and its haplotype analysis showed a strong statistical association with the disorder [21]. However, Saadat et al [22] too demonstrated that Preeclampsia disorder is linked with higher frequency of “194R-399Q” haplotype in *XRCC1* gene with a confidence of 95% as compared to the control. Moreover, variations in *ERCC5* repair gene are reported in gastric cancer which serves as an important marker for the disease [23]. In previous reports, similar studies on *ERCC1* and *ERCC2* DNA repair genes prominently demonstrates the relationship of two genes in lung adenocarcinoma

[24]. Above mentioned information suggests the involvement and association of LD and common haplotype patterns with countless DNA repair disorders [25]. Therefore, there is a need to analyze these indispensable genetic parameters in an efficient way to understand the mechanisms of DNA repair related disorders.

Additionally, the intrinsic properties of many DNA repair proteins are found to be affected by the altered phosphorylation sites, since its state may govern the risk of developing cancers [26-28]. The phosphorylation takes place at serine (S), threonine (T) or tyrosine (Y) residues [29] in the proteins which not only influences the structure but also affects the function, stability, sub-cellular localizations and interaction with other proteins [30, 31]. Few cases of key DNA repair proteins, where nsSNPs have already been associated with cancer risks such as S31R (CDKN1A) in endometrial [32], S326C (OGG1) in esophageal [33, 34] and T241M (XRCC3) in lung [35] and breast cancers [36] due to change in their phosphorylation states. The change may be from phosphorylated to dephosphorylated residue or vice-versa affecting the activity of the repair proteins. The change due to nsSNPs in DNA repair proteins is also found to be defensive against certain disorders, for example T241M mutation in XRCC3 protein is protective against bladder cancer in heavy smokers [37]. Since, the nsSNPs and phosphorylation states play an imperative role in the regulation of multitude of cellular processes, gene expression, signal transduction, apoptosis, homeostasis and DNA damage recognition and its repair [38], it is important to thoroughly analyze these constraints in diverse DNA repair proteins.

In human molecular systems, an extensive vision to thoroughly understand the DNA repair associated diseases is complex and requires high-end computations and resources therefore the main challenge lies in the development of a platform where one could easily access the integrated information for several genetic parameters concerned with DNA repair. Ultimately, there is a prerequisite for such a platform capable of storing huge amalgamation of DNA repair data and implementation of tools for its analysis to comprehensively understand human diseases. Currently, not many resources are available which provide information on DNA repair. Few such resources are REPAIRtoire [39] and Repair funmap [40], while to the best of our knowledge there is no database till date which provides all the information associated with



genetic parameters of human DNA repair system in a comprehensive way. Additionally, none of the available resources provide information concerning the functional association of nsSNPs and their phosphorylation states for human DNA repair system.

Based on the rationale to fill up this research gap for DNA repair systems, first a widespread computational analysis on the genotype data for 215 repair genes was performed and then a database named, DR-GAS (DNA Repair Genetic Association Studies) was compiled for various genetic features, for instance haplotype blocks, LD plots, essential genetic markers and their respective statistical parameters associated with repair genes. This database also includes nsSNPs and their putative functional effect on the genome through their phosphorylation states amongst all DNA repair mechanisms. DR-GAS database is a unique and most comprehensive catalog for DNA repair genes and its associated mechanisms, pathways and diseases. This information could be of utmost use to the researchers involved in the study of human DNA repair system. This database is integrated with a web interface and developed for 215 DNA repair genes/proteins.

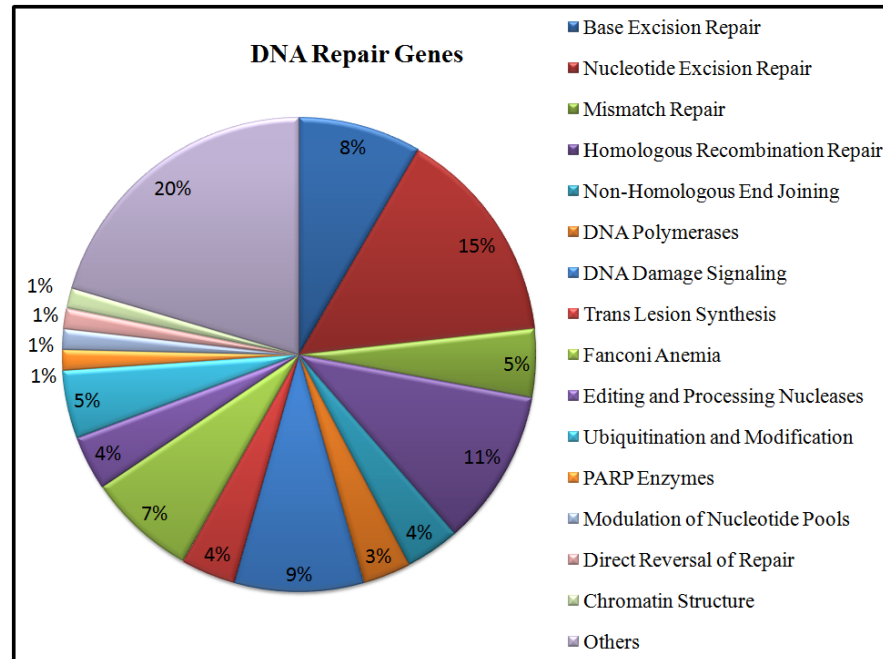
## **2.2 MATERIALS AND METHODS**

In the current study, we applied an integrated approach which is a combination of *in silico* and quantitative genetic studies; performed on 215 DNA repair genes, their proteins, associated pathways and diseases obtained from National Center for Biotechnology Information (NCBI) and other published studies. On the basis of literature and information collected from numerous relevant resources, we categorized the 215 DNA repair genes into 16 major classes as shown in **Figure 2.1**.

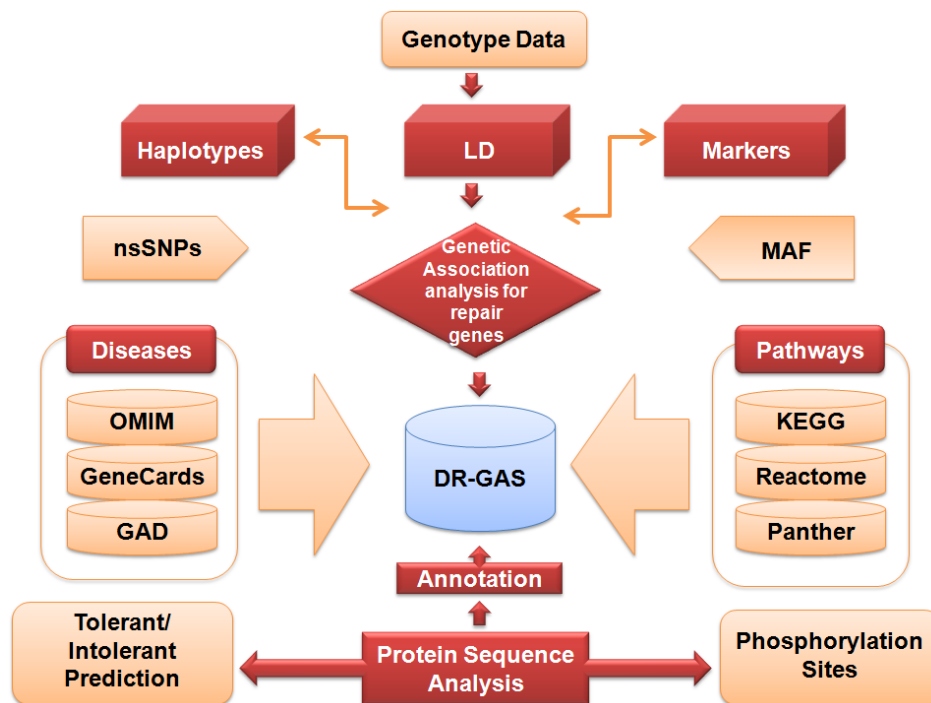
### **2.2.1 Database Design and Content**

To build DR-GAS database, we designed a comprehensive workflow (**Figure 2.2**) which explains the overall process of data collection, computational analysis, database design and implementation. It consists mainly of 4 parts i.e. (i) Collection of genotype data and quantitative genetic studies, (ii) Identification of nsSNPs and their functional effect on human repair system, (iii) Detection of putative phosphorylation sites in DNA repair proteins, (iv) Computational classification of human DNA repair associated diseases and pathways. For creating DR-GAS, the

applied approach puts into focus gene/protein interactions and molecular pathways to elucidate the functional and evolutionary clues latent in biological sequences of interest. The database and the web interface has been created using MySQL and PHP languages respectively.



**Figure 2.1** DNA repair genes and associated mechanisms; the percentage (%) shows the number of genes present in each mechanism.



**Figure 2.2** The pursued pipeline for DNA repair genetic association studies database.

### 2.2.2 Collection of genotype data and quantitative genetic studies

The genotype data for 215 DNA repair genes was retrieved from The International HapMap Project [41], a joint venture among six nations i.e. Canada, China, Japan, Nigeria, the United Kingdom and the United States to help identify genes related to human diseases. The data was analyzed for quantitative genetic parameters such as LD, haplotype blocks and tag SNPs using Haploview, a tool from MIT/Harvard Broad Institute for analyzing and visualizing genetic data [42]. These genetic parameters are vital in carrying out genetic association studies and in examining important variations that lead to diverse human diseases. The haplotypes suggest and reveals combination of alleles at neighboring loci on the chromosome that may be transmitted together and LD provides a measure for the involvement of alleles (genetic markers) in a non-random mode which is either more often or less often than would be expected in a population. These genetic patterns are found to be consistent in many genetic disorders and can predict the predisposition of a person to encompass a particular disease. The main parameters analyzed for genetic association were  $D'$  and  $r^2$ .  $D'$  is the measure of LD between the two blocks which is calculated from the equation:

$$D' = \frac{D}{D_{max}}$$

$$\text{Where, } D = [(F11)(F22) - (F12)(F21)]$$

and ' $D_{max}$ ' depends upon the sign of  $D$ . If  $D$  is positive, then

$$D_{max} = \min [(m1n2) \text{ or } (m2n1)]$$

While if,  $D$  is negative, then

$$D_{max} = \min [(m1n1) \text{ or } (m2n2)]$$

Value of  $D$  in the vicinity of zero provides greater amount of historical recombination between the two blocks. Here,  $m1$  and  $m2$  are the frequencies of alleles at SNP1,  $n1$  and  $n2$  are the frequencies of alleles at SNP2 and  $F11$ ,  $F12$ ,  $F21$ ,  $F22$  are the possible haplotype frequencies.

Another crucial parameter used was  $r^2$  i.e. the correlation coefficient which is calculated by:

$$r = \frac{D}{(m1m2n1n2)^{1/2}}$$

The squared coefficient of correlation ( $r^2$ ) is frequently used to eliminate the arbitrary sign thus introduced in the correlation value.

### **2.2.3 Identification of nsSNPs and their functional effect on human repair system**

The nsSNPs encoding different amino acids in protein sequences of 215 DNA repair genes was analyzed from single nucleotide polymorphism database (dbSNP) [43] and other SNP based databases [44]. Investigating the effect of SNPs in coding sequences is very complex and expensive through experimental methods, consequently the genetic mutations in genes have been linked to deviations in the phenotypes using a bioinformatics algorithm i.e. sorting intolerant from tolerant (SIFT) [45]. This algorithm focuses on the genetic variants that may affect the phenotypic characteristics. SIFT prediction tool has been used for the protein conservation analysis which is based on the principle that protein evolution has a strong correlation with protein function. The highly conserved positions suffer from fewer substitutions whereas the ones weakly conserved, can tolerate more substitutions. If the SIFT score is less than the threshold value, the substitution is said to have an effect on the protein otherwise no major changes in the protein are confirmed.

### **2.2.4 Detection of putative phosphorylation sites in DNA repair proteins**

The key amino acids within the repair proteins could be analyzed by identifying their phosphorylation sites and such predictions could provide insights into the biochemical actions of the analyzed proteins. For the prediction of potential phosphorylation sites at S, T, and Y residues in the 215 repair protein sequences, NetPhos [46] tool was used. It is based on artificial neural networks which have been extensively trained from various samples. For the training of NetPhos program, PhosphoBase [47], a database of comprehensive phosphorylated proteins is used for pattern recognition. The analysis of phosphorylation state of repair proteins is a significant phase of this study as many cellular processes and DNA damage recognition and its repair mechanism are affected by it.

### **2.2.5 Computational classification of human DNA repair associated diseases and pathways**

The information for a variety of diseases related to DNA repair such as multiple kind of cancers, XP, CS and other genetic aberrations were collected from the literature and databases like Online Mendelian Inheritance In Man (OMIM) [48], Genetic Association Database (GAD) [49] and Gene Cards [50]. The information thus obtained was integrated and represented in **Table 2.1** along with the number of predicted nsSNPs in each mechanism.

**Table 2.1 No. of predicted nsSNPs that alter protein sequence and associated diseases with various DNA repair mechanisms**

Mechanisms	No. of Genes	Gene Names	No. of Predicted Damaging nsSNPs	Associated Major Disease
<b>Base Excision Repair</b>	18	MBD4, MPG, MUTYH, NEIL1, NEIL2, NEIL3, NTHL1, OGG1, APEX1, APEX2, LIG3, PNKP, XRCC1, SMUG1, TDG, UNG, APLF, HUS1	40	Rett Syndrome, Angelman Syndrome, Nsclc, Lynch syndrome, XP, Parkinson disease, Alzheimers disease, Diphtheria, Bloom syndrome, Ataxia telangiectasia and various kinds of cancers.
<b>Nucleotide Excision Repair</b>	32	CCNH, CDK7, CETN2, DDB1, DDB2, ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, ERCC6, ERCC8, GTF2H1, GTF2H2, GTF2H3, GTF2H4, GTF2H5, LIG1, MMS19L, MNAT1, RAD23A, RAD23B, RPA1, RPA2, RPA3, TFIIH, XAB2, XPA, XPC, UVSSA, RFC1, CUL4A	122	Retinoblastoma, Leukemia, Lymphoma, XP, Alzheimers disease, Nsclc, Cockayne syndrome, Trichothiodystrophy, Glioma, Necrosis, Fanconianemia, Ataxia telangiectasia and various kinds of cancers.
<b>Mismatch Repair</b>	10	MLH1, MLH3, MSH2, MSH3, MSH4, MSH5, MSH6, PMS1, PMS2, PMS2P3	246	Lynch syndrome, HNPCC, Muir-torre syndrome, Turcot syndrome, Werner syndrome, Hnscc, Nsclc, Peutz-jeghers syndrome and various kinds of cancers.
<b>Non-Homologous End Joining</b>	8	DCLRE1C, LIG4, NHEJ1, PRKDC, XRCC4, XRCC5, XRCC6, WRN	48	Omenn syndrome, Ataxia telangiectasia, Nijmegen breakage syndrome, Lig4 syndrome, Leukemia, Bloom syndrome and various kinds of cancers.
<b>Homologous Recombination Repair</b>	23	BRCA1, DMC1, EME1, EME2, GIYD1, GIYD2, MRE11A, MUS81, NBN, RAD50, RAD51, RAD51L1, RAD51L3, RAD52, RAD54B, RAD54L, RBBP8, SHFM1, XRCC2, XRCC3, BLM, RAD51B, RAD51C	74	Fanconi anemia, Ataxia telangiectasia, Cowden disease, Nijmegen breakage syndrome, Bloom syndrome, HNPCC, Werner syndrome, Canavan disease, XP, Lynch syndrome, Sickle cell disease, Hodgkin disease, Leukemia, Anemia, Hnscc, Nsclc and various kinds of cancers.
<b>DNA Damage Signaling</b>	19	ATM, ATR, ATRIP, CDKN1A, CHEK1, CHEK2, DCLRE1A, DCLRE1B, GPS1, MDC1, RAD1, RAD9A, RAD17, RFC2, RFC3, RFC4, RFC5, TOPBP1, TP53	92	Ataxia telangiectasia, Osteoporosis, Atherosclerosis, Necrosis, Nsclc, Hnscc, XP, Cardiovascular diseases, Cockayne syndrome, Alzheimers disease, Malaria, Pituitary diseases, Burkitt lymphoma, Mental retardation, Thalassemia, Epilepsy, Gaucher disease, Aids and Liver cirrhosis.
<b>Trans Lesion Synthesis</b>	8	POLM, POLN, POLQ, REV1L, POLH, POLI, POLK, REV3L	76	Mitochondrial diseases, Werner syndrome, Glioma, Alpers syndrome, Epilepsy, Neurodegenerative diseases, Liver diseases, Parkinson disease, XP and various kinds of cancers.

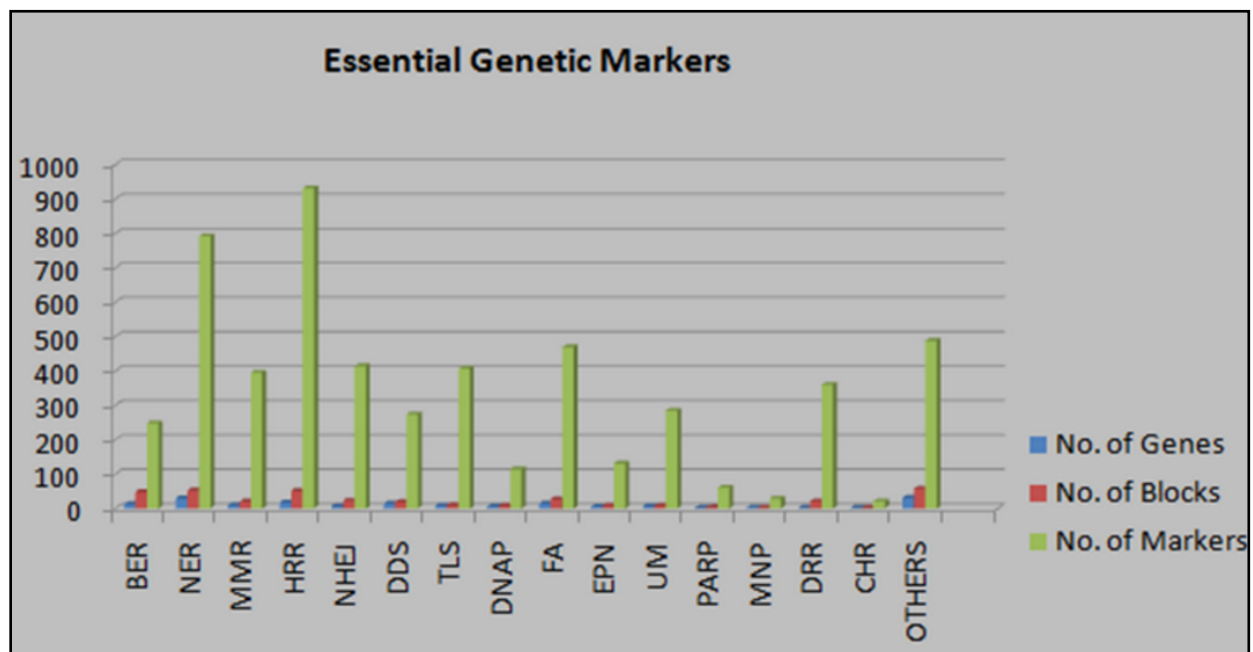
<b>Fanconi Anemia</b>	16	BRCA2, BRIP1, BTBD12, FAAP24, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCL, FANCM, FANCI, PALB2, C1ORF86	762	Fanconi anemia, Ataxia telangiectasia, Bloom syndrome, Lynch syndrome, HNPCC, Tay-sachs disease, Necrosis, Hodgkin disease, Hnscc, Nijmegen breakage syndrome, Alzheimers disease and various kinds of cancers.
<b>Editing and Processing Nucleases</b>	8	APTX, EXO1, FEN1, MTMR15, SPO11, TREX1, TREX2, TTRAP	14	Ataxia telangiectasia, HNPCC, Neurodegenerative diseases, Werner syndrome, Parkinson's disease and various kinds of cancers.
<b>Ubiquitination and Modification</b>	10	HLTF, RAD18, RNF4, RNF8, RNF168, SHPRH, UBE2A, UBE2B, UBE2N, UBE2V2	9	Colorectal cancer, Colon cancer, Gastric cancer, Adenoma, Papilloma.
<b>PARP Enzymes</b>	3	PARP1, PARP2, PARP3	14	Ataxia telangiectasia, Werner syndrome, Fanconi anemia, Cockayne syndrome, Parkinson disease, Diabetes mellitus, Asthma and various kinds of cancers.
<b>Modulation of Nucleotide Pools</b>	3	DUT, NUDT1, RRM2B	-	Herpes simplex, Hiv infections, Parkinson disease, Dysplasia and various kinds of cancers.
<b>Direct Reversal of Damage</b>	3	ALKBH2, ALKBH3, MGMT	10	Glioblastoma, Nsclc, Stable disease, XP, Viral infection and various kinds of cancers.
<b>Chromatin Structure</b>	3	CHAF1A, H2AFX, SETMAR	-	Ataxia telangiectasia, Nijmegen breakage syndrome, Bloom syndrome, Fanconi anemia, Necrosis, Leukemia and various kinds of cancers.
<b>DNA Polymerases</b>	7	PCNA, POLB, POLD1, POLE, POLG, MAD2L2, POLL	24	Werner syndrome, Aids, Glioma, Alpers syndrome, Mitochondrial diseases, Epilepsy, Neurodegenerative diseases, Liver diseases, Parkinson disease, XP, Burkitt lymphoma and various kinds of cancers.
<b>Others*</b>	44	TDP1, ABL1, ADA, ATRX, BARD1, TTDN1, CIB1, CLK2, COPS2, CRY1, GADD45A, GADD45G, HEL308, LHX3, OBFC2B, PER1, POLD3, POLDIP2, POLDIP3, POLE2, POLE3, POLE4, POLR2E, POLR2F, POLR2H, POLR2K, POLR2L, PRPF19, RAD21, RDM1, RECQL, RECQL4, RECQL5, RPA4, SIRT1, SMC1A, SUMO1, TP53BP1, UNG2, GEN1, PMS6, XSE6, YKU80P, ZFP276	62	Fanconi anemia, Werner syndrome, Bloom syndrome, Progeria, Nijmegen breakage syndrome, Cataract, Ataxia telangiectasia, Osteoporosis, Atherosclerosis, Necrosis, Nsclc, Hnscc, XP, Cardiovascular diseases, Cockayne syndrome, Alzheimers disease, Malaria, Kallmann syndrome, Pituitary diseases, Burkitt lymphoma, Mental retardation, Thalassemia, Epilepsy, Gaucher disease, Aids, Liver cirrhosis, Brain tumors, Rheumatoid Thyroid cancer, Hematologic disorders and various kinds of cancers.

\* This category includes all the DNA repair genes which have not been classified on the basis of mechanism in which they are involved.

The pathways wherein there is association of several DNA repair genes were collected and analyzed from GenBank annotations and variety of additional resources like Kyoto Encyclopedia of Genes and Genomes (KEGG) [51], Gene Cards and REACTOME [52] to verify their respective entities. This process applied manual curation of data to remove any likelihood of redundancy in the ultimate collected information.

### 2.3 RESULTS AND DISCUSSION

In this study, 215 DNA repair genes are categorized on the basis of mechanisms in which they are involved. DR-GAS database provides an easy and effective way for the search and retrieval of genetic essential markers (**Figure 2.3**) and associated information for repair genes. We have collected the LD, haplotype, markers, nsSNPs, pathways and disease related information for 215 repair genes which have been classified into main pathways such as BER, NER, MMR, HRR, NHEJ, DDS, TLS, DNA Polymerases, Ubiquitination and Modification, Fanconi anemia, Editing and Processing Nucleases. A category referred as ‘Others’ has been created for the genes which are reported as DNA repair genes but are not being classified in any of the main pathways.



**Figure 2.3** Genetic association details for 16 DNA repair mechanisms; the statistical distribution of number of genes in different mechanisms, important haplotype blocks and genetic markers associated with these genes.

### **2.3.1 The Web Interface**

The database is integrated with an inferable web interface which offers the facility to user for browsing the repository using seven different types of search options. One can search for mechanism, nsSNPs, haplotypes, LD, genetic markers, diseases and phosphorylation sites. It also provides an option for the advanced search for user's convenience where the hybrid data is provided for few categories.

#### ***Search for mechanism***

In the mechanism menu, user can search for any DNA repair mechanism and retrieve information regarding all genes involved in the browsed mechanism, their Gene ID's from GenBank database, OMIM ID's, implicated pathways, associated diseases and the appropriate literature references corresponding to the disorders which are linked to NCBI, OMIM, KEGG and PUBMED databases respectively.

#### ***Search for nsSNPs***

The nsSNPs in DNA repair genes are easily accessible through nsSNPs search option in DR-GAS. The nsSNP when introduced in the gene sequence can result in a change of encoded amino acid that could, in theory, disrupt function to promote inefficient DNA repair. By clicking on Get nsSNPs button, various nsSNPs for the selected gene (Gene ID) are displayed. SNP ID's are linked to dbSNP, amino acid change column gives the position of the change in the sequence and the amino acid being replaced, while the prediction column shows whether the change is damaging or not.

#### ***Search for haplotypes***

In the haplotypes search option, users can acquire exhaustive information regarding the critical haplotypes (combination of alleles at adjacent locations (loci) on the chromosome that are mostly transmitted together). Haplotypes are the blocks of associated SNPs which are conserved throughout the genome in form of patterns called "haplotype blocks". One can also explore parameters like block which gives the current number of blocks in a particular query gene, number of markers column gives the amount of markers present in a block. On clicking the block option, an easy and inferable view of the haplotypes in specified gene is generated where the



marker numbers are depicted on the top and the tag SNPs (if any) are highlighted with a triangular pointer. These blocks correspond to the set of consecutive sites which either has small or no indication of historical recombination. Population frequencies of each haplotype are shown and common crossings from one block to the next are represented by lines, where thicker lines portray more common crossings than the thinner ones. The blocks present the correlation of various residue states among the polymorphic sites across the genome. The multilocus  $D'$  value ( $D'$ ) is also being specified at the bottom of image.

### ***Search for LD***

The LD information of 215 repair genes is analyzed and congregated in the LD search option. Here, we have considered only those loci's whose  $r^2$  value i.e. the correlation coefficient between the two loci is  $\geq 0.6$  as these are the most substantial ones. It includes loci 1 and loci 2, which are the two loci under study,  $D'$  value between the two loci, log of likelihood odds ratio (LOD) i.e. a measure of confidence in the value of  $D'$ , correlation coefficient value between loci 1 and loci 2, CI-low and CI-hi column represents 95% confidence lower bound and upper bound on  $D'$ , distance (in bases) between the loci, LD image column gives an interactive image of the LD plot thus generated. In the markers search option, significant markers identified in study have been compiled and incorporated. It provides a facility to the user to access the marker specific parameters like Marker ID, fully genotyped family trios for the marker (0 for datasets with unrelated individuals), the marker's observed heterozygosity, predicted heterozygosity of the marker calculated from:

$$[2 * MAF * (1 - MAF)]$$

where, MAF is Minor Allele Frequency, Hardy-Weinberg (H-W) equilibrium  $p$  value, i.e. the probability that its deviation from H-W equilibrium could be explained by chance, the percentage of non-missing genotypes for the marker and MAF for the given marker.

### ***Search for Disease***

Additionally, moving further to the database options, there is a disease menu, which comprises of numerous diseases that have been reported because of mutations or any additional aberrations in DNA repair genes. The disease information includes details of the associated DNA repair mechanism, OMIM ID's and the pathways involved. Currently, the database supports browsing

for 36 DNA repair associated diseases which either directly or indirectly are concerned with aberrations or abnormalities in DNA repair system.

### ***Search for Phosphorylation Sites***

In the phosphorylation search option, user can acquire information on the phosphorylation sites in DNA repair proteins. The major information includes the protein ID of the repair protein sequence, position of the phosphorylation site in sequence, 9 character sequence representing the phosphorylated residue at exact center of the sequence, prediction scores above 0.5 has been chosen as a criteria for the selection of potential phosphorylation sites and the prediction column gives the putative phosphorylated residues (S or T or Y). The experimental validation for many phosphorylation sites in the DNA repair proteins has been made from diverse research articles and intense literature survey. Total number of phosphorylated residues (S, T, Y) for each mechanism, and PubMed IDs for few of these verified research articles are mentioned in **Table 2.2**.

These above mentioned search criterions can be easily and freely accessed at the homepage for DR-GAS database (illustrated in **Figure 2.4**.) at the following web address <http://www.bioinfoindia.org/drgas>. For effortless understanding of the results and their interpretations, following examples have been taken from the repository and illustrated in **Figure 2.5**. Here, “Translesion Synthesis” has been chosen from the mechanism menu as a testing mechanism, ‘675’ as Gene ID of *BRCA2* “GenBank ID: 675” for nsSNPs and markers search boxes, “*BRCA2*” as gene name for phosphorylation site prediction, “Gastric cancer” as disease name from disease menu, ‘7515’ as Gene ID of *XRCC1* “GenBank ID: 7515” for haplotype search option and ‘5985’ as Gene ID of *RFC5* “GenBank ID: 5985” for LD block generations. DR-GAS is first of its kind model where the users could easily retrieve and explore the quantitative genetic parameters and the phosphorylation states of DNA repair genes altogether along with various diseases, repair mechanism and published literature references for 215 genes of human DNA repair system.

**Table 2.2 Identification and literature verification of phosphorylation sites in DNA repair proteins**

Mechanisms	No. of Proteins	Predicted Phosphorylated Residues			Experimental Validations (PubMed IDs)*
		S	T	Y	
Base Excision Repair	18	295	98	57	18971944, 15073047, 18669648
Nucleotide Excision Repair	32	729	285	165	12140753, 17081983, 18669648, 16964243
Mismatch Repair	10	382	132	70	17525332, 18669648, 16964243
Non-Homologous End Joining	8	298	99	74	14599745, 18669648, 16097034
Homologous Recombination Repair	23	639	189	72	14701743, 22084686, 14749735
DNA Damage Signaling	19	705	269	156	22084686, 18971944, 8084608
Trans Lesion Synthesis	8	591	143	84	18669648
Fanconi Anemia	16	905	273	132	12815053, 18669648, 11855836, 17525332
Editing and Processing Nucleases	8	157	50	23	-
Ubiquitination and Modification	10	227	70	33	21150323, 21098111
PARP Enzymes	3	63	26	21	-
Modulation of Nucleotide Pools	3	28	7	7	8389461, 18669648
Direct Reversal of Damage	3	22	8	7	-
Chromatin Structure	3	58	28	16	15302935, 17525332
DNA Polymerases	7	158	72	65	18669648
Others <sup>†</sup>	44	1062	333	188	17081983, 18669648, 15298678, 17525332

\* The PubMed IDs for some of the research articles showing experimental validations for the occurrence of important phosphorylation sites in DNA repair proteins have been stated in this column.

<sup>†</sup> The category includes all the DNA repair genes which have not been classified on the basis of mechanism in which they are involved.

**DR-GAS**  
DNA Repair database for Genetic Association Studies

Home Overview Statistics Help Credits and Citations Contact Us

Search

DRGAS is a DNA Repair database of genetic association studies for human DNA repair systems. It includes the information for all the genetic parameters, pathways and disorders related to DNA repair genes. It provides you the facility to search by browsing through different mechanisms, diseases, genes, etc.

**Search By Mechanism:**  
Base Excision Repair (BER) Get By Mechanism

**Search for Non Synonymous SNPs:**  
Gene ID Get nsSNPs Clear

**Search for Haplotypes:**  
Gene ID Get Haplotype Clear

**Search for Linkage Disequilibrium:**  
Gene ID Get LD Clear

**Search for Genetic Marker(s):**  
Gene ID Get Marker Clear

**Search By Disease:**  
Xeroderma Pigmentosum Get By Disease

**Search for Phosphorylation Sites:**  
Gene Name Get Phos Sites Clear

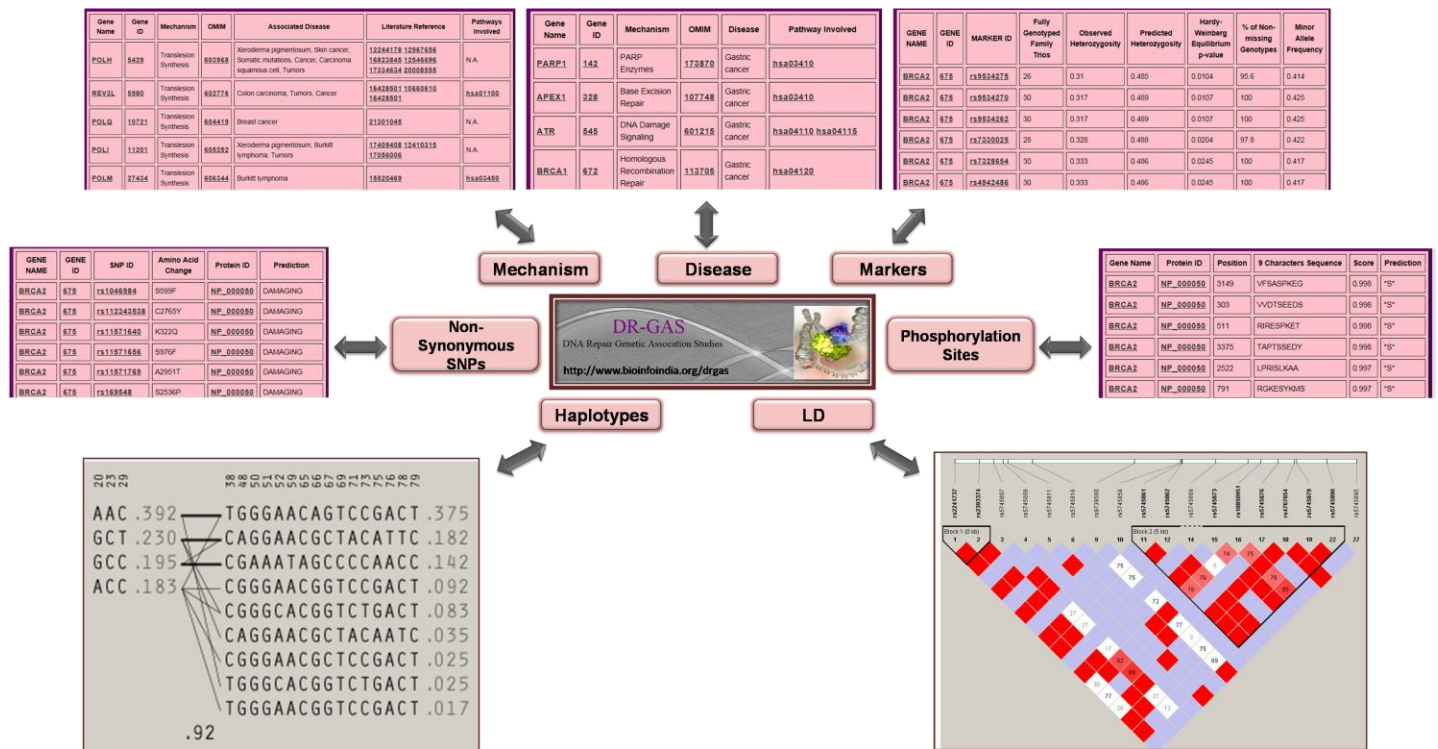
**Advanced Options**

To cite DRGAS: Manika Sehgal and Tiratha Raj Singh (2014) DR-GAS: A database of functional genetic variants and their phosphorylation states in human DNA repair systems. *DNA Repair*, 16: 97-103.

[Go to Top of the Page](#)

001617  
Visitor Number  
Copyright 2013 bioinfoindia.org | Design by Manika Sehgal

**Figure 2.4** The homepage for DNA Repair Genetic Association Studies (DR-GAS) database.



**Figure 2.5** Demonstration and implementation of DR-GAS with various available search and advanced options. The illustrated output from the repository is represented in a combined image for all the generated results.

## 2.4 CONCLUSION

DR-GAS is an extensive compendium for DNA repair genes comprising of quantitative genetic parameters such as LD, haplotypes, SNPs, disease allied information and their phosphorylation states. This repository is unique and first of its kind since there is no such resource available till date for DNA repair system which provides appropriate classification of DNA repair genes in associated mechanisms as presented by DR-GAS (i.e. 16 major classes along with quantitative genetic parameters and phosphorylation sites). This database will assist researchers to study the repair genes in depth and will provide useful insight for future analysis and studies. This repository will also help for easy understanding and investigation of many DNA repair related disorders and moreover will provide useful genes and proteins related information. This database will be of utmost use to the scientists focused in developing therapeutic targets for precarious diseases like multiple forms of cancers, skin diseases and neurodegenerative disorders through the genetic factor's information, which is the basic foundation for the analysis and treatment of

diseases. It is anticipated that this web based comprehensive resource would serve as a valuable accompaniment for analyzing DNA repair systems for human and will also contribute scientific knowledge towards better understanding of other mammalian repair systems. DR-GAS will be updated on regular basis and is believed that this resource will not only assist molecular biologists but also therapeutic developers and other scientific community to encounter biologically meaningful information.

## REFERENCES

- [1] G. A. Cromie, J. C. Connelly, and D. R. F. Leach, "Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans," *Molecular Cell*, vol. 8, pp. 1163-1174, 2001.
- [2] G. Slupphaug, B. Kavli, and H. E. Krokan, "The interacting pathways for prevention and repair of oxidative DNA damage," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 531, pp. 231-251, 2003.
- [3] A. Roulston, R. C. Marcellus, and P. E. Branton, "VIRUSES AND APOPTOSIS," *Annual Review of Microbiology*, vol. 53, pp. 577-628, 1999.
- [4] A. L. Jackson and L. A. Loeb, "The contribution of endogenous sources of DNA damage to the multiple mutations in cancer," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 477, pp. 7-21, 2001.
- [5] T. Lindahl, "DNA Repair Enzymes," *Annual Review of Biochemistry*, vol. 51, pp. 61-87, 1982.
- [6] C. E. Lilley, R. A. Schwartz, and M. D. Weitzman, "Using or abusing: viruses and the cellular DNA damage response," *Trends in Microbiology*, vol. 15, pp. 119-126, 2007.
- [7] M. C. Moraes, J. B. Neto, and C. F. Menck, "DNA repair mechanisms protect our genome from carcinogenesis," *Front Biosci*, vol. 17, pp. 1362-1388, 2012.
- [8] J. Knoch, Y. Kamenisch, C. Kubisch, and M. Berneburg, "Rare hereditary diseases with defects in DNA-repair," *European Journal of Dermatology*, vol. 22, pp. 443-455, 2012.
- [9] B. P. Best, "Nuclear DNA damage as a direct cause of aging," *Rejuvenation research*, vol. 12, pp. 199-208, 2009.
- [10] T. L. Timme and R. E. Moses, "Diseases with DMA Damage-Processing Defects," *The American journal of the medical sciences*, vol. 295, pp. 40-48, 1988.

- 
- [11] J. H. J. Hoeijmakers, "DNA damage, aging, and cancer," *New England Journal of Medicine*, vol. 361, pp. 1475-1485, 2009.
- [12] S. Hassen, N. Ali, and P. Chowdhury, "Molecular signaling mechanisms of apoptosis in hereditary non-polyposis colorectal cancer," *World journal of gastrointestinal pathophysiology*, vol. 3, pp. 71-79, 2012.
- [13] M. Sehgal and T. R. Singh, "Identification and analysis of biomarkers for mismatch repair proteins: A bioinformatic approach," *Journal of natural science, biology, and medicine*, vol. 3, p. 139, 2012.
- [14] D. Tamura, S. G. Khan, M. Merideth, J. J. DiGiovanna, M. A. Tucker, A. M. Goldstein, K.-S. Oh, T. Ueda, J. Boyle, and M. Sarihan, "Effect of mutations in XPD (ERCC2) on pregnancy and prenatal development in mothers of patients with trichothiodystrophy or xeroderma pigmentosum," *European Journal of Human Genetics*, vol. 20, pp. 1308-1310, 2012.
- [15] G. M. M. J. Oshima, F.M. Hisama, *Werner Syndrome*. University of Washington, Seattle: GeneReviews™, 2002.
- [16] A. N. Suhasini and R. M. Brosh Jr, "Fanconi anemia and Bloom's syndrome crosstalk through FANCD1/BLM helicase interaction," *Trends in Genetics*, vol. 28, pp. 7-13, 2012.
- [17] K. Savitsky, A. Bar-Shira, S. Gilad, G. Rotman, Y. Ziv, L. Vanagaite, D. A. Tagle, S. Smith, T. Uziel, and S. Sfez, "A single ataxia telangiectasia gene with a product similar to PI-3 kinase," *Science*, vol. 268, pp. 1749-1753, 1995.
- [18] C. A. Strathdee and M. Buchwald, "Molecular and cellular biology of Fanconi anemia," *Journal of Pediatric Hematology/Oncology*, vol. 14, pp. 177-185, 1992.
- [19] I. Kamileri, I. Karakasilioti, A. Sideri, T. Kosteas, A. Tatarakis, I. Talianidis, and G. A. Garinis, "Defective transcription initiation causes postnatal growth failure in a mouse model of nucleotide excision repair (NER) progeria," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 2995-3000, 2012.
- [20] H. Vogel, D.-S. Lim, G. Karsenty, M. Finegold, and P. Hasty, "Deletion of Ku86 causes early onset of senescence in mice," *Proceedings of the National Academy of Sciences*, vol. 96, pp. 10770-10775, 1999.

- 
- [21] M. Nakao, S. Hosono, H. Ito, M. Watanabe, N. Mizuno, S. Sato, Y. Yatabe, K. Yamao, R. Ueda, and K. Tajima, "Selected polymorphisms of base excision repair genes and pancreatic cancer risk in Japanese," *Journal of Epidemiology*, vol. 22, pp. 477-483, 2012.
- [22] I. Saadat, Z. Beyzaei, F. Aghaei, S. Kamrani, and M. Saadat, "Association between polymorphisms in DNA repair genes (XRCC1 and XRCC7) and risk of preeclampsia," *Archives of gynecology and obstetrics*, vol. 286, pp. 1459-1462, 2012.
- [23] Z. Duan, C. He, Y. Gong, P. Li, Q. Xu, L.-p. Sun, Z. Wang, C. Xing, and Y. Yuan, "Promoter polymorphisms in DNA repair gene ERCC5 and susceptibility to gastric cancer in Chinese," *Gene*, vol. 511, pp. 274-279, 2012.
- [24] J. Yin, U. Vogel, Y. Ma, R. Qi, H. Wang, L. Yue, D. Liang, C. Wang, X. Li, and T. Song, "HapMap-based study of a region encompassing *ERCC1* and *ERCC2* related to lung cancer susceptibility in a Chinese population," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 713, pp. 1-7, 2011.
- [25] P. Van Eerdewegh, R. D. Little, J. e. Dupuis, R. G. Del Mastro, K. Falls, J. Simon, D. Torrey, S. Pandit, J. McKenny, and K. Braunschweiger, "Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness," *Nature*, vol. 418, pp. 426-430, 2002.
- [26] A. Sarasin, "An overview of the mechanisms of mutagenesis and carcinogenesis," *Mutation Research/Reviews in Mutation Research*, vol. 544, pp. 99-106, 2003.
- [27] J. Thacker and M. g. Z. Zdzienicka, "The mammalian *XRCC* genes: their roles in DNA repair and genetic stability," *DNA repair*, vol. 2, pp. 655-672, 2003.
- [28] N. Motoyama and K. Naka, "DNA damage tumor suppressor genes and genomic instability," *Current opinion in genetics & development*, vol. 14, pp. 11-16, 2004.
- [29] L. N. Johnson and M. O'Reilly, "Control by phosphorylation," *Current opinion in structural biology*, vol. 6, pp. 762-769, 1996.
- [30] P. Cohen, "The regulation of protein function by multisite phosphorylation—a 25 year update," *Trends in biochemical sciences*, vol. 25, pp. 596-601, 2000.
- [31] T. Pawson, "Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems," *Cell*, vol. 116, pp. 191-203, 2004.

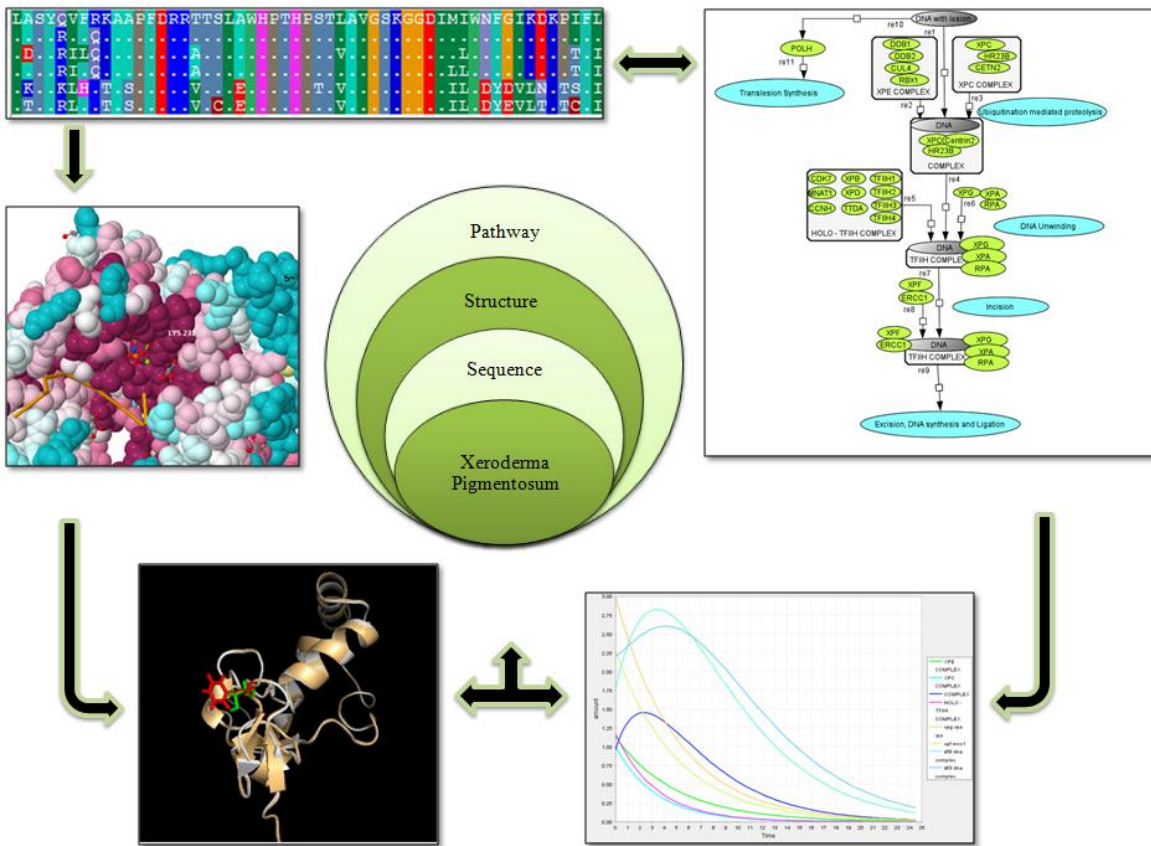


- [32] J. W. Roh, J. W. Kim, N. H. Park, Y. S. Song, I. Park, S.-Y. Park, S. B. Kang, and H. P. Lee, "*p53* and *p21* genetic polymorphisms and susceptibility to endometrial cancer," *Gynecologic oncology*, vol. 93, pp. 499-505, 2004.
- [33] M.-T. Wu, D.-C. Wu, H.-K. Hsu, E.-L. Kao, C.-H. Yang, and J.-M. Lee, "Association between p21 codon 31 polymorphism and esophageal cancer risk in a Taiwanese population," *Cancer Letters*, vol. 201, pp. 175-180, 2003.
- [34] D. Y. Xing, W. Tan, N. Song, and D. X. Lin, "Ser326Cys polymorphism in hOGG1 gene and risk of esophageal cancer in a Chinese population," *International journal of cancer*, vol. 95, pp. 140-143, 2001.
- [35] H. Sugimura, T. Kohno, K. Wakai, K. Nagura, K. Genka, H. Igarashi, B. J. Morris, S. Baba, Y. Ohno, and C. Gao, "hOGG1 Ser326Cys polymorphism and lung cancer susceptibility," *Cancer Epidemiology Biomarkers & Prevention*, vol. 8, pp. 669-674, 1999.
- [36] J. C. Figueiredo, J. A. Knight, L. Briollais, I. L. Andrulis, and H. Ozelik, "Polymorphisms XRCC1-R399Q and XRCC3-T241M and the risk of breast cancer at the Ontario site of the Breast Cancer Family Registry," *Cancer Epidemiology Biomarkers & Prevention*, vol. 13, pp. 583-591, 2004.
- [37] M. Shen, R. J. Hung, P. Brennan, C. Malaveille, F. Donato, D. Placidi, A. Carta, A. Hautefeuille, P. Boffetta, and S. Porru, "Polymorphisms of the DNA repair genes XRCC1, XRCC3, XPD, interaction with environmental exposures, and bladder cancer risk in a case-control study in northern Italy," *Cancer Epidemiology Biomarkers & Prevention*, vol. 12, pp. 1234-1240, 2003.
- [38] J. D. Graves and E. G. Krebs, "Protein phosphorylation and signal transduction," *Pharmacology & therapeutics*, vol. 82, pp. 111-121, 1999.
- [39] K. Milanowska, J. Krwawicz, G. Papaj, J. Kosińska, K. Poleszak, J. Lesiak, E. Osinska, K. Rother, and J. M. Bujnicki, "REPAIRtoire- a database of DNA repair pathways," *Nucleic Acids Research*, vol. 39, pp. D788-D792, 2011.
- [40] L. Wen and J.-A. Feng, "Repair-FunMap: a functional database of proteins of the DNA repair systems," *Bioinformatics*, vol. 20, pp. 2135-2137, 2004.
- [41] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The international HapMap project web site," *Genome research*, vol. 15, pp. 1592-1593, 2005.

- 
- [42] J. C. Barrett, B. Fry, J. Maller, and M. J. Daly, "Haploview: analysis and visualization of LD and haplotype maps," *Bioinformatics*, vol. 21, pp. 263-265, 2005.
- [43] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, "dbSNP: the NCBI database of genetic variation," *Nucleic acids research*, vol. 29, pp. 308-311, 2001.
- [44] A. J. Brookes, H. Lehtinen, M. Siegfried, J. G. Boehm, Y. P. Yuan, C. M. Sarkar, P. Bork, and F. Ortigao, "HGBASE: a database of SNPs and other variations in and around human genes," *Nucleic Acids Research*, vol. 28, pp. 356-360, 2000.
- [45] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature protocols*, vol. 4, pp. 1073-1081, 2009.
- [46] N. Blom, S. Gammeltoft, and S. R. Brunak, "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites," *Journal of molecular biology*, vol. 294, pp. 1351-1362, 1999.
- [47] A. Kreegipuu, N. Blom, and S. R. Brunak, "PhosphoBase, a database of phosphorylation sites: release 2.0," *Nucleic acids research*, vol. 27, pp. 237-239, 1999.
- [48] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, pp. D514-D517, 2005.
- [49] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database," *Nature genetics*, vol. 36, pp. 431-432, 2004.
- [50] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet, "GeneCards: integrating information about genes, proteins and diseases," *Trends in Genetics*, vol. 13, p. 163, 1997.
- [51] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic acids research*, vol. 40, pp. D109-114, 2012.
- [52] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, and B. Jassal, "Reactome: a database of reactions, pathways and biological processes," *Nucleic acids research*, vol. 39, pp. D691-697, 2010.

# CHAPTER-3

Systems biology approach for mutational and site-specific structural investigation of DNA repair genes for xeroderma pigmentosum



*“To acquire knowledge, one must study. But to acquire wisdom, one must observe.”*  
*-Marilyn vos Savant*

## **ABSTRACT**

Xeroderma pigmentosum (XP), a rare genetic disorder wherein extreme sensitivity to ultraviolet (UV) radiations is observed that causes damage to DNA. These damages lead to skin and often neurological abnormalities since the repair process is not triggered. The DNA repair implicated in fixing UV linked damages is nucleotide excision repair (NER) and abnormalities or mutations in genes of this mechanism form basis of the disorder. The analysis of mutations in NER genes is vital for understanding XP and the cancer genetics as it may facilitate the identification of crucial biomarkers and anticancer therapeutic drugs. In this study, non-synonymous single nucleotide polymorphisms (nsSNPs) and their deleterious effects have been detected which provides a fundamental association of genetic mutations with phenotypic variations. To systematically comprehend the effect of genetic variations in XP, diverse genetic parameters like linkage disequilibrium (LD), haplotypes and other genetic markers were examined. The abnormal role of these genetic deviations in the development of skin and other forms of cancers currently amounts to a sizzling issue in the field of molecular systems biology. Thus, we have proposed a model for the pathway drawn in XP and also executed the simulation studies for examining the role of associated entities. Moreover, the mutations have been examined at structure-level by altering the structure, estimating the secondary structure, solvent accessibility and performing site specific analysis. Crucial phosphorylation sites have also been identified to study its role in the disorder. The mutational and structural analysis offers valuable insight for the understanding of disease and will further assist experimental biologists to evaluate these mutations and their impact on the genome.

### 3.1 INTRODUCTION

The human genome is susceptible to numerous damages due to intrinsic and environmental factors like reactive oxygen species, harmful radiations and chemicals [1]. These lesions lead to chromosomal aberrations and interfere with the genomic integrity. To eradicate the effect of these genetic aberrations, an intricate process of DNA repair is involved including several DNA repair genes and a variety of repair mechanisms [2]. Nucleotide excision repair (NER) is one of the important DNA repair mechanisms that repair damages like ultraviolet (UV) induced cyclobutane pyrimidine dimers and chemically induced bulky base adducts. There are two forms of NER [1-3], i.e. transcription coupled repair (TCR), which swiftly repair regions of DNA which are “active” and are transcribed into RNA and the other one is the global genome repair (GGR) which slowly repairs the damage occurring in rest of the genome. Both the NER sub pathways differ only in their mode of recognition of damage via different set of proteins and have all other steps such as lesion incision, repair and ligation in common. The deficiency in NER system is mainly associated with xeroderma pigmentosum (XP), a disorder of dry and pigmented skin caused due to the exposure of UV radiations and failure in its repair [1, 3].

As discussed in our previous objective, there are innumerable human diseases reported due to aberrations in repair processes and there is a huge lack of knowledge underlying the disease mechanisms. Consequently we anticipated to analyze the different DNA repair genes and their involvement in critical DNA repair associated diseases. In the present study, systems biology approach was practiced on XP which is roughly known to exist from past 139 years and the prevalence is found all over the world. The disease is highly prevalent in the Japanese population with equal rate of incidence for both males and females with a ratio of occurrence as 1:250,000 [4-6]. The manifestation of the disorder include severe sunburns when exposed to even minute sunlight [3, 7], freckles on the face, irritation and pain in the eyes, corneal ulcerations and increased risk of skin cancer [7]. A number of cancers like cancers on face, lips, scalp, eyes, tip of the tongue and eyelids have been associated with XP [8, 9]. XP not only affect the eyes or skin but also induce neurological abnormalities such as loss in hearing ability, seizures and deprived movement and speech [7, 10]. The mode of inheritance for this genetic disorder of DNA repair is autosomal recessive where mainly NER enzymes are mutated leading to inappropriate functioning of NER. The genes involved in XP are crucial component of NER

---

system which recognizes and repairs DNA damage caused due to UV-induced photoproducts and large DNA adducts.

XP is caused by a combination of reasonably harmful mutations in different genes which leads to the disruption of involved functional networks and further result in human variation. The key XP associated mutations are linked to both the alleles of eight genes [5, 11, 12] namely XPA- XP complementation group A, XPB, XPC, XPD, XPE, XPF, XPG and XPV (POLH) and for small percentage of the disorder, the mutations in other genes are also known to exist. XPC and XPE (DDB2) are the DNA binding proteins that recognize the damages to DNA [13, 14] specifically in GGR, the unwinding of the damaged site is carried out by XPB (ERCC3) and XPD (ERCC2) helicases [15, 16] which are the component of TFIIH (Transcription factor II Human) complex [17]. XPA protein is then required for stabilizing the unwound region and XPG (ERCC5) and XPF (ERCC4) cuts down the DNA on either side of the damage so as the intact DNA can replace the damaged portion [18, 19]. The eighth gene, POLH which is known to cause XP is not a part of NER and also recognized as XP variant (XP-V) shows difficulty in the replication of DNA containing UV-induced damage [20, 21]. The XPC and XPE proteins are the major components of GGR whereas other NER-related XP groups (XP-A, B, D, F and G) are found in both the sub-pathways (GGR and TCR). If the mutations in any of the above genes are not rectified, the damage by UV may also cause mutations in the cell's DNA [22].

In XP affected individual, the association of genetic deficiencies in NER genes was also found to be interrelated with other disorders like cockayne syndrome (CS) and gastric cancers. In gastric cancer, a strong link with the associated haplotypes of XPD gene was reported [23, 24] and there are some cases where XP patients have been identified with CS because of mutations in XPB, XPD and XPG genes [25]. The LD and haplotype analysis of MLH1 gene in hereditary non-polyposis colorectal cancer (HNPCC) also provides useful insight in the understanding of the disorder at genetic level [26]. Phosphorylation is a crucial process which determines the effective functioning of various cellular mechanisms, signaling pathways, sub-cellular locations, DNA damage recognition and its repair [27]. The phosphorylation state of NER genes also governs the severity of disease and further the development of cancers. In earlier studies, Wu et al. [28] too demonstrated that the phosphorylation of 'S' residue at 196 position in XPA protein

transforms the cellular activity of NER to promote the cell survival on UV irradiation. Therefore, the analysis of phosphorylation states of XP related genes may provide some constructive imminent for the disease.

The system level perceptive of XP is required to study the interactions of NER related proteins which may be useful in the development of therapeutic targets. There are evidences where studies on a particular disease as a system and simulation of the biological pathways to analyze the various reaction species have been very fruitful [29-32]. Since, there is no defined pathway available for the disease although pathway implicated in NER is known; therefore there is a need to perform a comprehensive system level analysis for the disorder. We applied a novel integrative approach where we implemented reductionist approach towards the elucidation of components involved in XP which ranges from sequence and structural level site-specific mutations to modeling and simulation of XP associated pathways. It is anticipated that the combined outcome of this study would provide biologically meaningful results and will be of utmost use to the researchers and biomedical scientists. In this study, the comprehensive analyses for the genetic variations, their structural impact on XP disorder with their association in biological pathways have been examined. This extensive *in silico* analysis is assumed to improve the diagnosis, treatment and other therapeutics for skin cancer once the key components are detected.

## **3.2 MATERIALS AND METHODS**

### **3.2.1 Data Collection**

The data was collected for the genes responsible for causing XP from OMIM database and eight key genes were found to be associated with the disorder. The information regarding the single base changes in DNA repair genes was gathered from numerous databases like dbSNP, HGVBBase (Human genome variation database), etc. For our *in silico* analysis, the SNP's and the protein sequences for the eight DNA repair genes linked to XP were collected from NCBI.

### **3.2.2 Assessment of coding single nucleotide polymorphisms**

The coding SNP's were filtered out from all the known SNP's in genes related to XP based on their effect on the phenotypic variations. Various computational tools like SIFT [33] and

PolyPhen (Polymorphism Phenotyping) [34] comprising of complex algorithms for the prediction of functional consequences of nsSNPs have been used for extracting important non-synonymous SNP's [35]. SIFT uses an algorithm comprising of customized version of PSI-BLAST and Dirichlet mixture regularization for building multiple sequence alignments of the proteins. Tolerance scores ranging from 0 to 1 are applied to each residue and the scores  $\leq 0.05$  are predicted as intolerant or deleterious substitutions and those above this value are considered as tolerant. Further, for the identification and detection of putative functional nsSNPs, PolyPhen was used which incorporates sequence, structural and evolutionary annotation details along with the substitutions in the proteins. It computes position-specific independent counts (PSIC) scores and then figures out the difference in PSIC scores of the two variants [36]. Higher the difference in PSIC score, greater is the functional influence of an individual amino acid substitution. PolyPhen categorizes the substitutions as “benign”, “possibly damaging” and “probably damaging” based on the PSIC scores where the difference  $\geq 1.5$  is reflected as damaging.

### **3.2.3 Investigation of phenotypic impact of SNPs**

The functional impact of mutations were examined for DNA repair genes of XP through numerous tools like FastSNP (Function Analysis and Selection Tool for Single Nucleotide Polymorphisms) [37], SNPnexus [38] and PupaSuite [39]. FastSNP was used for the efficient identification and prioritization of high-risk SNPs based on their phenotypic risks and putative functional effects. SNPnexus provides annotations for known and novel SNPs on the chief transcriptome, proteome, regulatory and structural variation models to facilitate the identification of phenotypically important variants. The PupaSuite tool provides information regarding both coding and non-coding SNPs along with disease related mutation annotations from SwissProt. PupaSuite is a hybrid of PupaSNP and PupasView which helps in the detection of SNPs with putative phenotypic effect. It also includes the predictions from SNPeffect database.

### **3.2.4 Quantitative genetic association study and identification of phosphorylation sites as vital parameters implicated in XP**

The genetic investigation is particularly important when the mutations are major factors responsible for causing the disease. The genotype data was pulled out from The International HapMap Project [40]. The parameters like LD, haplotypes and vital markers in the eight DNA



repair genes involved in XP were analyzed for the detection of candidate markers for the disease. The LD provides information for the non-random association of alleles whereas the haplotypes are the combination of alleles at neighboring loci that are always transmitted together. The chief parameters under study for the genetic association were  $D'$  and  $r^2$ . These parameters are already well described in Chapter 2. The phosphorylation sites were investigated from NetPhos algorithm [41] which uses artificial neural networks trained on PhosphoBase [42], a database of phosphorylation sites in the proteins. In the eight DNA repair genes related to XP, the phosphorylated positions at S, T and Y residues have been identified. This may prove useful for the thorough investigation of the disease at the cellular level.

### **3.2.5 Detection of site-specific structural conservation for nsSNPs**

The variations identified in the XP related NER and TLS genes were classified as altering the function and resulting in different phenotypes. To analyze the locations of the nsSNPs, CONSURF server [43] was employed which illustrates whether the mutations are occurring at the conserved (the mutations are mostly not tolerated) or variable sites (impact of the mutation may be tolerated). The server is based on the principle that the degree of conservation at each amino-acid position is similar to the inverse of the site's rate of evolution i.e. slowly evolving sites are evolutionarily conserved and vice-versa.

### **3.2.6 Modeling nsSNPs on 3D protein structures and RMSD calculations**

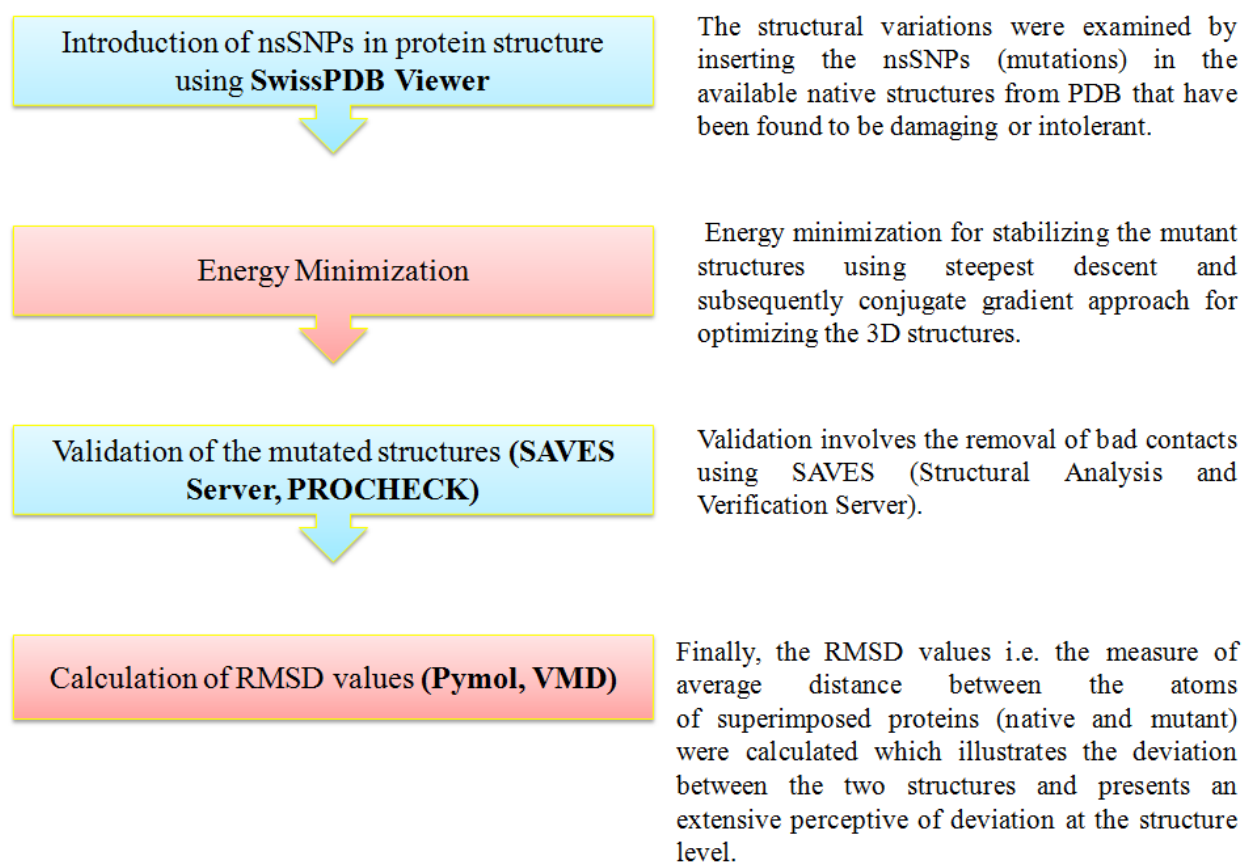
After assessing the site-specific conservation, further the examination of mutations for the structural variations was performed by inserting the mutations in the available native structures from Protein Data Bank (PDB) followed by energy minimization and calculation of Root Mean Square Deviation (RMSD) values for the detection of deviation between the two structures i.e. native and the mutant structures. Following tools and servers were used for this analysis which comprises of SwissPDB Viewer [44], SAVES (Structural Analysis and Verification Server), PROCHECK program [45], Pymol and VMD (Visual Molecular Dynamics). The entire workflow has been represented in **Figure 3.1**.

### 3.2.7 Effect of mutation on protein stability and secondary structures

The genetic variations or mutations may have a drastic influence on the stability of the protein and these variations may destabilize the proteins as well. Therefore, we attempted to analyze the stability through solvent accessibility for the XP related genes using NetSurfP [46] which calculates the relative solvent accessibility (RSA) by the equation:

$$RSA = \frac{ASA}{ASA_{MAX}} \cdot 100\%$$

Here, RSA is the ratio of the solvent Accessible Surface Area (ASA) of a given residue observed in the 3D structure, over the maximum obtainable solvent exposed area  $ASA_{MAX}$  for the given amino acid within an extended tri-peptide flanked with either glycine or alanine [47]. The secondary structure information was obtained from DSSP (Define Secondary Structure of Proteins) [48].



**Figure 3.1** The methodology for mapping the mutations onto the protein structures.

### 3.2.8 Quantitative study to simulate XP gene regulatory pathway

To understand the association of DDB2, ERCC2, ERCC3, ERCC4, ERCC5, XPA, XPC and POLH proteins in XP, the entire pathway i.e. involved in NER and TLS is examined. An attempt has been made through reductionist approach to design a model of the pathway involved in XP using CellDesigner [49]. The simulations were performed with respect to concentration and time scale parameters of different species (genes and proteins) for the proposed model and diverse graphs were obtained to study the behaviour of involved reaction species.

## 3.3 RESULTS AND DISCUSSION

### 3.3.1 Identification and evaluation of nsSNPs and their phenotypic effects

The experimental detection of complex disease associated mutations is very intricate hence bioinformatics analysis has been performed for its identification. From the literature survey and browsing through a number of databases, we found that there are eight genes related to XP namely DDB2, ERCC2, ERCC3, ERCC4, ERCC5, POLH, XPA and XPC. A total of 191 nsSNPs were detected in these genes and out of which 55 were predicted to be highly intolerant mutations. A table for genes implicated in XP, nsSNPs, haplotype blocks in these genes is shown in **Table 3.1**.

**Table 3.1 Compiled list of XP genes, vital genetic markers and phenotypic consequences**

Genes	No. of nsSNPs	Damaging nsSNPs	Haplotype Blocks	Functional Impact of mutations
<b>DDB2</b>	8	5	1	Missense (conservative); splicing regulation
<b>ERCC2</b>	8	4	3	Splicing site; missense (conservative); splicing regulation
<b>ERCC3</b>	9	3	1	Missense (conservative); splicing regulation
<b>ERCC4</b>	23	11	1	Missense (non-conservative); splicing regulation; missense (conservative)
<b>ERCC5</b>	40	8	1	Missense (non-conservative); splicing regulation
<b>POLH</b>	70	21	1	Missense (conservative); splicing regulation; missense (non-conservative)
<b>XPA</b>	7	1	1	Missense (conservative)
<b>XPC</b>	26	2	1	Missense (conservative)
<b>Total</b>	<b>191</b>	<b>55</b>	<b>10</b>	-

The functional influence of nsSNPs in eight DNA repair genes involved in NER that cause XP was exhaustively analysed. The nsSNPs were found to have damaging as well as tolerant effects. The consequence of these damages were evaluated from SIFT and PolyPhen tools. SIFT algorithm assign the scores to each nsSNP position where an outcome is classified as (0.00–0.05) for intolerant, (0.051–0.10) for possibly intolerant, (0.101–0.30) for borderline and (0.301–1.00) for tolerant mutations whereas PolyPhen tool assign the effect of the mutation based on the scores such as benign for (0-0.5), probably damaging for (0.5-0.9) and possibly damaging for (0.9-1). Those nsSNPs which were consensus in both PolyPhen and SIFT prediction were also validated from dbSNP and are represented in **Table 3.2**.

**Table 3.2 The association of genetic mutations and phenotypic variations in XP genes**

XP Related Genes			SIFT Prediction		POLYPHEN Prediction		Consensus Prediction
Gene Name	SNP ID	Amino Acid Change	Tolerance Index	Predicted Impact	Probability Score	Predicted Impact	SIFT + PolyPhen
DDB2	rs121434642	D307Y	0.04	Intolerant	1	Damaging	YES
	rs121434640	R273H	0.02	Intolerant	0.099	Benign	-
	rs121434639	K244E	0.17	Borderline	0.94	Probably damaging	YES
	rs78651238	A108G	0.35	Tolerant	0.013	Benign	-
	rs77897070	K427N	0	Intolerant	0.23	Benign	-
	rs4647751	A293T	0.14	Borderline	0.779	Possibly damaging	YES
	rs4647750	M215T	0.7	Tolerant	0.001	Benign	-
ERCC2	rs74792417	I569F	0	Intolerant	0.445	Benign	-
	rs41559922	F448S	0	Intolerant	0.149	Benign	-
	rs41556519	R683W	0	Intolerant	1	Damaging	YES
	rs34517175	A635V	0	Intolerant	0.969	Probably damaging	YES
	rs1799791	I199M	0.43	Tolerant	0.1	Benign	-
ERCC3	rs116713511	E105G	0	Intolerant	1	Damaging	YES
	rs115312738	D755N	0.12	Borderline	0.035	Benign	-
	rs115176021	I401V	0.28	Borderline	0.038	Benign	-
	rs114613120	P775L	0	Intolerant	0.055	Benign	-
	rs1805162	G402C	0	Intolerant	0.856	Possibly damaging	YES
ERCC4	rs112490976	E647G	0.03	Intolerant	0.521	Possibly damaging	YES
	rs61760161	F607L	0.06	Possibly Intolerant	0.937	Probably damaging	YES
	rs61731714	V34L	0.08	Possibly Intolerant	0.285	Benign	-
	rs55761944	V81F	0.07	Possibly Intolerant	0.891	Possibly damaging	YES
	rs41557814	R477W	0.01	Intolerant	0.617	Possibly damaging	-
	rs12932917	C745G	0	Intolerant	0.968	Probably damaging	YES
	rs12928616	S747F	0.02	Intolerant	0.941	Probably damaging	YES
	rs2020956	G912E	0.29	Borderline	0.012	Benign	-
	rs1800124	E875G	0.01	Intolerant	0.585	Possibly damaging	YES
	rs1800069	I706T	0	Intolerant	0.998	Probably damaging	YES

	rs1800067 rs1799802	R415Q P379S	0.01 0.02	Intolerant Intolerant	0.98 0.994	Probably damaging Probably damaging	YES YES
ERCC5	rs121434576	A874T	0	Intolerant	0.856	Possibly damaging	YES
	rs121434575	L858P	0	Intolerant	0.999	Probably damaging	YES
	rs121434574	P72H	0	Intolerant	0.971	Probably damaging	YES
	rs121434571	A792V	0	Intolerant	0.958	Probably damaging	YES
	rs113590348	L6P	0	Intolerant	1	Damaging	YES
	rs112571039	I295T	0.6	Tolerant	0.276	Benign	-
	rs56398372	A376V	0.15	Borderline	0.022	Benign	-
	rs56255799	R214C	0.01	Intolerant	1	Damaging	YES
	rs41564320	Q1002R	0.34	Tolerant	0.038	Benign	-
	rs41281674	R959S	0	Intolerant	0.11	Benign	-
	rs1047769 rs17655	M254V D1104H	0.11 0.04	Borderline Intolerant	0.365 0.819	Benign Possibly damaging	- YES
POLH	rs113074920	R81C	0	Intolerant	0.856	Possibly damaging	YES
	rs78080414	C321W	0	Intolerant	1	Damaging	YES
	rs77588969	K231N	0	Intolerant	1	Damaging	YES
	rs61756403	N233S	0.07	Possibly Intolerant	0.741	Possibly damaging	YES
	rs61748656	M14V	0.04	Intolerant	0.999	Probably damaging	YES
	rs35675573	T329I	0.16	Borderline	0.042	Benign	-
	rs2307456	G209V	0.01	Intolerant	0.956	Probably damaging	YES
	rs80185103	E1857G	0.01	Intolerant	0.997	Probably damaging	YES
	rs79332480	S276F	0.02	Intolerant	0.876	Possibly damaging	YES
	rs79146763	G2422V	0	Intolerant	1	Damaging	YES
	rs71329221	P2014L	0	Intolerant	0.82	Possibly damaging	YES
	rs61734794	D1784Y	0.04	Intolerant	0.014	Benign	-
	rs56104120	T817I	0.01	Intolerant	0.945	Probably damaging	YES
	rs55943551	V310G	0.03	Intolerant	0.335	Benign	-
	rs55923976	L197R	0	Intolerant	0.998	Probably damaging	YES
	rs55748151	A2464T	0	Intolerant	1	Damaging	YES
	rs41540016	G1751W	0.02	Intolerant	0.624	Possibly damaging	YES
	rs3218643	D1562Y	0.01	Intolerant	0.754	Possibly damaging	YES
	rs3218639	Q1565H	0	Intolerant	0.941	Probably damaging	YES
	rs3218635	E2465K	0.01	Intolerant	0.999	Probably damaging	YES
rs3218634 rs532411	L2538V A2304V	0.19 0.07	Borderline Possibly Intolerant	0.998 0.185	Probably damaging Benign	YES -	
XPA	rs104894131	C108F	0	Intolerant	1	Damaging	YES
XPC	rs121965091	F265S	0.16	Borderline	0.998	Probably damaging	YES
	rs56267823	R390M	0.06	Possibly Intolerant	0.37	Benign	-
	rs56012223	D68V	0.26	Borderline	0.015	Benign	-

The variations in NER related genes caused changes in the expression, functions and their important roles in the repair mechanism. Those variations or nsSNPs that alter the protein sequences were identified from sources like SNPnexus, FastSNP and Pupasuite. The phenotypic changes caused by these variations have been shown in **Table 3.3**.

**Table 3.3 The risk association of various SNPs with their phenotypic effects**

Gene	SNP ID	Allele 1	Allele 2	Region	Lower Risk*	Upper Risk*	Possible Functional Effects
DDB2	rs4647750	T	C	Coding	3	4	Missense (non-conservative); Splicing regulation
DDB2	rs11537594	G	A	Coding	2	3	Missense (conservative); Splicing regulation
DDB2	rs4647751	G	A	Coding	2	3	Missense (conservative); Splicing regulation
ERCC2	rs41556519	G	A	Coding	3	4	Splicing site
ERCC2	rs34517175	G	A	Coding	3	4	Splicing site
ERCC2	rs1799791	G	C	Coding	3	4	Splicing site
ERCC2	rs13181	G	T	Coding	2	3	Missense (conservative); Splicing regulation
ERCC2	rs41559922	A	G	Coding	2	3	Missense (conservative); Splicing regulation
ERCC2	rs1799793	G	A	Coding	2	3	Missense (conservative); Splicing regulation
ERCC2	rs1799792	T	C	Coding	2	3	Missense (conservative)
ERCC3	rs4150521	C	T	Coding	3	4	Missense (non-conservative)
ERCC3	rs4150522	T	C	Coding	2	3	Missense (conservative)
ERCC3	rs1805162	C	W	Coding	2	3	Missense (conservative); Splicing regulation
ERCC3	rs1805161	A	G	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs1799802	C	T	Coding	3	4	Missense (non-conservative)
ERCC4	rs1800067	G	A	Coding	3	4	Missense (non-conservative); Splicing regulation
ERCC4	rs1800068	G	C	Coding	3	4	Missense (non-conservative); Splicing regulation
ERCC4	rs1800069	T	C	Coding	3	4	Missense (non-conservative); Splicing regulation
ERCC4	rs12932917	T	G	Coding	3	4	Missense (non-conservative); Splicing regulation
ERCC4	rs12928616	C	T	Coding	3	4	Missense (non-conservative); Splicing regulation
ERCC4	rs12928650	C	T	Coding	3	4	Missense (non-conservative)
ERCC4	rs1800124	A	G	Coding	3	4	Missense (non-conservative); Splicing regulation
ERCC4	rs61760160	C	T	Coding	2	3	Missense (conservative)
ERCC4	rs34205098	C	G	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs61731714	G	C	Coding	2	3	Missense (conservative)
ERCC4	rs55761944	G	W	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs2020961	C	C/G/T	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs41557814	C	T	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs41552412	C	G	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs55736359	G	A	Coding	2	3	Missense (conservative)
ERCC4	rs61760161	T	G	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs2020955	T	C	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs56129764	G	A	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs4986933	C	A	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs2020957	A	G	Coding	2	3	Missense (conservative); Splicing regulation
ERCC4	rs2020956	G	A	Coding	2	3	Missense (conservative); Splicing regulation
ERCC5	rs34291397	C	T	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs2307456	G	T	Coding	3	4	Missense (non-conservative); Splicing regulation
POLH	rs9333548	C	C/G/T	Coding	3	4	Missense (non-conservative)
POLH	rs9296419	C	T	Coding	3	4	Missense (non-conservative); Splicing regulation
POLH	rs6941583	A	T	Coding	3	4	Missense (non-conservative)
POLH	rs61748656	A	G	Coding	2	3	Missense (conservative); Splicing regulation

\* Corresponds to the prioritization of SNPs based on their phenotypic risks.

POLH	rs61756403	A	G	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs35675573	C	T	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs61076173	C	T	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs56307355	A	G	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs56213129	C	T	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs9333554	T	C	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs55682447	A	C	Coding	2	3	Missense (conservative); Splicing regulation
POLH	rs9333555	A	G	Coding	2	3	Missense (conservative); Splicing regulation
XPA	rs3176750	C	G	Coding	2	3	Missense (conservative); Splicing regulation
XPA	rs3176749	G	T	Coding	2	3	Missense (conservative); Splicing regulation
XPA	rs1805160	G	A	Coding	2	3	Missense (conservative); Splicing regulation
XPA	rs10983315	C	T	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs2228001	A	C	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs3731177	T	G	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs55779831	C	G	Coding	2	3	Missense (conservative)
XPC	rs3731152	C	T	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs2227998	G	A/C/G	Coding	2	3	Sense/synonymous; Splicing regulation
XPC	rs3731140	C	W	Coding	2	3	Missense (conservative)
XPC	rs3731139	G	C	Coding	2	3	Missense (conservative)
XPC	rs3731130	G	C	Coding	2	3	Missense (conservative)
XPC	rs6413541	A	C	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs2228000	C	T	Coding	2	3	Missense (conservative)
XPC	rs2227999	A	G	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs56267823	C	A	Coding	2	3	Missense (conservative)
XPC	rs3731126	C	T	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs35629274	A	C	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs56115311	G	C	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs3731063	A	G	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs56012223	T	A	Coding	2	3	Missense (conservative)
XPC	rs1126482	A	G	Coding	2	3	Missense (conservative); Splicing regulation
XPC	rs1870134	G	C	Coding	2	3	Missense (conservative); Splicing regulation

### 3.3.2 Quantitative genetic analysis and putative phosphorylation sites

The vital genetic parameters have been detected in the eight NER DNA repair genes concerning XP. Important haplotypes, i.e. blocks of associated SNPs conserved across the genome have been identified in each of the eight genes. The haplotype blocks representing the correlation of several residue conditions among the polymorphic sites across the genome have been shown in **Appendix A.1**.

The important phosphorylation positions that may be involved in XP have been predicted from NetPhos algorithm and represented in **Appendix A.2**. A total of ‘338’ S residues, ‘103’ T residues and ‘53’ Y residues have been identified in the eight DNA repair genes that are

implicated in XP. The identification of these phosphorylation sites will provide useful insight for the discovery of essential biomarkers for the disease. Additionally, we mined experimentally verified phosphorylation sites for XP genes where we found many common entries between our predictions and experimental data.

### 3.3.3 Inspection of nsSNP locations on the protein structures

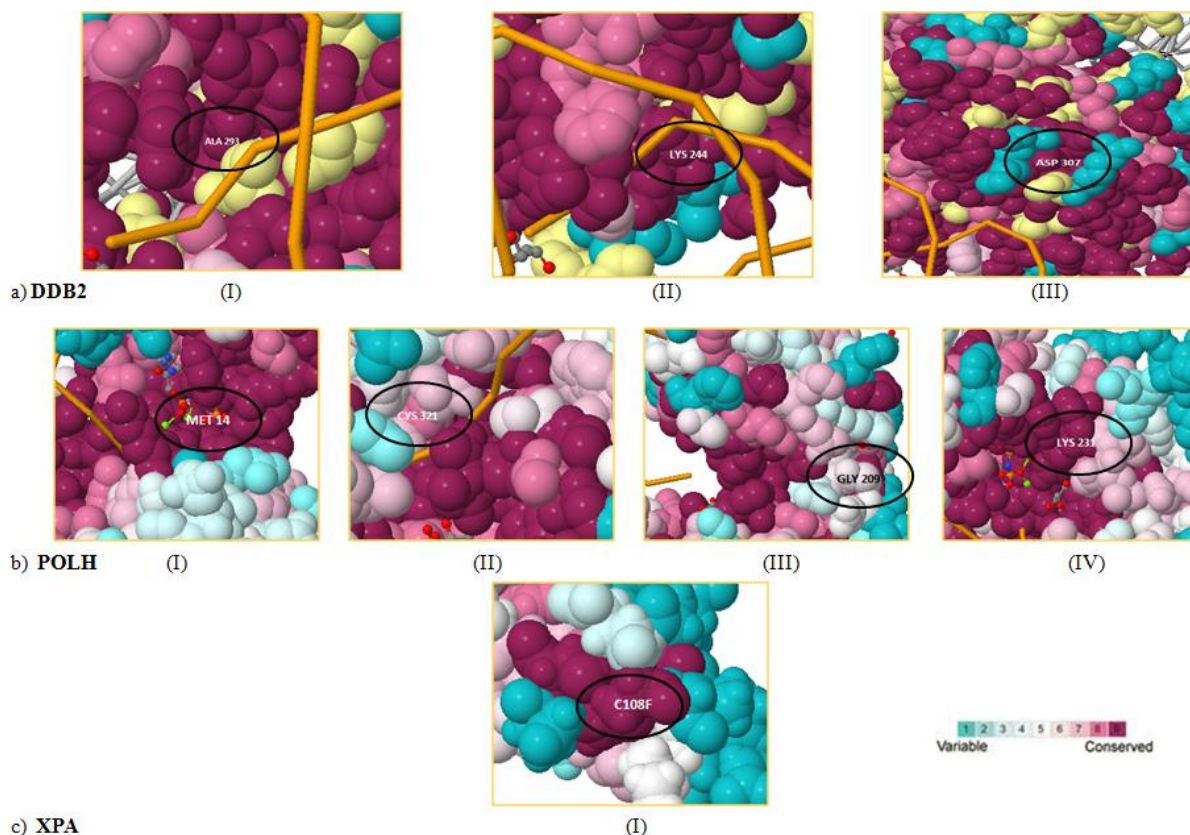
Using the CONSURF server, we observed the effect of nsSNPs at the structural level. Here, the site-specific mutations were observed on structures where the selection pressure works on the mutations. We found that for DDB2 gene, these mutations ('K244E', 'A293T', 'D307Y') have an intolerant impact on the structure as predicted by SIFT and PolyPhen therefore, we confirmed our analysis from CONSURF server where mutations were present in the evolutionary conserved portion of the proteins since the mutations were found in the region with bordeaux colours as seen in **Figure 3.2(a)**. Mutations in the conserved regions are evolutionary not preferential so they have damaging consequences. We also observed that these mutations are very much close to the interacting DNA molecule thus confirming the harmful effect of these mutations. Similarly for POLH gene, these mutations ('M14V', 'G209V', 'K231N', 'C321W') and for XPA gene, the mutation (C108F) was predicted to have damaging effect from SIFT and PolyPhen tools. In POLH gene, the mutations M14V and K231N were found to be present at highly conserved regions (bordeaux color) as compared to C321W (light pink) which is more conserved than G209V (very light pink) through CONSURF analysis as shown in **Figure 3.2(b)**. Similarly, in XPA protein, C108F mutation was found in the highly conserved region of the protein therefore the mutation is also not favorably accepted as may lead to damaging effects as observed in **Figure 3.2(c)**.

### 3.3.4 Modeling of mutations onto the protein structures

Here, the protein structures were available for three XP related DNA repair genes namely, DDB2 (PDB Id: 4E54), POLH (PDB Id: 3MR2), XPA (PDB Id: 1XPA). There is a need to perform a mutational study for the deleterious mutations occurring in these genes that may have an important role to play in XP. The nsSNPs in these three DNA repair genes involved in NER and TLS were analysed and then substitutions were made in the available native 3D structures from PDB. The individual mutations were mapped on the native structures to make them the mutant



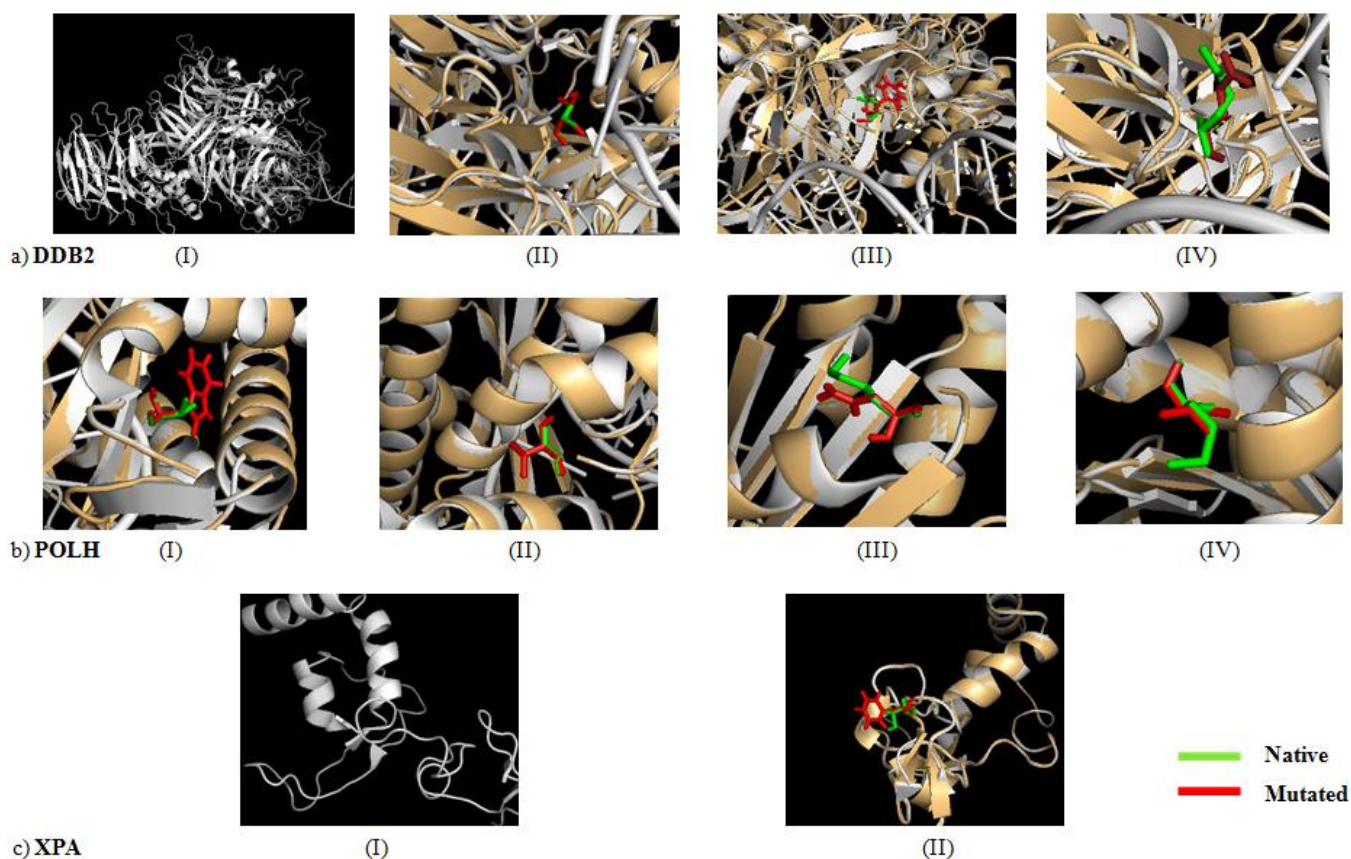
structures and then compared with the native ones. In DDB2, POLH and XPA gene, the mutations ('K244E', 'A293T', 'D307Y'), ('M14V', 'G209V', 'K231N', 'C321W') and ('C108F') were confirmed to have damaging impact and were analyzed by modeling the substitutions onto the PDB structures to create the mutant structures using Swiss PDB Viewer. The structures (native and mutant) were superimposed and the native and mutated residues were highlighted with green and red colour using Pymol respectively (**Figure 3.3**).



**Figure 3.2 (a)** In DDB2 protein, ('A293T', 'K244E', 'D307Y') mutations were mapped on the structure of DDB2 (PDB ID: 4E54) and the site specific positions have been labeled where the three mutations have been observed at highly conserved sites of the protein due to bordeaux colour; **(b)** The POLH mutations ('M14V', 'G209V', 'K231N', 'C321W') were introduced in the structure (PDB ID: 3MR2) and been labeled where M14V and K231N mutations were found in the conserved region whereas C321W and G209V in less conserved as compared to earlier mutations due to light pink and very light pink colours; **(c)** In XPA protein, the C108F mutation was mapped to the protein structure (PDB ID: 1XPA) and was also found in the highly conserved region.

The energy minimizations of the mutant structures were carried out initially using steepest descent and then conjugate gradient method **Appendix A.3**. The energy minimized

structures were validated from the SAVES server. The RMSD values for both the structures (native and mutant) of DDB2, POLH and XPA were thus calculated which gives the degree of deviation between the two structures. The RMSD between the native structure and K244E, A293T and D307Y mutations of DDB2 were estimated as ‘0.174’, ‘0.1675’ and ‘0.1813’ respectively. The RMSD between the native structure and M14V, G209V, K231N and C321W mutations of POLH came out to be ‘0.092’, ‘0.091’, ‘0.091’ and ‘0.083’ respectively. The RMSD value for the native and the mutant structure of XPA came out to be ‘0.2295’.



**Figure 3.3** (a) For DDB2 protein, (I) is the native structure of DDB2 in cartoon representation and in (II-IV) the native structure (white) is superimposed with the mutant structure (orange) where (II) symbolizes the A293T mutation (III) corresponds to D307Y mutation (IV) represents K244E mutation where the native residues are highlighted in green colour and the mutants in red; (b) Likewise, in POLH, (I) symbolizes C321W mutation (II) corresponds to G209V mutation (III) represents K231N mutation (IV) showing M14V mutation where the native residues are highlighted in green colour and the mutants in red; (c) For XPA protein, (I) characterizes the native structure of XPA and (II) represents the native structure (white) superimposed to the mutant structure (orange) with C108F mutation in cartoon representation where the native residues are highlighted in green colour and the mutants in red.

### 3.3.5 Secondary structure and protein solvent accessibility

For the systematic understanding of the protein structures, secondary structure information is extremely valuable. Therefore, the secondary structure related information has been retrieved from DSSP and many vital findings were detected such as C108F mutation in XPA gene alters the bent structure to hydrogen bonded turn which in turn has an impact on the 3D structure of the XPA protein. Protein solvent accessibility (solvent ASA) is also a crucial parameter to analyze the stability of protein which has been computed from NetsurfP tool and revealed in **Table 3.4**. In general, the solvent accessible surface area is the surface area of the biomolecule (protein) i.e. accessible to the solvent and in NetSurfP it is divided into two categories i.e. buried and exposed which indicate whether the amino acids have low or high accessibility to the solvent.

**Table 3.4 Impact of mutations on the protein solvent accessibility and its secondary structure**

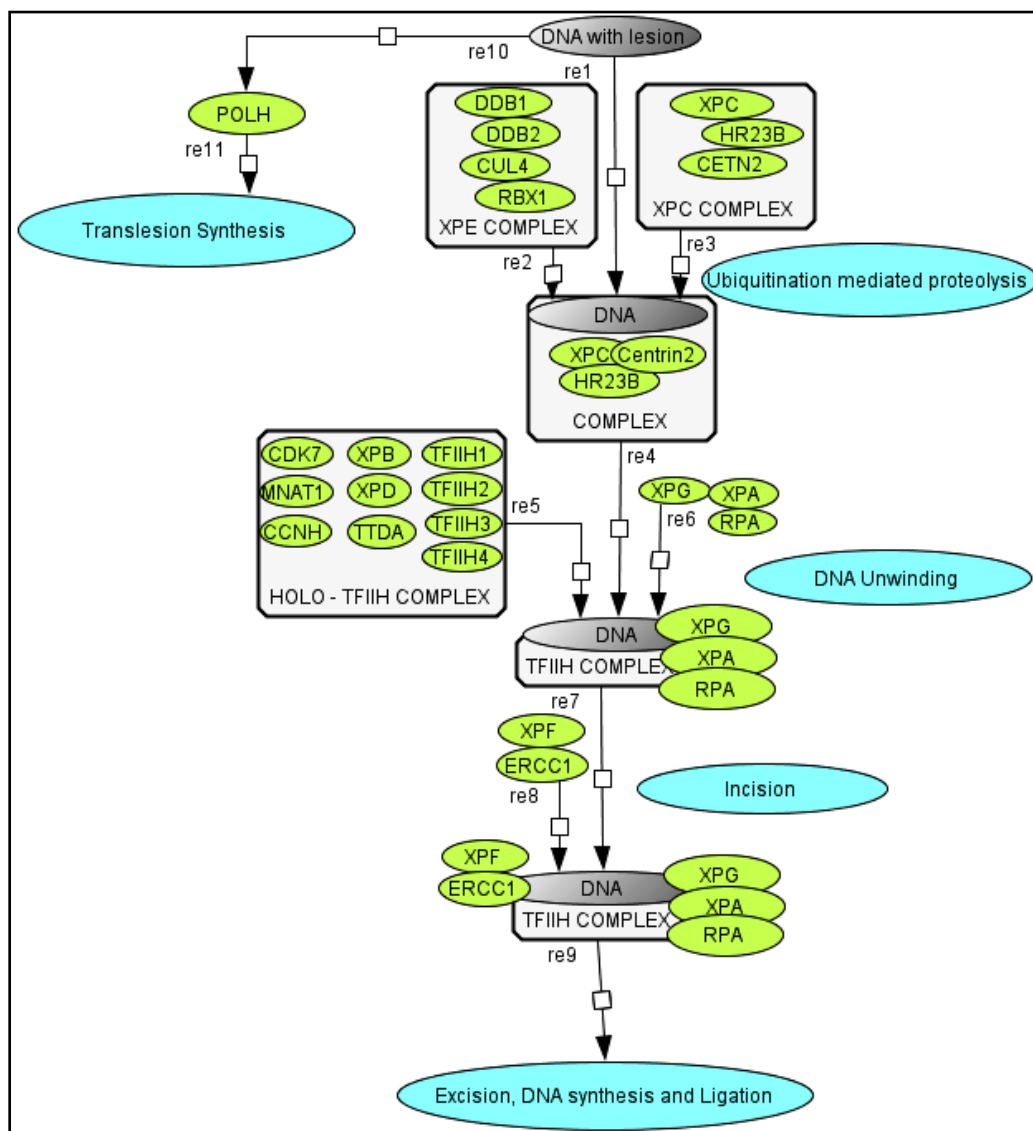
Gene	Mutation	Original class assignment	Secondary structure prediction *	Altered class assignment	Altered secondary structure *
DDB2	A293T	Buried	E	Buried	-
DDB2	D307Y	Buried	E	Buried	E
DDB2	K244E	Exposed	-	Exposed	-
XPA	C108F	Buried	S	Buried	T
POLH	C321W	Buried	E	Buried	E
POLH	G209V	Exposed	S	Exposed	S
POLH	K231N	Buried	-	Buried	-
POLH	M14V	Buried	E	Buried	E

### 3.3.6 XP regulatory pathways simulation studies

Additionally, there was no available report on the XP disease pathway; although interactions between involved entities were known via NER and TLS mechanisms; thus we reconstructed a putative XP associated pathway and carried out several simulation studies for understanding the interactions. We have designed a model of a biological pathway for XP depicted in **Figure 3.4**, representing all the proteins that are component of XP, comprising of those involved in NER system and POLH protein which is a part of TLS mechanism. In a XP patient, when damage to DNA is identified by a cell, two mechanisms may be affected depending on the type of damage i.e. NER or TLS. Here, we analyzed the various reaction species for their important interactions

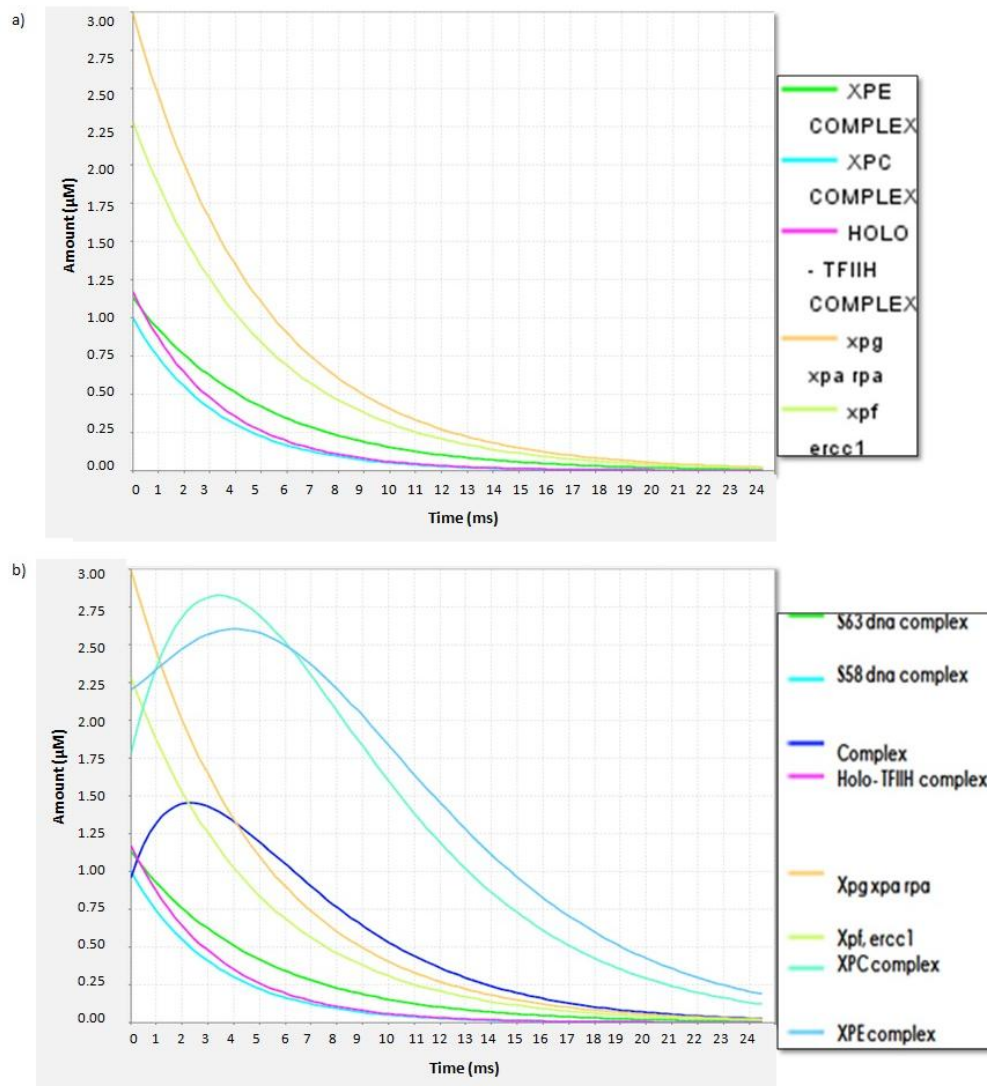
\* Here, S = Bend, T = hydrogen bonded turn and E = extended strand, participates in beta ladder.

and associations and obtained diverse graphs. In **Figure 3.5(a)**, an exponential decline in the graph was detected with the five complexes (XPE, XPC, Holo-TFIIH, XPG-XPA-RPA and XPF-ERCC1) with the end time of simulation as 24.5 milliseconds and concentration of each species as 1.13, 1.0, 1.17, 2.99 and 2.28 micro molar respectively. In **Figure 3.5(b)**, the graph in addition to the previous proteins also showed the DNA-protein complexes formed in the repair process and it was observed that the major complexes (formed before the incision of damaged DNA) reached a state of equilibrium at 25-30ms to further activate the repair process.



**Figure 3.4** A reconstructed pathway model for delineation of regulatory processes implicated in XP.

The comprehensive examination of XP facilitates the systems level study of current validated XP proteins involved in the disease, network components and the biological pathway which indeed will assist in the discovery of a novel regulatory process and further therapeutic strategies for skin cancer. Analyzing the deleterious effect of mutations, other genetic parameters like LD and haplotypes which are known to cause XP will help in inherent level understanding of the disease. The systems biology approach that we applied is expected to elucidate the molecular control of XP and further provide potential therapeutics for the treatment of XP disorder.



**Figure 3.5 (a)** A simulation performed through genes with respect to time and concentration where the genes corresponds to different coloured lines in the graph and are illustrated on right side of the image. **(b)** Includes the simulation studies executed with the DNA-protein complexes generated in the repair process.

### 3.4 CONCLUSION

The study characterizes the association of deleterious mutations resulting in XP and also proposes its impact on the genome. In our investigation, we have proposed a XP associated pathway along with the simulation analysis for the various interactions that may perhaps be affected due to these mutations. This comprehensive analysis of the disorder facilitated the characterization of the disease and revealed some unknown parameters such as LD, haplotypes, nsSNPs and important phosphorylation sites associated with XP. We postulated new computational annotations for the description of the vast genomic data from high-throughput sequencing, genotyping and haplotype diversity. The perceptive of the effect of mutations at protein structure level and its diverse phenotypes has been progressively important area in computational biology and therefore, our annotation initiated with the explanation of various phenotypic behaviours linked with specific SNPs, crucial mutated phosphorylation sites that may influence the genome and further mapping these mutations at structural level for the identification of vital candidate markers implicated in XP and other skin diseases. Additionally, the impact of these mutations for causing XP may be experimentally verified and validated by geneticists and experimental biologists.

### REFERENCES

- [1] J. H. J. Hoeijmakers, "Genome maintenance mechanisms for preventing cancer," *Nature*, vol. 411, pp. 366-374, 2001.
- [2] E. C. Friedberg, "How nucleotide excision repair protects against cancer," *Nature Reviews Cancer*, vol. 1, pp. 22-33, 2001.
- [3] L. Feller, N. H. Wood, M. H. Motswaledi, R. A. Khammissa, M. Meyer, and J. Lemmer, "Xeroderma pigmentosum: a case report and review of the literature," *J Prev Med Hygiene*, vol. 51, pp. 87-91, 2010.
- [4] Y. Hirai, Y. Kodama, S. I. Moriwaki, A. Noda, H. M. Cullings, D. G. MacPhee, K. Kodama, K. Mabuchi, K. H. Kraemer, and C. E. Land, "Heterozygous individuals bearing a founder mutation in the XPA DNA repair gene comprise nearly 1% of the Japanese population," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 601, pp. 171-178, 2006.

- 
- [5] P. Mahindra, J. J. DiGiovanna, D. Tamura, J. S. Brahim, T. J. Hornyak, J. B. Stern, C. C. R. Lee, S. G. Khan, B. P. Brooks, and J. A. Smith, "Skin cancers, blindness and anterior tongue mass in African brothers," *Journal of the American Academy of Dermatology*, vol. 59, pp. 881-886, 2008.
- [6] A. R. Lehmann, D. McGibbon, and M. Stefanini, "Xeroderma pigmentosum," *Orphanet J Rare Dis*, vol. 6, p. 70, 2011.
- [7] P. T. Bradford, A. M. Goldstein, D. Tamura, S. G. Khan, T. Ueda, J. Boyle, K. S. Oh, K. Imoto, H. Inui, and S. I. Moriwaki, "Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair," *Journal of medical genetics*, vol. 48, pp. 168-176, 2011.
- [8] G. Liu and X. Chen, "DNA polymerase  $\eta$ , the product of the xeroderma pigmentosum variant gene and a target of p53, modulates the DNA damage checkpoint and p53 activation," *Molecular and cellular biology*, vol. 26, pp. 1398-1413, 2006.
- [9] K. H. Kraemer, N. J. Patronas, R. Schiffmann, B. P. Brooks, D. Tamura, and J. J. DiGiovanna, "Xeroderma pigmentosum, trichothiodystrophy and Cockayne syndrome: a complex genotype-phenotype relationship," *Neuroscience*, vol. 145, pp. 1388-1396, 2007.
- [10] J. H. Robbins, K. H. Kraemer, M. A. Lutzner, B. W. Festoff, and H. G. Coon, "Xeroderma Pigmentosum An Inherited Disease with Sun Sensitivity, Multiple Cutaneous Neoplasms, and Abnormal DNA Repair," *Annals of internal medicine*, vol. 80, pp. 221-248, 1974.
- [11] D. Bootsma, H.K. Kraemer, E.J. Cleaver, and J. H. J. Hoeijmakers, *Nucleotide excision repair syndromes: xeroderma pigmentosum, Cockayne syndrome, and trichothiodystrophy* vol. 1. New York: McGraw-Hill Book Co., 2001.
- [12] G. Kulaksız, Joyce T. Reardon, and A. Sancar, "Xeroderma pigmentosum complementation group E protein (XPE/DDB2): purification of various complexes of XPE and analyses of their damaged DNA binding and putative DNA repair properties," *Molecular and cellular biology*, vol. 25, pp. 9784-9792, 2005.
- [13] M. Volker, M. J. Moné, P. Karmakar, A. van Hoffen, W. Schul, W. Vermeulen, J. H. J. Hoeijmakers, R. van Driel, A. A. van Zeeland, and L. H. F. Mullenders, "Sequential

- assembly of the nucleotide excision repair factors in vivo," *Molecular cell*, vol. 8, pp. 213-224, 2001.
- [14] K. Sugasawa, Y. Okuda, M. Saijo, R. Nishi, N. Matsuda, G. Chu, T. Mori, S. Iwai, K. Tanaka, and K. Tanaka, "UV-induced ubiquitylation of XPC protein mediated by UV-DDB-ubiquitin ligase complex," *Cell*, vol. 121, pp. 387-400, 2005.
- [15] J. R. Hwang, V. Moncollin, W. Vermeulen, T. Seroz, H. van Vuuren, J. H. J. Hoeijmakers, and J. M. Egly, "A 3'→5' XPB helicase defect in repair/transcription factor TFIIH of xeroderma pigmentosum group B affects both DNA repair and transcription," *Journal of Biological Chemistry*, vol. 271, pp. 15898-15904, 1996.
- [16] G. S. Winkler, S. J. Araújo, U. Fiedler, W. Vermeulen, F. Coin, J.-M. Egly, J. H. J. Hoeijmakers, R. D. Wood, H. T. M. Timmers, and G. Weeda, "TFIIH with inactive XPD helicase functions in transcription initiation but is defective in DNA repair," *Journal of Biological Chemistry*, vol. 275, pp. 4258-4266, 2000.
- [17] L. Schaeffer, V. Moncollin, R. Roy, A. Staub, M. Mezzina, A. Sarasin, G. Weeda, J. H. Hoeijmakers, and J. M. Egly, "The ERCC2/DNA repair protein is associated with the class II BTF2/TFIIH transcription factor," *The EMBO journal*, vol. 13, pp. 2388-2392, 1994.
- [18] A. O'Donovan, A. A. Davies, J. G. Moggs, S. C. West, and R. D. Wood, "XPG endonuclease makes the 3' incision in human DNA nucleotide excision repair," 1994.
- [19] T. Matsunaga, D. Mu, C. H. Park, J. T. Reardon, and A. Sancar, "Human DNA repair excision nuclease analysis of the roles of the subunits involved in dual incisions by using anti-XPG and anti-ERCC1 antibodies," *Journal of Biological Chemistry*, vol. 270, pp. 20862-20869, 1995.
- [20] A. R. Lehman, S. Kirk Bell, C. F. Arlett, M. C. Paterson, P. H. Lohman, E. A. de Weerd-Kastelein, and D. Bootsma, "Xeroderma pigmentosum cells with normal levels of excision repair have a defect in DNA synthesis after UV-irradiation," *Proceedings of the National Academy of Sciences*, vol. 72, pp. 219-223, 1975.
- [21] C. Masutani, R. Kusumoto, A. Yamada, N. Dohmae, M. Yokoi, M. Yuasa, M. Araki, S. Iwai, K. Takio, and F. Hanaoka, "The XPV (xeroderma pigmentosum variant) gene encodes human DNA polymerase  $\eta$ ," *Nature*, vol. 399, pp. 700-704, 1999.



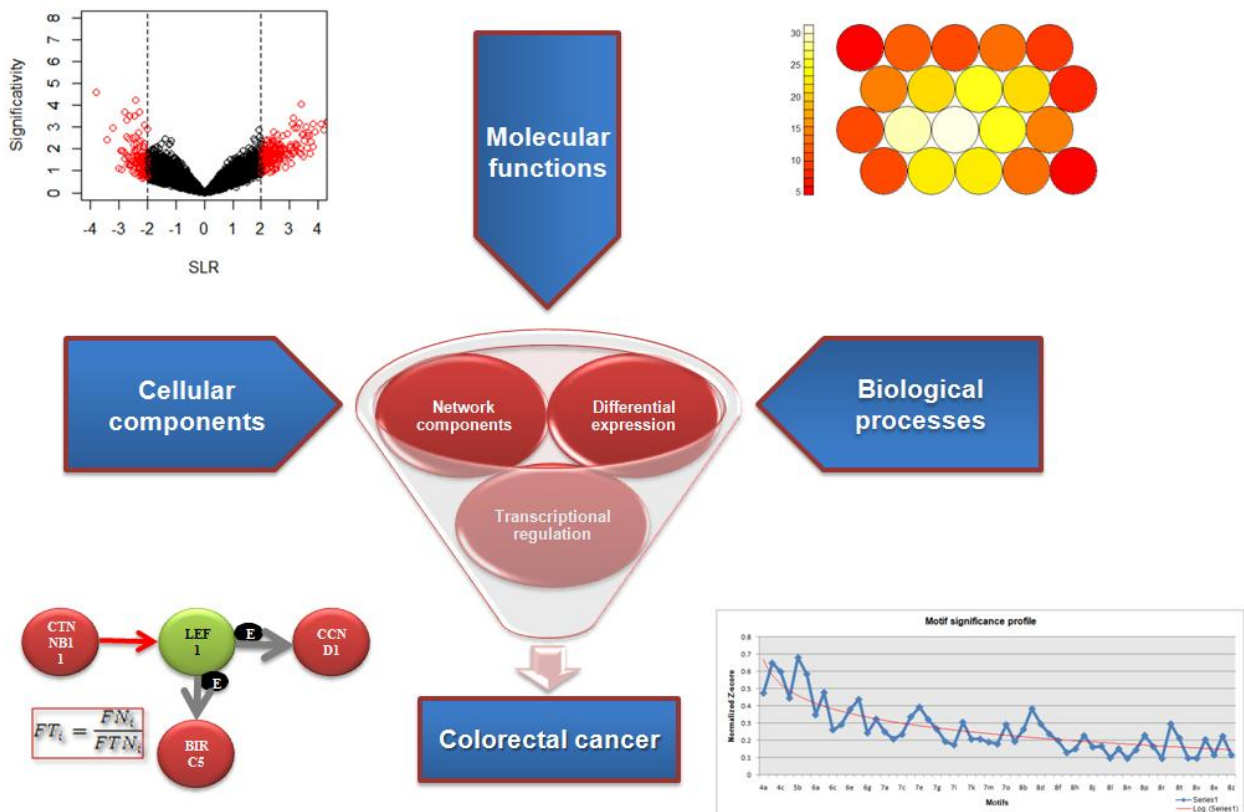
- 
- [22] K. Sugasawa, "Xeroderma pigmentosum genes: functions inside and outside DNA repair," *Carcinogenesis*, vol. 29, pp. 455-465, 2008.
- [23] C. Z. Zhang, Z. P. Chen, C. Q. Xu, T. Ning, D. P. Li, and R. P. Hou, "Correlation of XPD gene with susceptibility to gastric cancer," *Ai zheng= Aizheng= Chinese journal of cancer*, vol. 28, pp. 1163-1167, 2009.
- [24] Z. Chen, C. Zhang, C. Xu, K. Li, R. Hou, D. Li, and X. Cheng, "Effects of selected genetic polymorphisms in xeroderma pigmentosum complementary group D on gastric cancer," *Molecular biology reports*, vol. 38, pp. 1507-1513, 2011.
- [25] I. Rapin, Y. Lindenbaum, D. W. Dickson, K. H. Kraemer, and J. H. Robbins, "Cockayne syndrome and xeroderma pigmentosum DNA repair disorders with overlaps and paradoxes," *Neurology*, vol. 55, pp. 1442-1449, 2000.
- [26] M. Sehgal and T. R. Singh, "Identification and analysis of biomarkers for mismatch repair proteins: A bioinformatic approach," *Journal of natural science, biology, and medicine*, vol. 3, pp. 139-146, 2012.
- [27] J. D. Graves and E. G. Krebs, "Protein phosphorylation and signal transduction," *Pharmacology & therapeutics*, vol. 82, pp. 111-121, 1999.
- [28] X. Wu, S. M. Shell, Z. Yang, and Y. Zou, "Phosphorylation of Nucleotide Excision Repair Factor Xeroderma Pigmentosum Group A by Ataxia Telangiectasia Mutated and Rad3-Related-Dependent Checkpoint Pathway Promotes Cell Survival in Response to UV Irradiation," *Cancer research*, vol. 66, pp. 2997-3005, 2006.
- [29] A. Kumar and T. R. Singh, "A quantitative study of gene regulatory pathways in *Bacillus subtilis* for virulence and competence phenotype by quorum sensing," *Systems and synthetic biology*, vol. 7, pp. 33-39, 2013.
- [30] E. Nilsson, B. Zhang, and M. K. Skinner, "Gene bionetworks that regulate ovarian primordial follicle assembly," *BMC genomics*, vol. 14, p. 496, 2013.
- [31] P. P. Panigrahi and T. R. Singh, "Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: revealed association with aging and other diseases," *Journal of theoretical biology*, vol. 334, pp. 109-121, 2013.
- [32] C. Röhr, M. Kerick, A. Fischer, A. Kühn, K. Kashofer, B. Timmermann, A. Daskalaki, T. Meinel, D. Drichel, S. T. Börno, A. Nowka, S. Krobisch, A. C. McHardy, C. Kratsch, T.

- Becker, A. Wunderlich, C. Barmeyer, C. Viertler, K. Zatloukal, C. Wierling, H. Lehrach, and M. R. Schweiger, "High-throughput miRNA and mRNA sequencing of paired colorectal normal, tumor and metastasis tissues and bioinformatic modeling of miRNA-1 therapeutic applications," *PloS one*, vol. 8, p. e67461, 2013.
- [33] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm," *Nature protocols*, vol. 4, pp. 1073-1081, 2009.
- [34] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature methods*, vol. 7, pp. 248-249, 2010.
- [35] T. R. Singh, A. Gupta, A. Riju, M. Mahalaxmi, A. Seal, and V. Arunachalam, "Computational identification and analysis of single-nucleotide polymorphisms and insertions/deletions in expressed sequence tag data of Eucalyptus," *Journal of genetics*, vol. 90, pp. e34-e38, 2011.
- [36] S. R. Sunyaev, F. Eisenhaber, I. V. Rodchenkov, B. Eisenhaber, V. G. Tumanyan, and E. N. Kuznetsov, "PSIC: profile extraction from sequence alignments with position-specific counts of independent observations," *Protein engineering*, vol. 12, pp. 387-394, 1999.
- [37] H. Y. Yuan, J. J. Chiou, W. H. Tseng, C. H. Liu, C. K. Liu, Y. J. Lin, H. H. Wang, A. Yao, Y. T. Chen, and C. N. Hsu, "FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization," *Nucleic acids research*, vol. 34, pp. W635-W641, 2006.
- [38] A. Z. Dayem Ullah, N. R. Lemoine, and C. Chelala, "SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update)," *Nucleic Acids Research*, vol. 40, pp. W65-W70, 2012.
- [39] L. Conde, J. M. Vaquerizas, H. Dopazo, L. Arbiza, J. Reumers, F. Rousseau, J. Schymkowitz, and J. Dopazo, "PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes," *Nucleic acids research*, vol. 34, pp. W621-W625, 2006.
- [40] G. A. Thorisson, A. V. Smith, L. Krishnan, and L. D. Stein, "The international HapMap project web site," *Genome research*, vol. 15, pp. 1592-1593, 2005.

- 
- [41] N. Blom, S. Gammeltoft, and S. R. Brunak, "Sequence and structure-based prediction of eukaryotic protein phosphorylation sites," *Journal of molecular biology*, vol. 294, pp. 1351-1362, 1999.
- [42] A. Kreegipuu, N. Blom, and S. R. Brunak, "PhosphoBase, a database of phosphorylation sites: release 2.0," *Nucleic acids research*, vol. 27, pp. 237-239, 1999.
- [43] M. Landau, I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures," *Nucleic acids research*, vol. 33, pp. W299-W302, 2005.
- [44] N. Guex and M. C. Peitsch, "SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modeling," *electrophoresis*, vol. 18, pp. 2714-2723, 1997.
- [45] R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: a program to check the stereochemical quality of protein structures," *Journal of applied crystallography*, vol. 26, pp. 283-291, 1993.
- [46] B. Petersen, T. N. Petersen, P. Andersen, M. Nielsen, and C. Lundegaard, "A generic method for assignment of reliability scores applied to solvent accessibility predictions," *BMC Structural Biology*, vol. 9, p. 51, 2009.
- [47] J. A. Marsh and S. A. Teichmann, "Relative solvent accessible surface area predicts protein conformational changes upon binding," *Structure*, vol. 19, pp. 859-867, 2011.
- [48] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-2637, 1983.
- [49] A. Funahashi, M. Morohashi, H. Kitano, and N. Tanimura, "CellDesigner: a process diagram editor for gene-regulatory and biochemical networks," *Biosilico*, vol. 1, pp. 159-162, 2003.

# CHAPTER-4

An integrative approach for mapping differentially expressed genes and network components to elucidate key regulatory genes using novel parameters in colorectal cancer



*“You can’t cross the sea merely by staring at the water.” -Rabindranath Tagore*

## **ABSTRACT**

For examining the intricate biological processes concerned with colorectal cancer (CRC), a systems biology approach integrating several biological components and other influencing factors is essential to understand. We performed a comprehensive system level analysis for CRC which assisted in unravelling crucial network components and many regulatory elements through a coordinated view. Using this integrative approach, the perceptiveness of complexity hidden in biological phenomenon is extensively simplified. The microarray analyses facilitated differential expression of 631 significant genes employed in the progression of disease and supplied interesting associated up and down regulated genes like JUN, FOS and MAPK1. The transcriptional regulation of these genes was deliberated widely by examining diverse transcription factors such as HNF4, NR2F1, ZNF219, and DR1 influencing the expression. Further, interactions of the products for genes were evaluated and crucial network motifs were detected to associate with the pathophysiology of CRC. The available standard statistical parameters such as  $z$ -score,  $p$ -value and significance profile were explored for the identification of key signatures from CRC pathway whereas a few novel parameters exemplifying over-represented structures are also presented in the study. Thus, the applied approach revealed 5 key genes i.e. KRAS, ARAF, PIK3R5, RALGDS and AKT3 via our novel designed parameters illustrating high statistical significance. Further, investigating and targeting these proposed genes for experimental validations, instead being spellbound by the complicated pathway will certainly endow valuable insight in a well-timed systematic understanding of the disease.

## 4.1 INTRODUCTION

In this chapter, an integrated *in silico* analyses was executed for gaining a broad perspective on colorectal cancer (CRC), a DNA repair associated disease. Worldwide, CRC influences millions of people and exists as the most commonly diagnosed cancers along with lung and breast cancer [1]. CRC contributes to second largest cause of death in males and third highest in females, also prevalence of the disorder is observed mostly in economically developed regions [2, 3] probably due to lifestyle and dietary issues. The incidence and mortality rate for CRC is approximately 35-40 percent higher in men as compared to women [4]. As per the United States scenario for 2013, approximately 102,480 people suffered and 50,830 died of CRC which reflects the severity of disease [5]. CRC mainly manifests as abnormal growth of cells occurring at the lining of colon or rectum and the disease progression takes place by replacing a non-cancerous polyp to cancerous tumour. Previous reports [6-8] suggest a variety of factors linked to the disease pattern such as inflammatory bowel disease, polyps, obesity, smoking and genetic history of cancer. Also, the disease is characterized by rectal bleeding, obstruction, abdominal pain, iron-deficiency anaemia, lack of appetite and subsequent weight loss [7, 9]. None of the above symptoms alone can be considered as an assurance for incidence of CRC and often there are no observable symptoms in early CRC. Therefore, appropriate screening for the disease is required [10] to facilitate an early detection and timely removal of polyps [11].

In addition, it is evident that CRC is mainly associated with two important biological pathways i.e. chromosome instability (CIN) pathway [12] and microsatellite instability (MSI) pathway [13, 14]. Genetic aberrations in genes involved in CIN pathway leads to the activation of oncogenes like KRAS and inactivate certain tumor suppressor genes such as SMAD4, P53, SMAD2, BAX and APC that are known to cause the disease [15]. Therefore, investigating up and down regulated important genes may help identify candidate biomarkers for CRC as observed in other studies for different diseases [16]. Moreover, Wheeler et al. [17] and a database on DNA repair genetic association studies as developed in Chapter 2 [18] suggests that mutations in DNA repair genes such as MLH1, MSH2, MSH3, MSH6 implicated in mismatch repair (MMR) system contributes to CRC. In recent decades, many studies on screening, diagnosis and treatment for CRC [19, 20] have conceded but still, there lays a mystery regarding

the genetic and initiation factors accountable for the disease [21]. Also, there is a huge lack of knowledge in mechanisms underlying the progression of CRC from non-cancerous polyp to a tumor and their responsible pathways [22]. Further, a comprehensive perceptive on the genes and related pathways is required for designing specific and effective therapies for CRC [23].

There is already a massive accumulation of gene expression data via DNA microarray studies and several computational techniques have been applied for its analysis. But, the ultimate challenge lies in extracting vital biological information or markers from this amalgamation of data [24]. The microarray technique not only provides a valuable measure for the expression level of numerous genes at once but also offers vital molecular clues regarding mechanisms underlying the pathophysiology of disease [25]. Subsequently, the strategy we pursued includes identification of biologically significant genes and elucidation of key patterns that may govern the functional impact of various biological processes. Each identified gene was then annotated and a search for group of genes sharing a particular characteristic was made. The gene annotation focussed on categorizing the genes by means of three major ontological entities: biological processes, molecular functions and cellular components for their association and involvement in CRC [26].

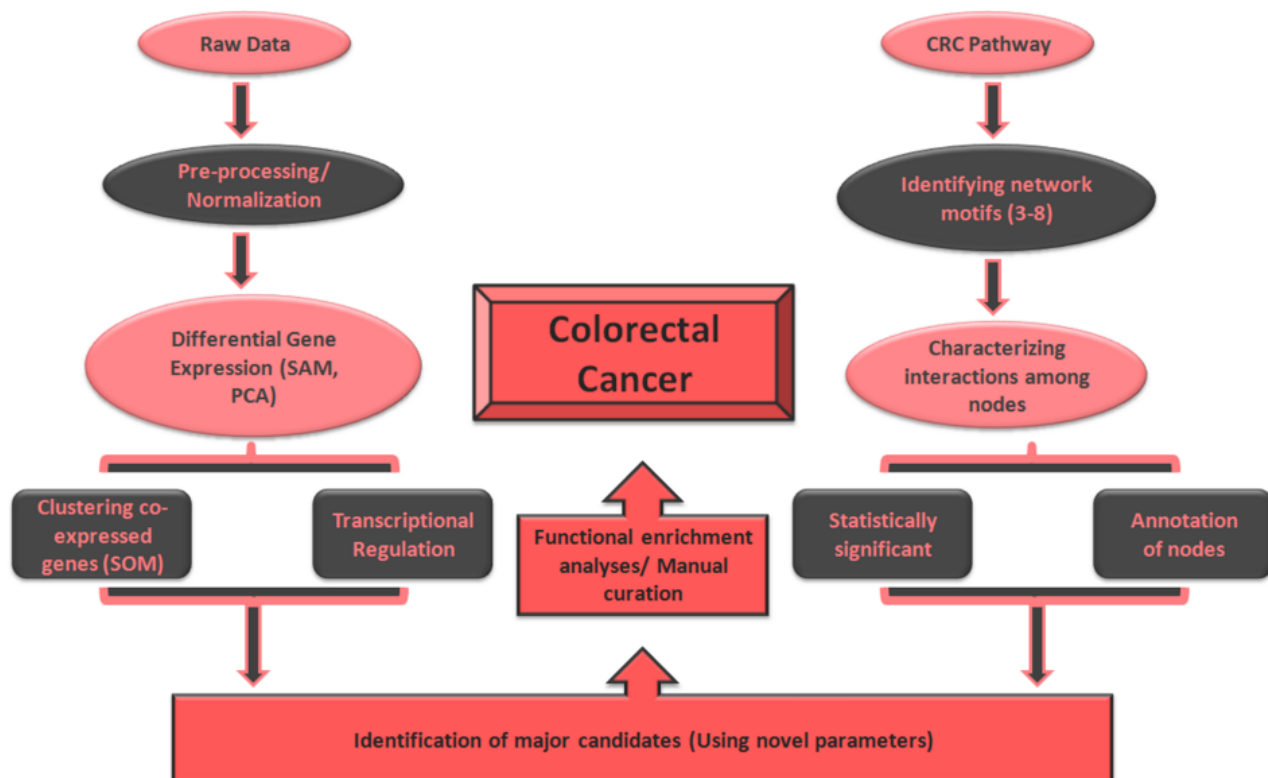
Additionally, an attempt was made to identify vital network components also referred as network motifs, found in elevated frequencies that could generally be expected by chance in a pathway. These network motifs provide statistically overrepresented sub-structures (sub-graphs) in a network and are recognized as simple building blocks of a complicated network. These network motifs play a central role in recognition and analysis of specific patterns in biological networks and yield significant insights into understanding complex biological processes involved in intricate human diseases [27]. We applied computational and statistical criterion for the efficient detection of biological network motifs in CRC and their functional evaluation measures were utilized to reduce the complexity for recognizing best appropriate candidates in the proposed study.

The main perspective of our study was system-component analyses for CRC with several biological components comprising the expression of genes involved, their annotations, and analyses in form of complex network motifs for carrying out a particular function. The foremost

objective was to manually curate and annotate all genes, network components, processes, molecular functions and pathways involved in CRC and then facilitate identification of a few key genes that may serve as vital targets for CRC. On the whole, an integrative approach was practised that includes various aspects of molecular data, networks and pathways for uncovering the intricacy in CRC pathway and then confining the search to only a few genes or network components that may answer diverse biological queries concerning CRC. Also, such *in silico* approach could be applied to other diseases in quest for identifying biomarkers and the study will not only assist experimental biologists, geneticists and other scientific community to identify novel candidate markers for diseases but also has implications for the pharmaceutical industry to target important molecules and design appropriate target based drugs for medications.

## 4.2 MATERIALS AND METHODS

An *in silico* approach has been applied using different forms of raw data, computational tools, software and databases for extensive understanding of mechanisms involved in CRC. A myriad of in-house Perl scripts and various statistical techniques were employed for characterization of candidate markers for the disease. Entire workflow representing different parameters and biological aspects considered for the study is presented in **Figure 4.1**.





---

**Figure 4.1** The workflow deliberated for recognizing candidate markers in colorectal cancer.

#### **4.2.1 Biological data**

The DNA microarray analysis was performed on raw data retrieved from Gene Expression Omnibus (GEO) [28] for the early onset of CRC [29]. The main priority for studying expression of genes at an early stage was to identify markers for early detection of disease which consequently could then be aptly managed. The ultimate goal of the study was to detect additional differentially expressed genes in early onset CRC since the one's involved in familial adenomatous polyposis (FAP) [30] and hereditary non-polyposis colorectal cancer (HNPCC) [31, 32] are already well illustrated. The extracted dataset was then analyzed using GeneChip U133-Plus 2.0 Array. Furthermore, the network motifs for CRC were detected by retrieving biological pathways from KEGG [33], Reactome [34], BioGRID [35] and other pathway databases [36].

#### **4.2.2 Pre-processing of data**

For the identification of differentially expressed genes by DNA microarray analysis, first and the foremost step was pre-processing followed by normalization of raw data which then was subjected to further analysis. The process of data normalization minimizes the effects resulting from technical variations and subsequently permits the data to be compared for determining the actual biological changes. The implementation of data normalization is indispensable in order to stabilize unequal quantities of starting RNA, differences in labelling or detection efficiencies between the used fluorescent dyes and systematic biases in expression levels. Hence, the data congregated from each available CRC disease chip has been normalized using the robust multi average analysis (RMA) algorithm [37].

#### **4.2.3 Identification of differentially expressed genes**

Subsequent to microarray experiments, recognizing genes with increased or decreased expression between the diverse conditions is an imperative and tedious task to perform. This is due to the fact that there are a few conditions, many observations and thousands of hypotheses to be explicitly tested which leads to multiple testing problems. For this purpose, an appropriate statistic has been chosen for testing each gene in the dataset and then the corresponding  $p$ -value was computed. An adjustment process was applied to the resulting raw  $p$ -values to avoid errors

from hypotheses multiplicity [38] and a Quantile-Quantile plot (QQ plot) was generated to plot the observed test statistics against the expected values of test statistics under a combination of null hypotheses. In our study, the expressed genes for control as well as the diseased state were considered for significance analysis of microarrays (SAM) and volcano plot analyses to measure the substantial gap leading to the identification of important regulatory genes [39, 40].

#### **4.2.4 Cluster analysis for co-expressed genes**

The clustering of differentially expressed genes was characterized using hierarchical clustering algorithm. Genes sharing similar expression profiles and other biological features were clustered together and vice-versa. In earlier studies, this kind of classification is achieved for diverse forms of cancers but for CRC, a poor classification has been observed [41]. Moreover, hierarchical clustering was performed in order to deduce the significance of differential expression selection step in classifying the co-regulated genes. Further, for the identification of important patterns in multi-dimensional microarray data, principal component analysis (PCA) was accomplished [42]. This technique facilitated the detection of patterns and aided in analyzing and visualizing genes with similar expression profiles.

#### **4.2.5 Transcriptional regulation of CRC genes**

Since, gene regulation plays a crucial role at the transcriptional level using a variety of transcription factors (TFs) and their target genes; a broad knowledge of transcriptional regulatory elements (REs) is necessary for thorough understanding of gene regulation and underlying complex regulatory processes. Available, *in silico* tools such as DiRE (Distant Regulatory Elements) [43] and oPOSSUM [44] were surveyed for the involved REs among these differentially expressed genes. For a broad perspective on the concerned regulatory process of CRC, these REs were detected that include proximal promoters and distant REs such as enhancers, repressors and silencers.

#### **4.2.6 Functional enrichment for differentially expressed genes**

The enrichment analysis focused on manual curation and annotation via a WEB-based Gene Set AnaLysis Toolkit (WebGestalt) [45] that comprises of genomics, proteomics and large-scale genetic studies for functional annotation of differentially expressed and co-expressed datasets.

This toolkit integrates information from several public resources and often provides accurate and sensitive results aiding in identification of biological processes, their cellular compartments and molecular functions associated with corresponding genes. In addition, GOrilla [46] tool was also explored for detecting the functional characteristics of the gene sets which makes computation on the basis of exact  $p$ -values without simulation analyses. Both the tools make use of same statistical approach i.e. hyper-geometric distribution (HGD) for significance testing and functional enrichment of genes whereas WebGestalt furthermore exploits Fisher's exact test for the annotation analyses. Mathematically, for HGD if there are ' $N$ ' number of genes in a group where ' $A$ ' genes are related to a particular GO term and a sample of ' $n$ ' genes from ' $N$ ' is taken, then the probability of acquiring ' $a$ ' genes associated with ' $a$ ' or more GO terms in a sample ' $n$ ' is deliberated using HGD:

$$p\text{-value} = 1 - \sum_{i=0}^{a-1} f_{HG}(i; N, A, n) = 1 - \sum_{i=0}^{a-1} \frac{\binom{A}{i} \binom{N-A}{n-i}}{\binom{N}{n}}$$

GOrilla displays the statistically significant and enriched genes at the top of ranked gene list and uses a variant of regular HGD named mHG (minimum hypergeometric) for the enrichment analyses of ranked gene lists [47]. In many cases, a fixed threshold ( $n$ ) doesn't work and ranking of all the elements (genes) is required for finding the value of ' $n$ ' that further minimizes HGD. For instance, consider a ranked gene list say  $g_1, \dots, g_N$  in place of a target set, and defined label vector:  $\lambda = \lambda_1, \dots, \lambda_N \in \{0, 1\}^N$  as indicated by the association of ranked genes to a given GO term,  $\lambda_i = 1$  if  $g_i$  is associated with the term [47]. Then, mHG score is given by:

$$mHG(\lambda) = \min_{1 \leq n \leq N} (HGT(N, K, n, k_n(\lambda)))$$

$$\text{Where } k_n(\lambda) = \sum_{i=1}^n \lambda_i$$

Here, the cut-off between top rated genes and rest of the genes is calibrated in a precise manner to maximize the gene enrichment analyses.

#### 4.2.7 Detection of important regulatory patterns in CRC pathway

Examination of vital network motifs, an important aspect to recognize the modularity and to solve large-scale structure of complicated biological networks drawn in CRC was facilitated from complex CRC pathway. A variety of motif detection tools like MFinder [48], MAVisto [49] and FANMOD [50] were employed to identify motifs; where all these tools implement different

algorithms. MFinder uses a semi-dynamic programming algorithm in order to reduce the run time in detecting network motifs and performs full enumeration of the sub-graphs whereas the MAVisto tool employs a flexible algorithm for identification of network motifs and also includes an advanced force-directed layout algorithm [51] for its analyses. Moreover, FANMOD runs a much sophisticated algorithm named RAND-ESU [52] that works on both directed as well as undirected networks for specification and sampling of sub-graphs. This algorithm performs better than its contemporary algorithms [48] for the identification of network motifs from complex biological networks.

The statistical implications of these generated motifs were evaluated using available standard constraints such as  $z$ -scores,  $p$ -values and significance profile (SP). The  $p$ -value and  $z$ -score for each motif was calculated and those having  $z$ -score $>2$  and  $p$ -value $<0.05$  were classified as significant motifs. Further, the SP furnishes normalized  $z$ -score values for a particular network motif ( $m_i$ ) which is given by:

$$SP(m_i) = \frac{z(m_i)}{\sqrt{\sum_{i=1}^n z(m_i)^2}}$$

Where  $Z(m_i)$  corresponds to the  $z$ -score value for each network motif.

All the generated 4-8 node sub-graphs with unique network motif IDs were then extensively analysed for examining proteins and their complex interactions in CRC using our newly designed parameters such as ' $FN_i$ ', ' $FTN_i$ ' and ' $FT_i$ '. Here, ' $FN_i$ ' corresponds to the number of genes present in a given network motif ID; ' $FTN_i$ ' is the sum of frequencies for all the genes occurring in a given network motif ID and ' $FT_i$ ' is defined as the ratio of number of genes for a particular network motif ID and the sum of frequencies for all genes in a given network motif. For a given network motif ID say ' $n_i$ ', where  $i=1,2,3,\dots,n$ ; ' $FT_i$ ' is given by:

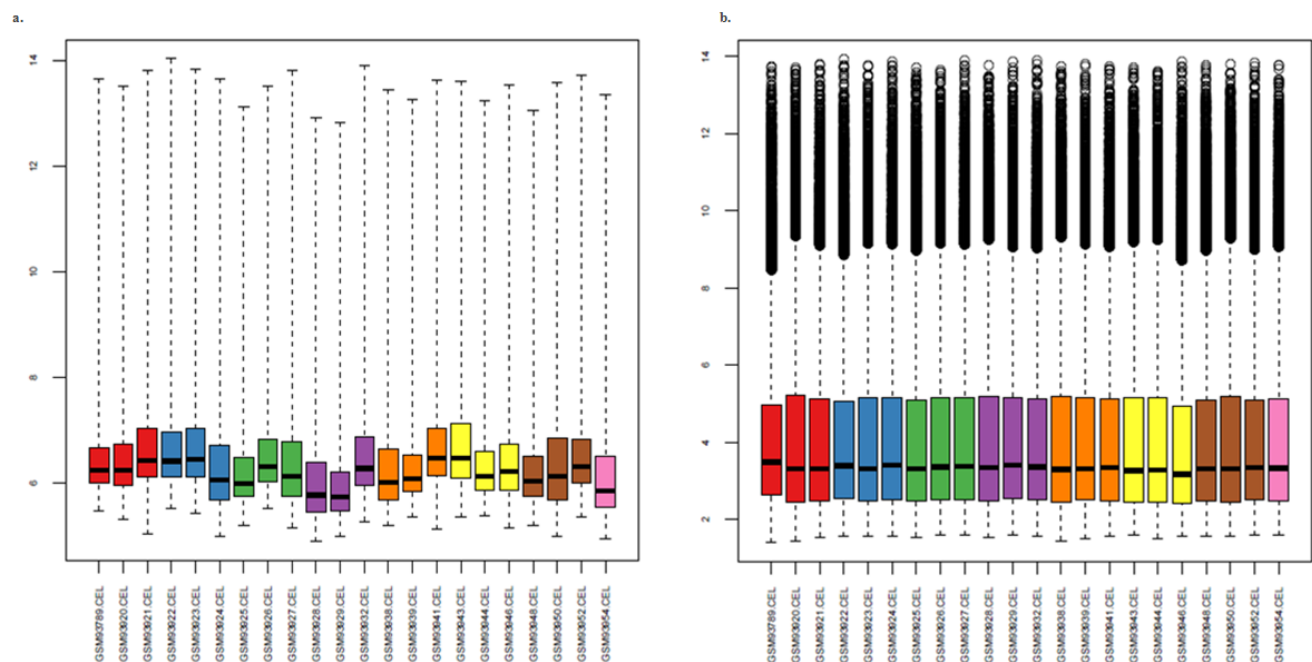
$$FT_i = \frac{FN_i}{FTN_i}$$

Each ' $FT_i$ ' value for a particular network motif ID provides the magnitude of all genes involved in a particular network motif. Thus, the applied methodology comprises of both top-down and bottom-up approaches for detecting the key players in CRC pathway. Using the top-

down approach, first the entire CRC pathway was partitioned into smaller sub-graphs with small functional modules and then the involved nodes were identified and annotated. We attempted to identify motifs ranging from 3-8 node sub-graphs but did not locate 3-node sub-graphs in the CRC pathway. On the other hand, a bottom-up approach was applied for classifying the interactions and relationships among the nodes. Ultimately, outcome from both the approaches were incorporated to identify key nodes in CRC pathway in order to deduce the crucial proteins employed in the progression and regulation of disease.

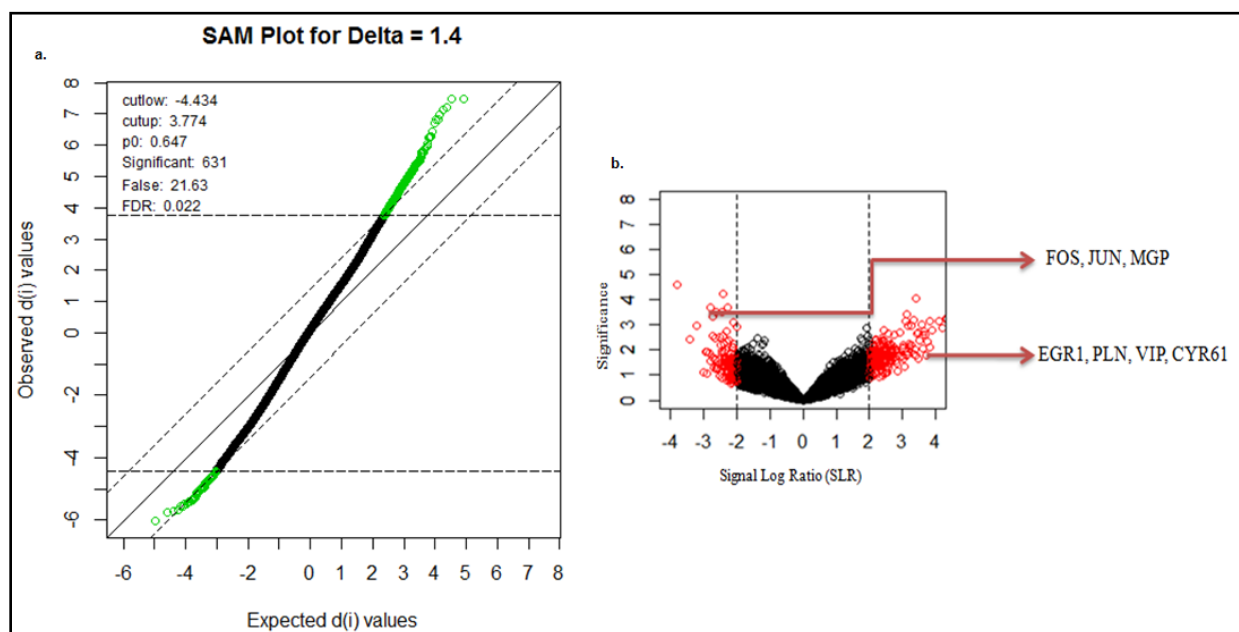
### 4.3 RESULTS AND DISCUSSION

In this study, a comprehensive analysis for differentially expressed genes, TFs, interacting proteins, putative network motifs and their implications in diverse pathways related to CRC has been extensively carried out. Selected CRC dataset for DNA microarray was considered for the process of normalization for removal of errors and noise from the dataset as depicted in **Figure 4.2**. The figure illustrates the box plot for all four affymetrix chips before and after normalization using quantile normalization and clearly demonstrates the impact of normalization step by rectifying the signal of genes across all chips.



**Figure 4.2** Pre-processing and normalization of DNA microarray data; **a.** shows the distribution of the microarray files before normalization; **b.** explains uniform distribution obtained after implementing normalization for removing noise from the data.

The microarray dataset has been examined for identifying specific patterns or markers that may differentiate a normal individual to a diseased one for signifying the susceptibility and facilitate early diagnosis of CRC. After the preliminary pre-processing and manual inspection based on the proportional analysis, final set composed of only the robust candidates. SAM revealed a total of '631' genes (**Figure 4.3a**) from the microarray dataset which were differentially expressed among the tested conditions since points lie aside the diagonal line in a substantial way. The volcano plot between control and the diseased state for CRC clearly elucidated the difference between the genes that were differentially expressed in the two groups as shown in **Figure 4.3b**. Here, the spots represented in black are the genes showing normal expression whereas the red ones with signal log ratio (SLR) $>2$  are over expressed and those with SLR $<-2$  are under expressed genes in the diseased state. Moreover, PCA revealed the projections for 3 different conditions, i.e. over-expressed genes, under-expressed genes and genes showing normal expression which is well described in **Appendix B1 and B2**.



**Figure 4.3** Identification of differential expression via significance analysis of microarray and volcano plot.

After characterizing the differential expression pattern of crucial genes implicated in early CRC progression, role of REs and transcriptional regulation was essential to recognize. We identified a total of '108' TFs in the gene expression dataset for CRC (**Appendix B3**), represented in descending order of their occurrence in the frequency column. Also importance of

these TFs were calculated using an optimization procedure that considers a weight ' $w_i$ ' for each  $i^{th}$  TF, as a measure of its association with the input gene set and further calculates the importance value as the product of TF occurrence (frequency) and TF weight. We also classified TFs (see **Appendix B4**) found in each differentially expressed gene from CRC dataset, providing total number of TFs for each gene, locus, their names, position and their associated types. Moreover, families for all the important TFs have been recognized and illustrated in **Appendix B5**. We also compiled a list for top 10 TFs implicated in genes responsible for differential expression in early CRC with their frequencies of occurrence, importance and other essential details as depicted in **Table 4.1**.

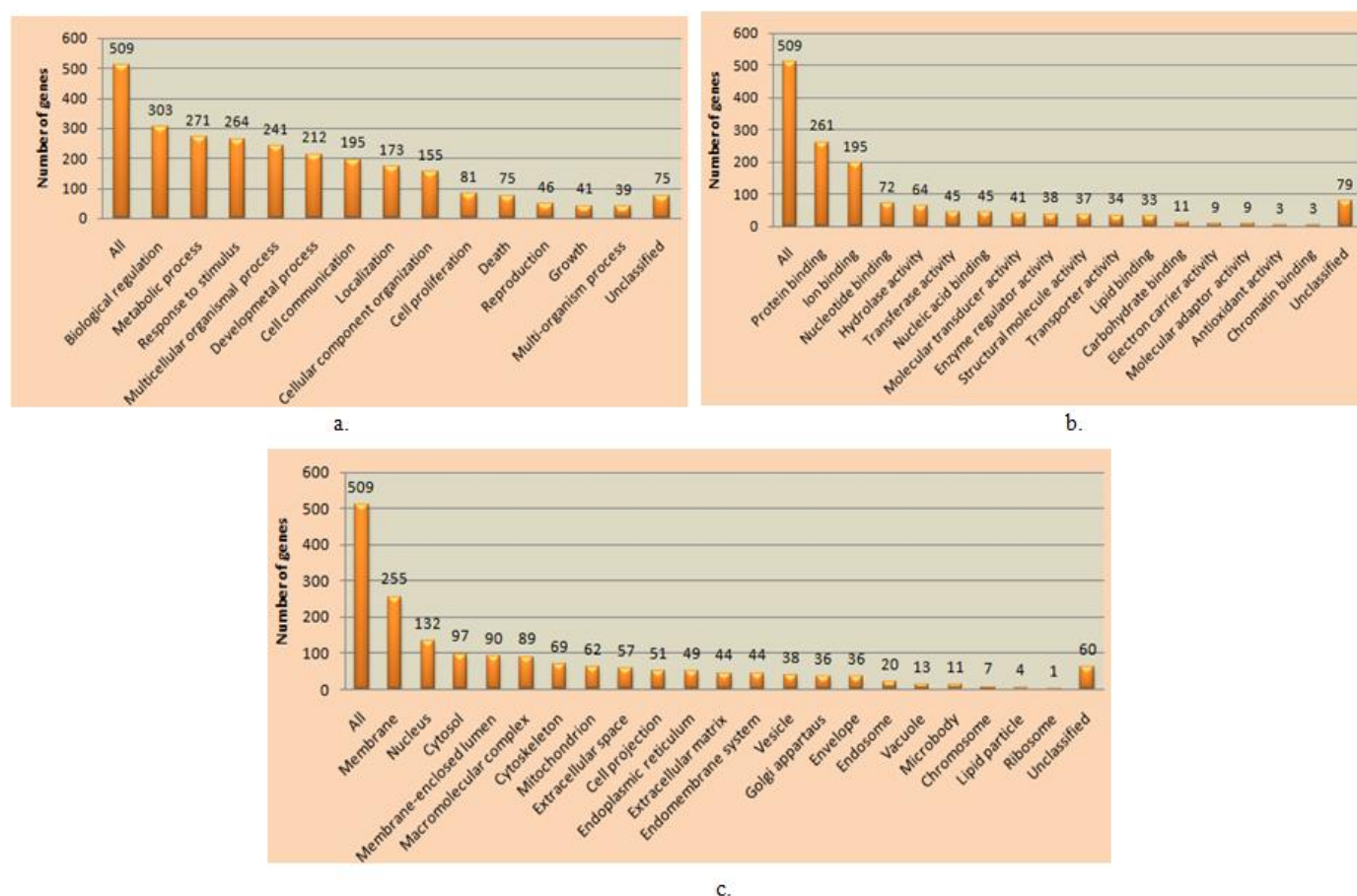
**Table 4.1 Major transcription factors identified in early colorectal cancer progression**

Transcription Factor	Frequency	Importance	JASPAR ID*	Class	Family
HNF4	31.80%	0.31802	MA0114.1	Zinc-coordinating	Hormone-nuclear Receptor
NR2F1	19.43%	0.50044	MA0017.1	Zinc-coordinating	Hormone-nuclear Receptor
DR1	17.31%	0.04112	-	Zinc-coordinating	Hormone-nuclear Receptor
PPARG	14.49%	0.03622	MA0066.1	Zinc-coordinating	Hormone-nuclear Receptor
HNF1	14.13%	0.36064	MA0046.1, MA0153.1	Helix-Turn-Helix	Homeo
HNF4_DR1	13.78%	0.16882	-	Zinc-coordinating	Hormone-nuclear Receptor
PPAR_DR1	13.43%	0.13428	-	Zinc-coordinating	Hormone-nuclear Receptor
HNF4ALPHA	12.01%	0.29848	MA0114.1	Zinc-coordinating	Hormone-nuclear Receptor
PAX4	12.01%	0.18322	MA0068.1	Helix-Turn-Helix	Homeo
ER	10.60%	0.08216	MA0112.2, MA0258.1	Zinc-coordinating	Hormone-nuclear Receptor

The majority of identified TFs belonged to zinc-coordinating class and hormone-nuclear receptor family of transcriptional regulatory system. Hepatocyte nuclear factor 4 (HNF4), nuclear receptor subfamily 2 group F member 1 (NR2F1) and down-regulator of transcription 1 (DR1) are the most frequent TFs regulating the genes in early CRC dataset and are the members of same class as well as family of TFs. All these TFs either bind directly or in the form of a complex to control the rate of transcription. This kind of information is primarily required to understand the gene regulation in a comprehensive manner. It is anticipated that for the

\* The JASPAR IDs correspond to the transcription factors from standard JASPAR database

regulation of genes involved in CRC, manipulation of regulatory region of genes specifically for the identified important TFs such as HNF4, NR2F1, DR1, and their classes could provide biological insight to experimental biologists and geneticists. Further, an attempt was made to annotate and manually curate the genes for their biological roles, functions, cellular components and their implication in diverse complex biological pathways. Out of ‘631’ differentially expressed genes, functional enrichment for ‘509’ genes was engendered. Maximum number of genes had vital roles in biological regulation, protein binding and were present at membranes of the cell (**Figure 4.4**). This particular section of the chapter provides an insight to diverse mechanisms and pathways elucidated by the regulation of genes involved in CRC pathway.

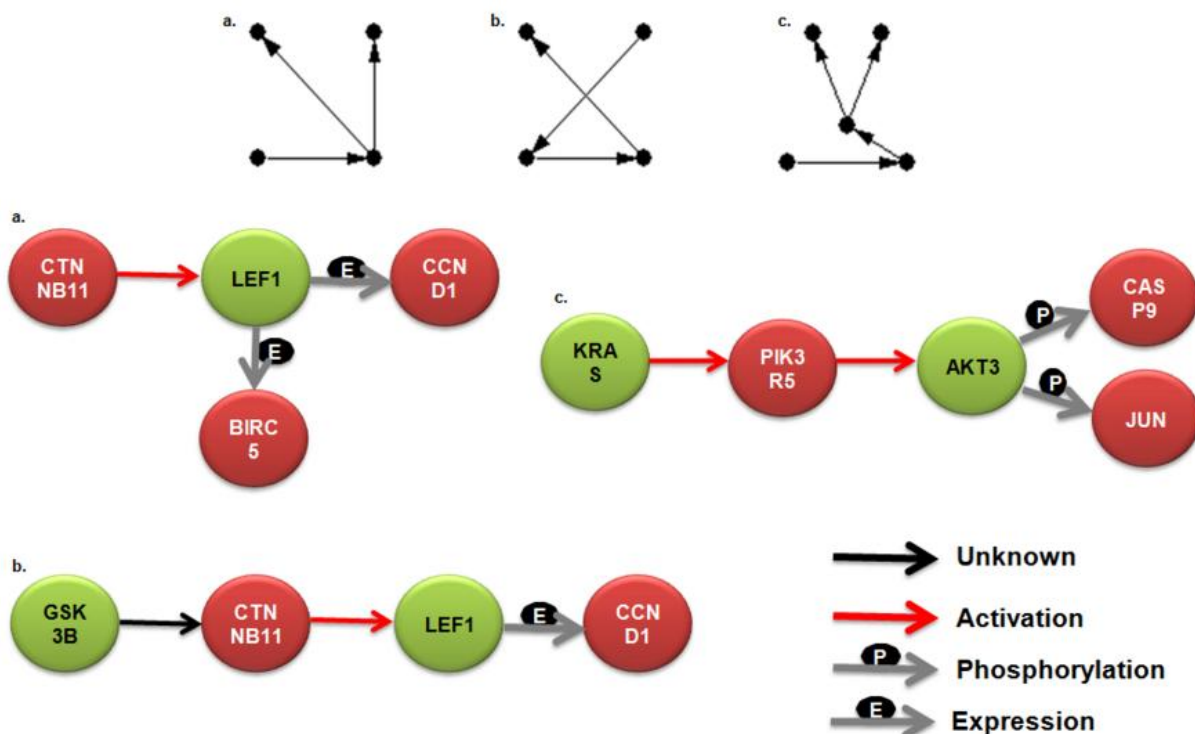


**Figure 4.4** Functional enrichment and annotation analyses for differentially expressed genes.

After acquiring the differential expression pattern, we were intended to identify chief sub-networks configured by these genes that may facilitate the annotation of intricate biological network implicated in CRC. Based on the rationale, detection of crucial network motifs and



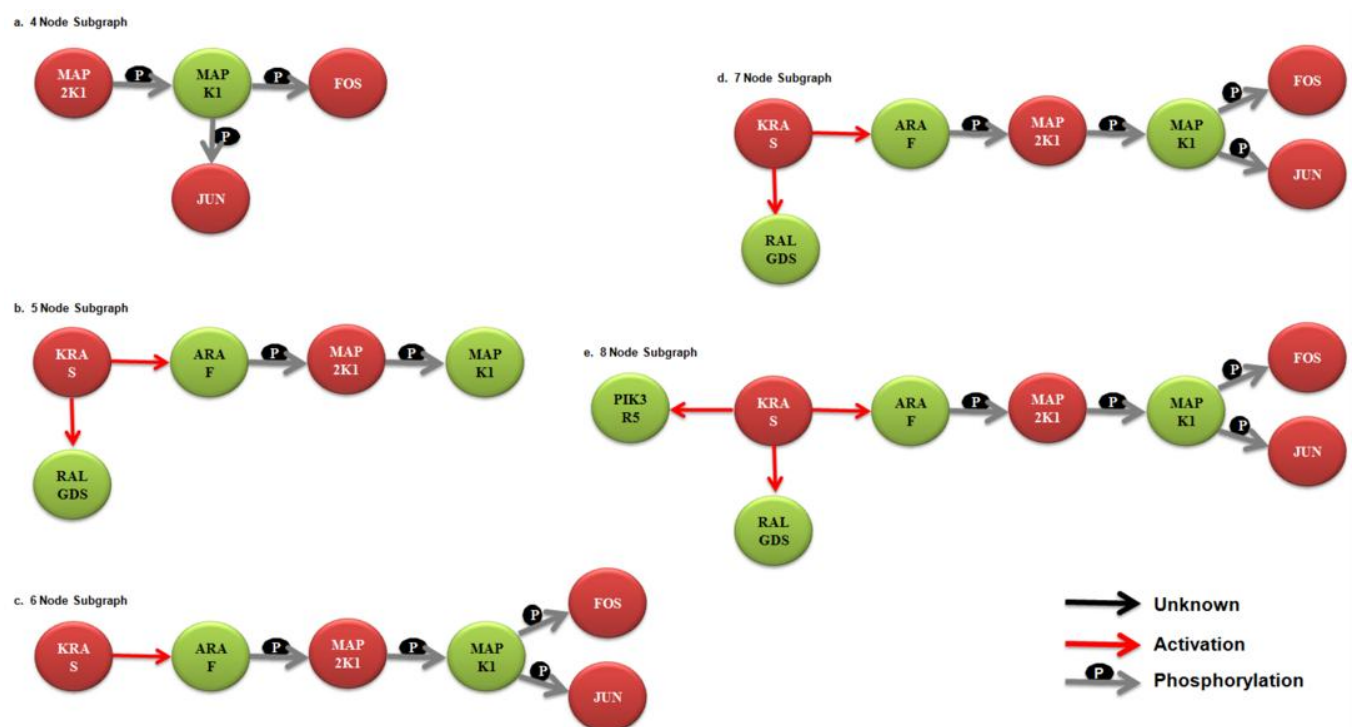
network patterns was made; providing essential clues concerning the hierarchical decomposition of CRC network. Here the patterns being referred are small connected sub-networks occurring in significantly higher frequencies in a network than would be expected for a given random network. These patterns or motifs are considerably overrepresented and characterize certain essential functional aspects associated with CRC related pathways and its progression. Several motifs ranging from 4-8 sub-graph nodes were generated and annotated for the CRC pathway which is available as supplementary data (available at: <http://www.bioinfoindia.org/CRCData>), and a few have been depicted in **Figure 4.5**. The applied bottom-up approach is clearly demonstrated in **Figure 4.6** starting from 4-node sub-graphs and then proceeding one by one till 8-node sub-graphs were generated; all the interacting genes were annotated along with their functional relationships.



**Figure 4.5** Identified vital patterns (network motifs) from colorectal cancer pathway.

The network motifs thus obtained from CRC pathway contained 4-chain motifs, single input module (SIM), multiple input module (MIM), bifan motifs and other important biological signatures that were supported by significant  $z$ -scores and  $p$ -values for their statistical relevance.

These network motifs were then subjected to further annotation and disease-specific analyses since, they have important functions to execute; as in case of SIM motif, several genes are controlled by a single master gene and the master gene is known to be autoregulatory. Whereas, in MIM motif (a generalization of SIM), a single gene is being controlled by multiple genes. Also, other regular 4-node motifs were detected that confirmed the presence of diamond, biparallel and bifan motifs i.e. often built by two regulatory and two regulated genes. Further, these nodes were annotated for their biological significance using in-house Perl scripts for identifying genes involved in these patterns.



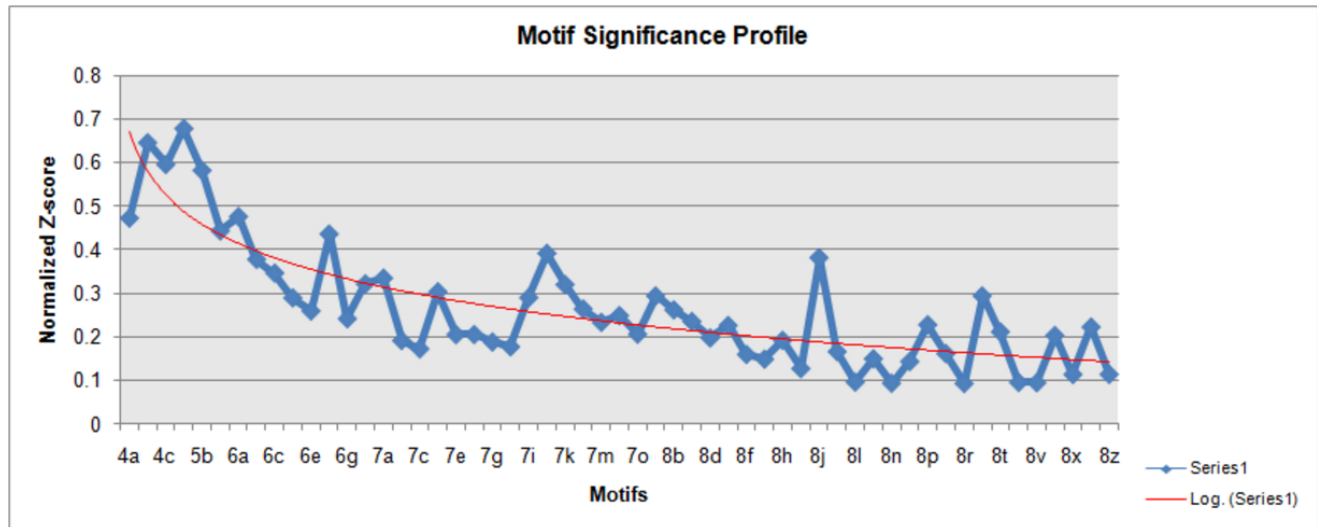
**Figure 4.6** Bottom-up approach for classifying the network motifs.

Similar type of motif graphs were generated for the sub-networks of other node sizes and annotation of these graphs were based on statistical criterion using mean-frequencies, standard deviation,  $z$ -scores and  $p$ -values as demonstrated in **Table 4.2**. The SP thus calculated could be superlatively observed by plotting it on a graph against the different motifs as illustrated in **Figure 4.7**. The motif SP graph depicts that when the number of nodes in a motif increase, the complexity increases and further the trend declines representing smaller normalized  $z$ -score values towards large motif sizes.

Table 4.2 Network motifs with their respective standard statistical parameters

Network motif ID (Adjacency matrix)	z-score	p-value	Significance Profile
000001000011000	2.2224	0.009	0.474
0000000000011100	3.0346	0.001	0.647
0000000000001110	2.8002	0.006	0.597
000000000000100100010100	2.2899	0.011	0.445
0000010000010000000101000	3.5006	0.001	0.680
000000000000010000111000	3.0053	0.001	0.583
00000000000001000000010100000001100	3.235	0.01	0.348
00000000000000010000001000001110000	4.4425	0.002	0.477
0000000000000000000000010011000100100	2.4236	0.025	0.260
000000000000000000010000010000001110000	2.7065	0.013	0.291
000000000000000000000000010001000110100	3.5275	0.003	0.379
000000100000000010000001010000010000	4.0716	0.002	0.438
000000100000010000000001010000000010	2.2603	0.028	0.243
0000000000000000000000001000001111000	3.0051	0.007	0.323
00000000000000000000000001000000001001000001001100	3.9553	0.007	0.250
00000000000000000000010000000100000001001000001010000	3.2784	0.011	0.207
00000000000000000000000100010000000001010000000011000	3.7112	0.008	0.235
0000000000000000000000000100000001000100001101000	5.3149	0.001	0.336
0000000000000000000001000000010100000001000000011000	6.2161	0.001	0.393
0000000000000000000001000000010001000000001001010000	5.08	0.005	0.321
000000000000000000000000010000000001011000000001100	4.1933	0.005	0.265
0000000000000000000000000000000000000001000110001100100	3.0485	0.011	0.193
00000000000000000000010000000010000000100000011100000	2.7445	0.02	0.174
0000000000000000000001000000010000000100000011100000	4.8218	0.002	0.305
0000000100000000000100000001010000001000000000100	3.2633	0.016	0.206
000000010000000100000000010010000000000010000100	3.2667	0.016	0.207
000000000000000000000000000000000000000100011001110000	2.9992	0.011	0.190
0000000000000000000000000000000001000001000010001110000	2.8224	0.019	0.178
00000000000000000000000000000100000100010001110000	4.6079	0.002	0.291
00000000000000000000000000000100001000000000010010000010011000	4.5789	0.003	0.193
0000000000000000000000000000010000000010001000000100000010011000	6.2385	0.001	0.263
00000000000000000000010000000010000100000000000100000100010100000	9.0783	0.001	0.383
00000000000000000000000000000001000000010000100000110000010000100	6.9995	0.001	0.295
0000000000000000000000000000010100000000000100001100000000010100	5.6065	0.004	0.237
00000000000000000000000000000000010000000000100110000010001100	4.7134	0.003	0.199
00000000000000000000010000000010000000010000000100100000010100000	3.0432	0.013	0.128
0000000000000000000001000000001000010000000000010000000010100000	3.5566	0.011	0.150
0000000000000000000000000000000001000000000010000010001100000000010100	5.3903	0.006	0.227
0000000000000000000000000000000001000000001000100000110000010000100	3.817	0.013	0.161
000000000000000000000000000000000100010000001000000100000000011100	3.9599	0.01	0.167
0000000000000000000000000000000001000000000010100000000100010000011000	2.3068	0.036	0.097
000000000000000000000000000000000100000000100000000100010000011010000	3.5739	0.011	0.151
00000000000000000000000000000000010000000010000100000000000111010000	2.2481	0.028	0.095
000000000000000000000000000000000100000000100000000010000000111000000	3.4235	0.016	0.144
0000000010000000010000000000010001000000000000100000000100001000	5.4301	0.004	0.229
0000000000000000000001000000001000010000000000010000000111000000	3.8602	0.01	0.163
0000000000000000000000000000000001000000001000000100000000000111000000	2.2299	0.036	0.094
0000000000000000000000000000000001000000001000000000000000010011010000	7.003	0.001	0.295
0000000010000000000010000000000101000000010000000000010000000010	5.0413	0.005	0.213
000000000000000000000100000000100000000100010000000100000010100000	2.292	0.038	0.097
0000000000000000000001000001000000000000101000000000100010100000	2.2751	0.038	0.096
0000000000000000000000000000000000000000000000000000000000010011100000	4.8459	0.003	0.204
00000000000000000000000000000000000000000000000000000000000100011100001000100	2.7287	0.02	0.115
0000000000000000000000000000000000000000000000000000000000010000000010000000	5.3047	0.001	0.224
0000000000000000000000000000000000000000000000000000000000010000000010000000111100000	2.7255	0.026	0.115

Therefore, our approach of reducing the entire CRC pathway into smaller sub-graphs and subsequently identifying key players proves quite valuable. Based upon this SP analysis we suggest that network motifs with smaller node size (3 or 4) are more functionally allied towards their role in pathways while motifs of larger size ( $\geq 5$  nodes) are less functional (**Figure 4.7**). It is believed that the observed trend might be similar in many such biological networks if analyzed.



**Figure 4.7** Significance profile for all 4-8 node generated sub-graphs based on normalized z-scores; the motif significance profile evidently exemplifies that when the complexity in CRC pathway increases, the interactions among the nodes and intricacy in recognition of genes amplifies immensely. Lesser the node size, it becomes easy to annotate the nodes (genes) and their associations with stronger statistical significance (greater normalized z-scores).

From the novel deliberated parameters depicted in **Table 4.3**, it was observed; the lower ' $FT_i$ ' value proves to be more statistically significant as it signified greater involvement of a few genes that explains complex interactions among different nodes in a given motif. Further, the motif showing least ' $FT_i$ ' value i.e. 0.171 for motif ID '7n' was chosen for identifying the key players in the given motif. This information was attained by mapping all the genes from the complex CRC pathway onto the network motifs and then frequency of each gene for each network motif was calculated. Then the frequencies for all genes in the above mentioned motif (with least ' $FT_i$ ' value) were calculated and presented in **Table 4.4**.

Table 4.3 Values of the designed parameters for each recurrent motif in the CRC pathway

Network Motif Image ID (Adjacency matrix)	Abbreviation	$FTN_i$	$FN_i$	$FT_i$
'000001000011000'	4a	76	25	0.329
'000000000011100'	4b	48	16	0.333
'000000000001110'	4c	16	16	1
'000001000001000000101000'	5a	30	8	0.267
'000000000000010000111000'	5b	15	6	0.4
'0000000000001001000010100'	5c	60	14	0.233
'00000000000000010000001010000'	6a	36	8	0.222
'000000000000000000000010001000110100'	6b	36	14	0.389
'00000000000010000000010100000001100'	6c	60	12	0.2
'00000000000001000000100000111000'	6d	36	14	0.389
'000000000000000000000010011000100100'	6e	36	8	0.222
'000000100000000010000001010000010000'	6f	18	8	0.444
'000000100000010000000001010000000010'	6g	18	8	0.444
'00000000000000000000001000001111000'	6h	6	6	1
'0000000000000000000000010000001000100001101000'	7a	49	18	0.367
'00000000000000000000000000000000000000001000110001100100'	7b	21	8	0.381
'0000000000000000010000000010000000100000011100000'	7c	21	8	0.381
'000000000000000000000001000000010000000100000011100000'	7d	21	8	0.381
'0000000100000000000100000001010000001000000000100'	7e	21	9	0.429
'00000001000000010000000001001000000000010000100'	7f	21	9	0.429
'0000000000000000000000000000000000000000100011001110000'	7g	14	8	0.571
'0000000000000000000000000000000000000000100001000010001110000'	7h	14	14	1
'00000000000000000000000000000000000000001000000100010001110000'	7i	14	8	0.571
'000000000000000000000001000000010100000001000000011000'	7j	49	12	0.245
'0000000000000000000100000010001000000001001010000'	7k	42	10	0.238
'00000000000000000000000000000000000000001011000000001100'	7l	42	10	0.238
'00000000000000000000000100010000000001010000000011000'	7m	56	13	0.232
'00000000000000000000000000000000000000001001000001001100'	<b>7n</b>	<b>70</b>	<b>12</b>	<b>0.171</b>
'000000000000000001000000100000001001000001010000'	7o	50	13	0.26
'000000000000000000000000000000000000000010000001000010000011000001000100'	8a	48	10	0.208
'0000000000000000000000000000000000000000100000001000100000100000010011000'	8b	56	12	0.214
'0000000000000000000000000000000000000000101000000000010000110000000010100'	8c	48	12	0.25
'0000000000000000000000000000000000000000100000000000100110000010001100'	8d	48	10	0.208
'000000000000000000000000000000000000000010000000001000001000110000000010100'	8e	48	11	0.229
'0000000000000000000000000000000000000000100000001000100000110000010000100'	8f	48	10	0.208
'000000000000000000000000000000000000000010000000100010000000000000000000110100000'	8g	48	11	0.229
'0000000000000000000000000000000000000000100001000000000001000000010011000'	8h	64	13	0.203
'000000000000000000000000000000000000000010000000010000000010010000010100000'	8i	48	11	0.229
'00000000000000000000000000000000000000001000000000000100000100010100000'	8j	48	11	0.229
'00000000000000000000000000000000000000001000100000010000001000000000011100'	8k	40	12	0.3
'0000000000000000000000000000000000000000100000000001010000000100010000011000'	8l	32	11	0.344
'00000000000000000000000000000000000000001000000010000000100010000011010000'	8m	29	13	0.448
'000000000000000000000000000000000000000010000000010000100000000000000000111010000'	8n	24	10	0.417
'00000000000000000000000000000000000000001000000010000000010000000111000000'	8o	24	9	0.375
'000000001000000001000000000001000100000000000100000000100000000100001000'	8p	24	10	0.417
'0000000000000000000000000000000000000000100001000000000010000000111000000'	8q	24	9	0.375
'00000000000000000000000000000000000000001000000100000100000000000111000000'	8r	24	9	0.375
'0000000000000000000000000000000000000000100000001000010000000000010011010000'	8s	24	10	0.417
'00000000100000000000100000000001010000000100000000000100000000010'	8t	24	10	0.417
'000000000000000000000000000000000000000010000000010000000100000010100000'	8u	16	9	0.563
'00000000000000000000000000000000000000001000001000000000001000010100000'	8v	16	9	0.563
'00000000000000000000000000000000000000001000100000000110011100000'	8w	16	10	0.625
'000000000000000000000000000000000000000010000000000000000000000000000000100011100011000100'	8x	8	8	1
'000000000000000000000000000000000000000010000000010000000010000000111100000'	8y	8	8	1
'000000000000000000000000000000000000000010000000010000000010000000111100000'	8z	8	8	1

**Table 4.4 Putative over-represented genes from CRC pathway as indicated by the most recurrent network motif**

S. No.	Genes/ Proteins	Gene/Protein Details	Gene Size	Frequency	Molecular Functions	Pubmed IDs*
1	KRAS	Kirsten rat sarcoma viral oncogene homolog	21656 Da, 189 amino acids	10	GTPase activity, LRR domain binding, protein binding	19515263, 15069679, 19832985
2	ARAF	V-raf murine sarcoma 3611 viral oncogene homolog	67585 Da, 606 amino acids	10	Protein kinase activity, protein binding, ATP binding, transferase activity, metal ion binding	<b>20145135</b>
3	PIK3R5	Phosphoinositide-3-kinase, regulatory subunit 5	97348 Da, 880 amino acids	10	G-protein beta/gamma-subunit complex binding, 1-phosphatidylinositol-3-kinase regulator activity	-
4	RALGDS	Ral guanine nucleotide dissociation stimulator	100607 Da, 914 amino acids	10	Small GTPase regulator activity, protein binding, guanyl-nucleotide exchange factor activity	<b>15766656, 17568777</b>
5	AKT3	V-akt murine thymoma viral oncogene homolog 3	55775 Da, 479 amino acids	8	Protein kinase activity, ATP binding, protein binding, transferase activity	18813315
6	RHOA	Ras homolog family member A	21768 Da, 193 amino acids	6	GTPase activity, protein binding, myosin binding, protein domain specific binding	19374769, 11844789, 11953197, 19499974
7	MAP2K1	Mitogen-activated protein kinase kinase 1	43439 Da, 393 amino acids	6	Protein kinase activity, ATP binding, protein binding, transferase activity, RAS GTPase binding	<b>17667937</b>
8	MAPK1	Mitogen-activated protein kinase 1	41390 Da, 360 amino acids	2	Phosphotyrosine binding, DNA binding, protein kinase activity, transferase activity, ATP binding, transcription factor binding	9690379, 11992399
9	GSK3B	Glycogen synthase kinase 3 beta	46744 Da, 420 amino acids	2	Protein kinase activity, beta-catenin binding, tau protein binding, transferase activity, p53 binding, NF-kappaB binding	<b>17640304</b>
10	BAD	BCL2-associated agonist of cell death	18392 Da, 168 amino acids	2	Protein binding, phospholipid binding, protein heterodimerization activity, protein kinase binding, protein phosphatase binding	17583570, 17393317
11	CASP9	Caspase 9, apoptosis-related cysteine peptidase	46281 Da, 416 amino acids	2	Cysteine-type endopeptidase activity, enzyme activator activity, protein binding, peptidase activity, SH3 domain binding, protein kinase binding	11912124, 23303631
12	MAPK8	Mitogen-activated protein kinase 8	48296 Da, 427 amino acids	2	Catalytic activity, JUN kinase activity, MAP kinase activity, protein kinase activity, ATP binding, phosphotransferase activity, transferase activity, histone deacetylase binding	<b>19352384, 12819185</b>

\* Pubmed IDs correspond to the published literature illustrating role of these genes in colorectal cancer, whereas for some genes, experimental evidences were not found and a few depicted in bold explains their occurrence in colon cancer and further their role in colorectal cancer may be confirmed.

Analyzing complex biological pathway of CRC is a convoluted process and requires an integrative approach for identifying biomarkers for the disease. Thus, the approach we applied not only performs enrichment analyses but also presents observations from many different methods, applications and tools existing for gene expression and network data analyses. The current study intended for identification of vital components in pursuit of reducing the complexity hidden in intricate CRC pathway and their associated biological processes. Identification of crucial network motifs will help systems biologists to find key components from whole pathways and analyze their behaviour against different experimental conditions. Although genes involved in MMR system like MLH1, MSH2, MSH6, PMS2 and other genes such as APC and MUTYH have already shown their influence on CRC but still cause and progression of the disease remains unrequited. Consequently, we made an effort to identify certain other genes that may potentially impact meticulous understanding of CRC. Many important genes as revealed in **Table 4.4** like kirsten rat sarcoma viral oncogene homolog (KRAS), v-raf murine sarcoma 3611 viral oncogene homolog (ARAF), phosphoinositide-3-kinase, regulatory subunit 5 (PIK3R5), ral guanine nucleotide dissociation stimulator (RALGDS) and v-akt murine thymoma viral oncogene homolog 3 (AKT3) were observed to contribute maximum complexity in the CRC pathway. These genes illustrated higher frequencies and numerous interactions among nodes and are proposed to be vital for CRC disease progression. Here, the CRC pathway complexity has been reduced to a few key genes that may be explored further for their putative roles in the disease.

Previous reports suggest that the mutational analyses of KRAS and BRAF are highly correlated with the development of CRC by activating MAP kinase pathway [53]. The BRAF gene, an isoform of ARAF (suggested from the pathway level analysis) also has its influence on a number of tumors especially in colorectal and gastric cancer whereas role of ARAF still remains a mystery [54]. Although there have been contradictory results reported earlier [55] which states that mutations in ARAF gene may not be associated with pathogenesis of various human cancers. But we found 97% similarity among the two protein sequences (ARAF and BRAF) and the two isoforms share several domains such as Raf\_RBD, Pkinase, SPS1, TyrKc and biological properties including binding sites; so intending ARAF as one of the key genes in CRC for its association in disease may prove vital for understanding cancer genetics.

FBJ murine osteosarcoma viral oncogene homolog (FOS) and jun proto-oncogene (JUN) with ample frequencies are identified in network motifs as well as in the differential expression dataset depicting their putative crucial roles in forming the convoluted CRC pathway (**Figure 4.5-4.6**). As deciphered in the Figures, these genes demonstrate vital interactions among themselves and other genes focussing on activating certain genes, phosphorylating and affecting expression of other genes.

#### **4.4 CONCLUSION**

The study reveals important markers and a few novel genes and its variants that are believed to associate with CRC and its progression. The 5 genes reported in the study namely, KRAS, ARAF, PIK3R5, RALGDS and AKT3 along with 2 other genes JUN and FOS can be studied broadly for its association in CRC since, the former genes illustrated complex associations and latter signified high differential expression in diseased state. Moreover, the anticipated genes, JUN, FOS, MAPK1 and their REs ZNF219, HNF4, PPARG and DR1 could be utilized further to control the transcriptional regulation and other regulatory actions executed by these genes. The identified genes from early progression dataset and network analyses for CRC may be explored further and experimentally tested to reveal crucial insights in understanding the disease in an extensive mode. The novel parameters designed present the dependence of an entire system on a few key genes, proteins and metabolites for examining the statistical significance. Hence, the 5 proposed genes from comprehensive theoretical and computational analysis implicated in CRC may serve as imperative therapeutic targets for CRC. There is an imperative need to apply this approach on other diseases as well to identify crucial network components and candidate markers. It is believed that besides key genes proposed in this study, we provide novel methodology to analyze small components of large and complex biological networks. Further, investigating and targeting these proposed genes for experimental validations, instead being spellbound by the complicated pathway will certainly endow valuable insight in a well-timed systematic understanding of the disease.

#### **REFERENCES**

- [1] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "GLOBOCAN 2012 v1.0, Cancer Incidence and



- Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer..” vol. 2014, 2013.
- [2] M. M. Center, A. Jemal, R. A. Smith, and E. Ward, "Worldwide variations in colorectal cancer," *CA Cancer J Clin*, vol. 59, pp. 366-378, 2009.
- [3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA Cancer J Clin*, vol. 61, pp. 69-90, 2011.
- [4] D. Cunningham, W. Atkin, H. J. Lenz, H. T. Lynch, B. Minsky, B. Nordlinger, and N. Starling, "Colorectal cancer," *Lancet*, vol. 375, pp. 1030-1047, 2010.
- [5] A. C. Society, "Cancer Facts & Figures Atlanta: American Cancer Society 2013." vol. 2014, 2013.
- [6] A. J. M. Watson and P. D. Collins, "Colon cancer: a civilization disorder," *Dig Dis*, vol. 29, pp. 222-228, 2011.
- [7] P. Ferrari, M. Jenab, T. Norat, A. Moskal, N. Slimani, A. Olsen, A. Tjonneland, K. Overvad, M. K. Jensen, M. C. Boutron-Ruault, et al., "Lifetime and baseline alcohol intake and risk of colon and rectal cancers in the European prospective investigation into cancer and nutrition (EPIC)," *Int J Cancer*, vol. 121, pp. 2065-2072, 2007.
- [8] N. Jawad, N. Direkze, and S. J. Leedham, "Inflammatory bowel disease and colon cancer," *Recent Results Cancer Res*, vol. 185, pp. 99-115, 2011.
- [9] M. Astin, T. Griffin, R. D. Neal, P. Rose, and W. Hamilton, "The diagnostic value of symptoms for colorectal cancer in primary care: a systematic review," *Br J Gen Pract*, vol. 61, pp. 231-243, 2011.
- [10] B. K. Edwards, E. Ward, B. A. Kohler, C. Ehemann, A. G. Zauber, R. N. Anderson, A. Jemal, M. J. Schymura, I. Lansdorp-Vogelaar, L. C. Seeff, et al., "Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates," *Cancer*, vol. 116, pp. 544-573, 2010.
- [11] P. Boyle and J. S. Langman, "ABC of colorectal cancer: Epidemiology," *British Medical Journal*, vol. 321, pp. 805-808, 2000.
- [12] M. S. Pino and D. C. Chung, "The chromosomal instability pathway in colon cancer," *Gastroenterology*, vol. 138, pp. 2059-2072, 2010.

- 
- [13] C. R. Boland and A. Goel, "Microsatellite instability in colorectal cancer," *Gastroenterology*, vol. 138, pp. 2073-2087.e3, 2010.
- [14] F. A. Sinicrope and D. J. Sargent, "Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications," *Clin Cancer Res*, vol. 18, pp. 1506-1512, 2012.
- [15] T. Armaghany, J. D. Wilson, Q. Chu, and G. Mills, "Genetic alterations in colorectal cancer," *Gastrointest Cancer Res*, vol. 5, pp. 19-27, 2012.
- [16] P. P. Panigrahi and T. R. Singh, "Computational studies on Alzheimer's disease associated pathways and regulatory patterns using microarray gene expression and network data: revealed association with aging and other diseases," *J Theor Biol*, vol. 334, pp. 109-121, 2013.
- [17] J. M. Wheeler, W. F. Bodmer, and N. J. Mortensen, "DNA mismatch repair genes and colorectal cancer," *Gut*, vol. 47, pp. 148-153, 2000.
- [18] M. Sehgal and T. R. Singh, "DR-GAS: a database of functional genetic variants and their phosphorylation states in human DNA repair systems," *DNA Repair (Amst)*, vol. 16, pp. 97-9103, 2014.
- [19] B. Levin, D. A. Lieberman, B. McFarland, K. S. Andrews, D. Brooks, J. Bond, C. Dash, F. M. Giardiello, S. Glick, D. Johnson, et al., "Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps, 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology," *Gastroenterology*, vol. 134, pp. 1570-1595, 2008.
- [20] R. W. Burt, J. S. Barthel, K. B. Dunn, D. S. David, E. Drelichman, J. M. Ford, F. M. Giardiello, S. B. Gruber, A. L. Halverson, S. R. Hamilton, et al., "NCCN clinical practice guidelines in oncology. Colorectal cancer screening," *J Natl Compr Canc Netw*, vol. 8, pp. 8-61, 2010.
- [21] E. P. Whitlock, J. S. Lin, E. Liles, T. L. Beil, and R. Fu, "Screening for colorectal cancer: a targeted, updated systematic review for the U.S. Preventive Services Task Force," *Annals of Internal Medicine*, vol. 149, pp. 638-658, 2008.
- [22] E. G. Pulido, A. R. Oliveira, J. B. BARGUES, C. G. Ponce, and A. Carrato, "Molecular biology of colorectal cancer," in *The Challenge of Colorectal Cancer: A Review Book*, E. U. Cidon, Ed. India: Research Signpost, 2011, pp. 35-51.

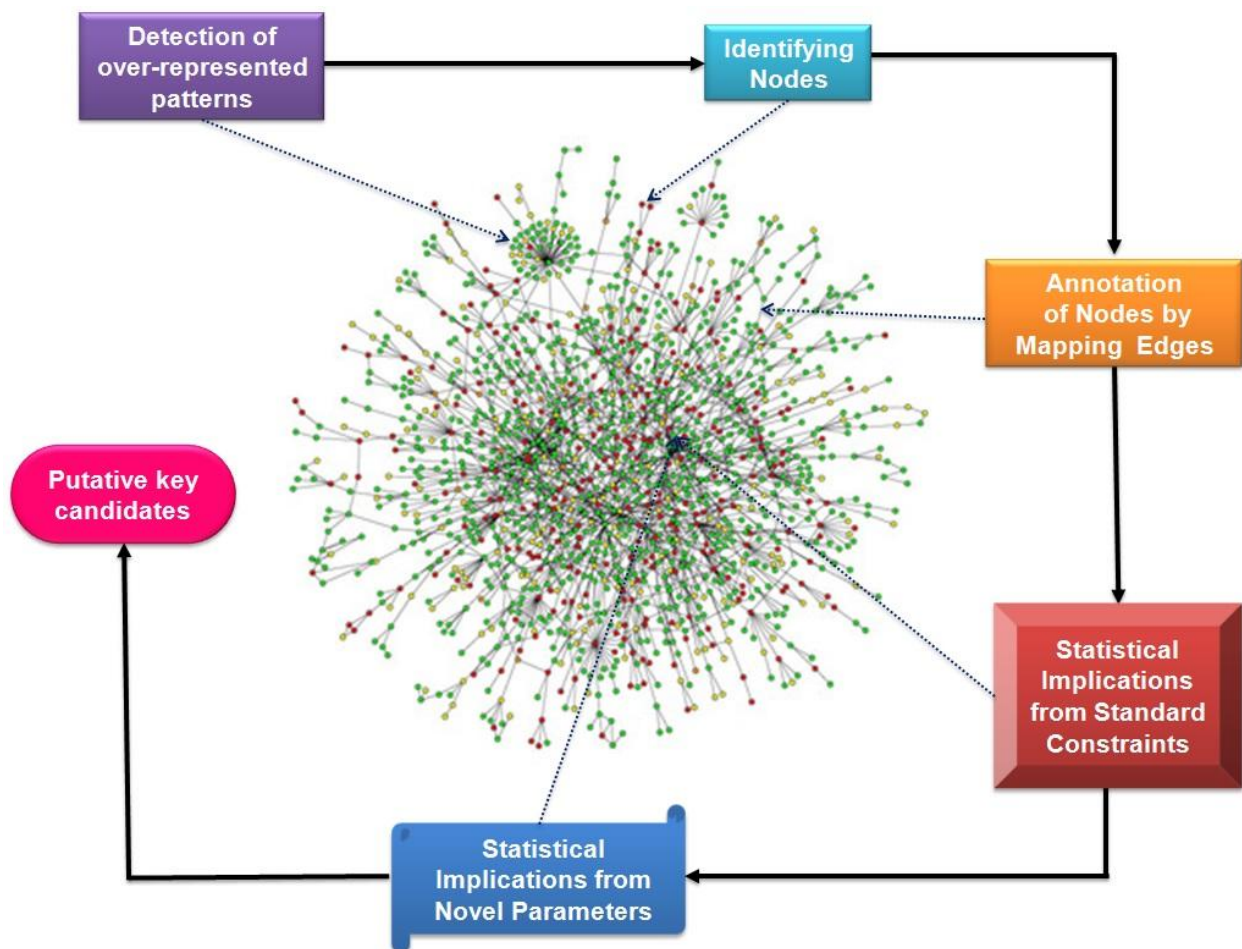
- 
- [23] W. Y. Tan and X. W. Yan, "A new stochastic and state space model of human colon cancer incorporating multiple pathways," *Biol Direct*, vol. 5, p. 26, 2010.
- [24] P. Hegde, R. Qi, R. Gaspard, K. Abernathy, S. Dharap, J. Earle-Hughes, C. Gay, N. U. Nwokekeh, T. Chen, A. I. Saeed, et al., "Identification of tumor markers in models of human colorectal cancer using a 19,200-element complementary DNA microarray," *Cancer Research* vol. 61, pp. 7792-7797, 2001.
- [25] T. T. Zou, F. M. Selaru, Y. Xu, V. Shustova, J. Yin, Y. Mori, D. Shibata, F. Sato, S. Wang, A. Olaru, E. Deacu, T. C. Liu, J. M. Abraham, and S. J. Meltzer, "Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon," *Oncogene*, vol. 21, pp. 4855-4862, 2002.
- [26] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nature Genetics*, vol. 25 pp. 25-29, 2000.
- [27] A. Pratap, S. Taliyan, and T. R. Singh, "NMDB: Network Motif Database envisaged and explicated from human disease specific pathways.," *Journal of biological systems*, vol. 22, pp. 89-100, 2014.
- [28] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, pp. D991–D995, 2013.
- [29] Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah, "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis," *Clin Cancer Res*, vol. 13, pp. 1107-1114, 2007.
- [30] S. Baglioni and M. Genuardi, "Simple and complex genetics of colorectal cancer susceptibility," *American journal of medical genetics Part C, Seminars in medical genetics*, vol. 129C, pp. 35-43, 2004.
- [31] H. T. Lynch and A. de la Chapelle, "Hereditary colorectal cancer," *N Engl J Med*, vol. 348, pp. 919-932, 2003.
- [32] M. Sehgal and T. R. Singh, "Identification and analysis of biomarkers for mismatch repair proteins: A bioinformatic approach," *J Nat Sci Biol Med*, vol. 3, pp. 139-146, 2012.

- 
- [33] M Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research* vol. 28, pp. 27-30, 2000.
- [34] G. Joshi-Tope, I. Vastrik, G. R. Gopinath, L. Matthews, E. Schmidt, M. Gillespie, P. D'Eustachio, B. Jassal, S. Lewis, G. Wu, E. Birney, and L. Stein, "The Genome Knowledgebase: a resource for biologists and bioinformaticists," *Cold Spring Harb Symp Quant Biol*, vol. 68, pp. 237-243, 2003.
- [35] C Stark, B J Breitkreutz, T Reguly, L Boucher, A Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets.," *Nucleic Acids Research* vol. 34 pp. D535-D539, 2006.
- [36] K. Kandasamy, S. S. Mohan, R. Raju, S. Keerthikumar, G. S. S. Kumar, A. K. Venugopal, D. Telikicherla, J. D. Navarro, S. Mathivanan, C. Pecquet, et al., "NetPath: a public resource of curated signal transduction pathways," *Genome Biol*, vol. 11, 2010.
- [37] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249-264, 2003.
- [38] R. Bender and S. Lange, "Adjusting for multiple testing--when and how?," *J Clin Epidemiol*, vol. 54, pp. 343-349, 2001.
- [39] S. Zang, R. Guo, L. Zhang, and Y. Lu, "Integration of statistical inference methods and a novel control measure to improve sensitivity and specificity of data analysis in expression profiling studies," *J Biomed Inform*, vol. 40, pp. 552-560, 2007.
- [40] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci U S A*, vol. 98, pp. 5116-5121, 2001.
- [41] D. G. Covell, A. Wallqvist, A. A. Rabow, and N. Thanki, "Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data," *Mol Cancer Ther*, vol. 2, pp. 317-332, 2003.
- [42] H. Hotelling, "Analysis of a complex of statistical variables into principle components," *Journal of Educational Psychology*, vol. 24, pp. 417-441, 1933.
- [43] V Gotea and I. Ovcharenko, "DiRE: identifying distant regulatory elements of co-expressed genes," *Nucleic Acids Research*, vol. 36, pp. W133-139, 2008.

- 
- [44] S. J. Ho Sui, D. L. Fulton, D. J. Arenillas, A. T. Kwon, and W. W. Wasserman, "oPOSSUM: integrated tools for analysis of regulatory motif over-representation," *Nucleic Acids Res*, vol. 35, pp. 245-252, 2007.
- [45] B Zhang, S Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," *Nucleic Acids Research*, vol. 33, pp. W741-748, 2005.
- [46] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, pp. 48-48, 2009.
- [47] E. Eden, D. Lipson, S. Yogev, and Z. Yakhini, "Discovering motifs in ranked lists of DNA sequences," *PLoS Comput Biol*, vol. 3, 2007.
- [48] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, pp. 1746-1758, 2004.
- [49] F Schreiber and H. Schwobbermeyer, "MAVisto: a tool for the exploration of network motifs," *Bioinformatics* vol. 21, pp. 3572-3574, 2005.
- [50] S Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, pp. 1152-1153, 2006.
- [51] T M J Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: practice and experience*, vol. 21, pp. 1129-1164, 1991.
- [52] S. Wernicke, "Efficient detection of network motifs," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 3 pp. 347-359, 2006.
- [53] K Fransén, M Klintenäs, A Osterström, J Dimberg, H J Monstein, and P. Söderkvist, "Mutation analysis of the BRAF, ARAF and RAF-1 genes in human colorectal adenocarcinomas," *Carcinogenesis*, vol. 25, pp. 527-533, 2004.
- [54] D. Matallanas, M. Birtwistle, D. Romano, A. Zebisch, J. Rauch, A. von Kriegsheim, and W. Kolch, "Raf Family Kinases: Old Dogs Have Learned New Tricks," *Genes & Cancer*, vol. 2, pp. 232-260, 2011.
- [55] J. W. Lee, Y. H. Soung, S. Y. Kim, W. S. Park, S. W. Nam, W. S. Min, S. H. Kim, J. Y. Lee, N. J. Yoo, and S. H. Lee, "Mutational analysis of the ARAF gene in human cancers," *APMIS*, vol. 113, pp. 54-7, 2005.

# CHAPTER-5

Decoding the intricate biological pathways in quest of candidate markers implicated in human DNA repair system



*“The roots of true achievement lie in the will to become the best that you can be.”  
-Harold Taylor*

## **ABSTRACT**

The biological network complexity is growing enormously and in order to reveal confined properties of these intricate networks, detection of crucial network components i.e. network motifs (over-represented sub-graphs) may assist in gaining effortless perceptiveness on the underlying biological processes. Analyzing complex human DNA repair pathways for their disease association is still a drawn-out process and requires an integrative approach for comprehensive examination of proteins and interactions to identify candidate markers underlying major malignancies and genetic disorders. Besides, the available standard statistical parameters such as  $z$ -score,  $p$ -value and significance profile, a few novel parameters representing over-represented structures were designed for the identification of key signatures such as MIM, BIFAN, SIM, FFL, 3-chain and 4-chain motifs from DNA repair pathways. Thus, the novel pipelined approach for the designed parameters will help comprehend the dependence of an entire system on a few key genes, proteins and metabolites for examining their statistical significance. For investigating DNA repair proteins and their interacting partners, the entire DNA repair pathway has been subjected to a pathway level analysis with diverse parameters under consideration like closeness centrality, shortest path length and one to one interacting entities. Further, on applying clustering technique, smaller sub-graphs from the pathway were generated to represent the crucial nodes and their interactions. In the study, we switched from the component level to systems level by annotating all the components and their associated interactions in a biological system and finally proposed and confined the search to only a few key regulatory players by uncovering the intricacy implicated in DNA repair associated diseases. It is anticipated that the proposed methodology would serve as a valuable complement for analyzing biomarkers in human repair specific disease pathways and will also contribute scientific knowledge towards their better understanding. These proposed biomolecules could be targeted for designing appropriate experiments and hence reducing the time and resources.

## 5.1 INTRODUCTION

The damage to human genome due to endogenous aberrant processes, environmental agents and genetic defects is purged from the human system via a process that recognizes and removes damages/lesions from DNA known as DNA repair [1]. There is a high rate of recurrence for endogenous DNA damage as compared to exogenous damage and the type of damages produced due to both factors is roughly indistinguishable [2]. The damage to the DNA is caused by multiple factors such as oxidation of bases, generation of DNA strand interruptions, alkylation of bases [3], bulky adduct formation, mismatches and pyrimidine dimers that often trigger viral interactions [4]. Automatically the cell tries to repair damages such as replication errors, ROS, harmful radiations and thermal disruption to maintain the integrity of genome using diverse repair mechanisms [1]. Eliminating and fixing the damaged DNA from the genome is a complicated process involving myriad of repair proteins like DDB2, MLH1, XPA and different associated mechanisms for diverse type of lesions. As already mentioned, the numerous known mechanisms by which the damaged DNA is repaired includes BER, NER, MMR, HRR, NHEJ, DDS and TLS that incorporates different set of genes, enzymes and pathways [5, 6]. These mechanisms not only maintain the genetic stability but also prevent the genome from carcinogenesis, pre-mature aging and other genetic disorders like CS, XP and progeria [1, 5, 7]. To broadly comprehend the role of DNA repair in associated diseases and perceiving crucial markers implicated in diseases, the final objective was designed to focus on the better understanding of entire process.

Since, variety of lesions triggers diverse repair pathways and involve complex factors for damage removal therefore the biological complexity enhances multi-folds. In order to deduce crucial aberrations in DNA repair associated diseases, dividing the entire network into number of distinct clusters would help in revealing some specific confined properties of these intricate networks. A variety of networks are composed of different sets of local structures as patterns which are referred to as network motifs [8]. Network motifs are over-represented sub-structures in a network that are assumed as recurring pattern of interaction from where the networks are believed to be built. The network motifs were first described in *Escherichia coli* [9] where they stated them as patterns occurring in the biological network more frequently than would have been expected in arbitrary random versions of the same networks. These motifs are an important



parameter to analyze since it not only describes the local properties of a network but also are small linked sub-graphs that appear most frequently and distinctively in a biological network. These network motifs are supposed to have functional significance related to biological pathways. As per the existing knowledge regarding network motifs, there are different patterns classified as network motifs including auto-regulation (positive or negative), feedforward loops (FFL), single-input modules (SIM), multiple input modules (MIM), feedback loops, dense overlapping regulons (DOR) and other putative regulatory motifs [10].

All kind of network motifs have their own importance in biological networks. For instance, FFL and bifan network motifs are recognized as distinctive patterns in different types of biological networks. The FFL is a model consisting of 3 genes, where two input transcription factors exists, each of which regulating one another, and then jointly regulates a target gene. The FFL has eight promising structural types where the four FFLs correspond to incoherent FFLs, acting as sign-sensitive accelerators i.e. off to on but not vice-versa. The other four types, coherent FFLs, act as sign-sensitive delays. This kind of biologically meaningful information should be associated with all network motif types and there is a need to correlate this information with their functional significance. Keeping in view the requisite for a systems biology approach for DNA repair related disorders and pathways; we proposed our final objective to fill the research gap. The biological network complexity is growing enormously and in order to reveal specific confined properties of these intricate networks, the breakdown of network into distinct clusters and components would be extremely valuable [10]. The diverse networks comprise of different set of local structures as patterns which are referred to as network motifs.

Identification and annotation of these biological entities in DNA repair pathways may help in analyzing vital interactions involved in diseases associated to DNA repair and may further aid in resolving critical functions performed by these networks. In this study, we have switched from the systems level to the component level by annotating all components and their associated interactions in a biological system and finally proposed a few major players implicated in DNA repair associated diseases. There is a need to utilize computational resources and implementation of appropriate statistical criteria for the efficient detection of biological network motifs with several estimation measures. The quantitative and qualitative studies at

systems level could provide insights to various biological systems ranging from organism specific, disease specific and to mechanism specific. We will apply a similar approach to deal with biological networks and their components as network motifs to analyze them for their functional consequences in repair specific disease pathways.

Till date, there is no such study based on DNA repair pathways that connects these vital biological entities to the complicated interacting pathways. In addition, there are a few in silico studies related to diseases caused due to DNA repair system. As per our knowledge, there has been no such study performed for examining the enrichment of network motifs in different biological networks for human disease specific pathways related to repair. It is believed that the proposed work will assist molecular and system biologists, biotechnologists and other scientific community to encounter biologically meaningful information. In the present study, 13 DNA repair related diseases i.e. Fanconi anemia, Glioma, Non-small cell lung cancer, Colorectal cancer, Acute myeloid leukemia, Brain tumor, Chronic myeloid leukemia, Endometrial cancer, Necrosis, Pancreatic cancer, Prostate cancer, Renal cell carcinoma and Small-cell lung cancer have been meticulously scrutinized. The biological pathways concerning these diseases were analyzed and finally key players are proposed that may serve as potential biomarkers for these diseases on further experimentation and validations.

## **5.2 MATERIALS AND METHODS**

### **5.2.1 Reconstruction and statistical estimation of DNA repair pathway**

A variety of computational resources and statistical criteria were implemented for the efficient detection of biological network motifs and highly interconnected patterns with several estimation measures. The DNA repair pathway was elucidated by mapping the human interactome with the available information on DNA repair proteins. The intricate network thus generated revealed specifically the DNA repair proteins and its associated protein-protein interactions (PPIs). This network comprised of numerous one to one, one to many and many to many interactions. Such, an analysis is exceptionally valuable in identifying interacting partners and other influencing factors. Thus, the detection of neighborhood proteins were performed using Cytoscape platform [11]. In addition, the complexity of generated pathway was simplified by applying clustering technique to identify highly inter-connected networks using the MCODE plug-in [12] of the

software. Various statistical inferences such as short path length, neighborhood connectivity and average clustering coefficient were generated for the DNA repair pathway using network analyzer [13].

The shortest path length  $L(n,m)$  is the length of shortest path between two nodes  $n$  and  $m$ . The shortest path length distribution gives the number of node pairs  $(n,m)$  with  $L(n,m) = k$  for  $k = 1, 2, \dots, z$ . Whereas, the neighborhood connectivity of a node  $n$  is defined as the average connectivity of all its neighbours [14]. The neighborhood connectivity distribution gives the average of the neighborhood connectivity's of all nodes  $n$  with  $k$  neighbors for  $k = 0, 1, \dots, z$ . The Indegree of a node  $n$  is the number of incoming edges and Outdegree of a node  $n$  is the number of outgoing edges. The closeness centrality  $C_c(n)$  of a node  $n$  is defined as the reciprocal of the average shortest path length [15] and is computed as:

$$C_c(n) = 1 / \text{avg}(L(n,m))$$

The closeness centrality of each node is a number between 0 and 1 where 0 corresponds to an isolate node and 1 represent nodes directly connected to other nodes. The average clustering coefficient distribution gives the average of clustering coefficients for all nodes  $n$  with  $k$  neighbors for  $k = 2, \dots, z$ . In directed network, the clustering coefficient  $C_n$  of a node  $n$  is defined as:

$$C_n = e_n / (k_n(k_n - 1))$$

where  $k_n$  is the number of neighbors of  $n$  and  $e_n$  is the number of connected pairs between all neighbors of  $n$  [16, 17]. The clustering coefficient is a ratio  $N / M$ , where  $N$  is the number of edges between the neighbors of  $n$ , and  $M$  is the maximum number of edges that could possibly exist between the neighbors of  $n$ . The clustering coefficient of a node always ranges from 0-1. The average clustering coefficient has already presented its utilization in metabolic networks where the modular organization has been studied [18].

The betweenness centrality of a node reflects the amount of control exerted by this node over interactions of other nodes in the network [19]. The betweenness centrality  $C_b(n)$  of a node  $n$  is computed as follows:

$$C_b(n) = \sum_{s \neq n \neq t} (\sigma_{st}(n) / \sigma_{st})$$

where  $s$  and  $t$  are nodes in the network different from  $n$ ,  $\sigma_{st}$  denotes the number of shortest paths from  $s$  to  $t$ , and  $\sigma_{st}(n)$  is the number of shortest paths from  $s$  to  $t$  that  $n$  lies on [20]. The other analyzed crucial feature of the pathway was stress centrality which is the number of shortest paths passing through a node  $n$ . This is an important feature to analyze since a node has high stress if it is traversed by a high number of shortest paths.

### 5.2.2 Biological pathways and detection of network motifs

The biological pathways for 13 major DNA repair associated diseases were retrieved from online resources and databases like KEGG [21], Reactome [22] and BioGrid [23]. From these intricate pathways, identification of vital network patterns/components was achieved via diverse tools such as FANMOD [24], MAVisto [25] and mfinder [26]. The principle of all these tools relies on different complex algorithms each representing some common as well as unique features. The MAVisto tool employs a flexible algorithm for identification of network motifs and also includes an advanced force-directed layout algorithm [27] for its analyses whereas mFinder uses a semi-dynamic programming algorithm in order to reduce the run time in detecting network motifs and performs full enumeration of the sub-graphs [26]. Moreover, FANMOD tool runs a much sophisticated algorithm named randomized enumeration of sub-graphs (RAND-ESU) algorithm [28] that works on both directed as well as undirected networks for specification and sampling of sub-graphs. This algorithm performs better than its counter algorithms [26] for the identification of network motifs from complex biological networks. FANMOD is based on node-sampling strategy and a pattern growth tree using node-extension and is fast compared to other tools in detecting up to size-8 motifs in both directed/undirected networks [24].

### 5.2.3 Annotation of vital network components

The information on network motifs from all these diverse tools was compiled for identification of promising patterns with high assurance. The retrieved motif sub-graphs ranged from 3-8 nodes that were further subjected to annotation and disease-specific analyses. After the elucidation of important patterns (motifs), manually curation and annotation of above interactions along with their molecular functions was performed via literature survey and available online resources. For functional enrichment of all the pathway entities, in-house perl scripts and diverse resources

were utilized for identifying genes and implicated interactions among nodes such as activation, inhibition, phosphorylation and gene expression. The network motifs thus obtained from the disease associated pathways contained SIM, MIM, Bifan and 4-chain motifs and other important biological signatures which were also supported by  $z$ -scores for their statistical significance. These network motifs and diverse signatures have important functions to perform as in SIM motif, several genes are controlled by a single master gene and the master gene is known to be auto-regulatory whereas MIM is a generalization of SIM. Other regular 4-node motifs that were detected confirmed the presence of diamond, biparallel and bifan motifs which are usually built by two regulators and two regulated genes. These patterns were generated and annotated for all the 13 DNA repair associated diseases.

#### 5.2.4 Biological inference form available statistical constraints

The statistical implication of these generated motifs was evaluated using diverse available constraints such as  $z$ -scores,  $p$ -values and significance profile (SP). The  $p$ -value and  $z$ -score for each motif was calculated and those having  $z$ -score  $> 2$  and  $p$ -value  $< 0.05$ ; i.e. significant motifs were selected. The network motif ID corresponds to the adjacency matrix created for each motif whereas  $z$ -score and  $p$ -values represent the significance of these motifs. Further, the SP furnishes normalized  $z$ -score values for a particular network motif ( $m_i$ ) which is given by:

$$SP(m_i) = \frac{Z(m_i)}{\sqrt{\sum_{i=1}^n Z(m_i)^2}}$$

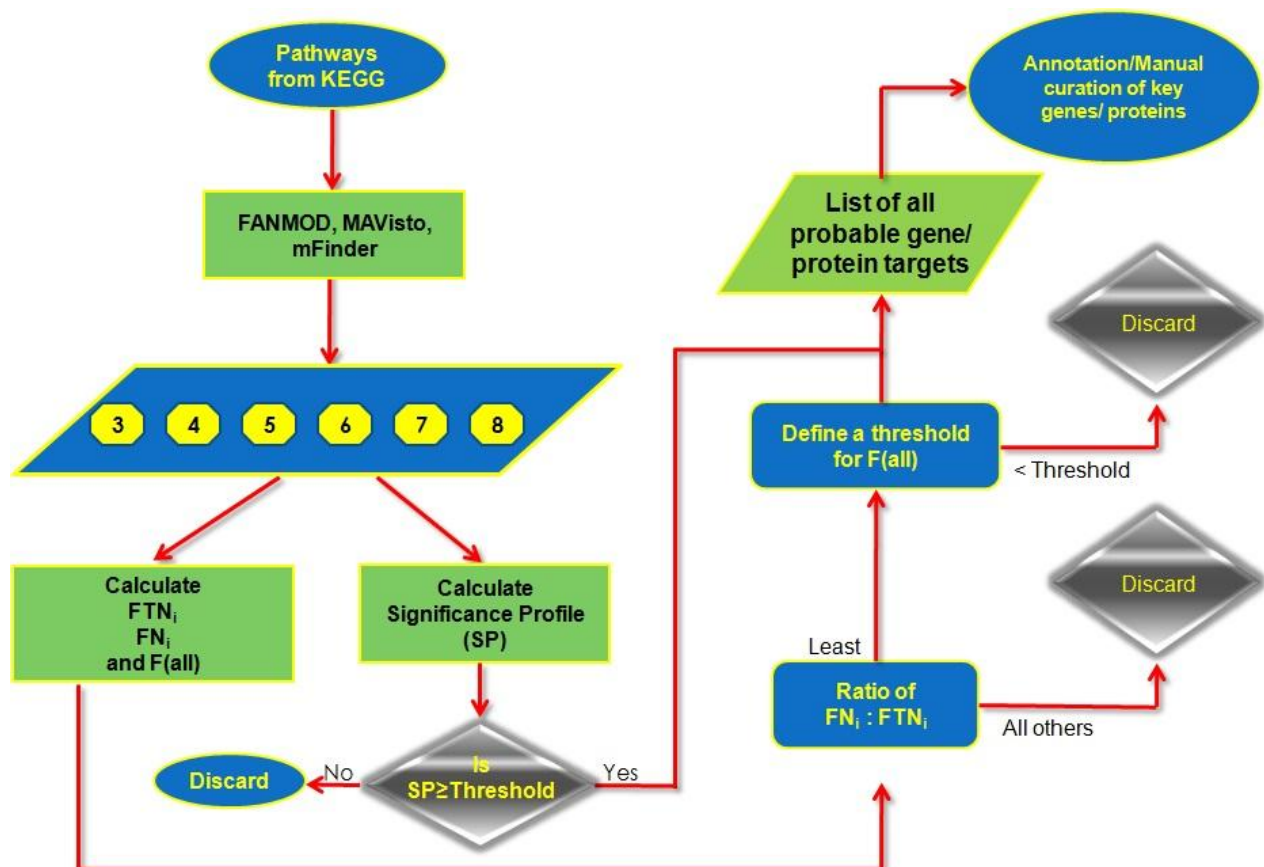
Where  $Z(m_i)$  corresponds to the  $z$ -score value for each network motif.

#### 5.2.5 Novel designed parameters for deducing crucial patterns

All the generated 3-8 node sub-graphs with unique network motif IDs were then extensively analysed for examining proteins and their complex interactions in disease associated pathways using our novel designed parameters such as ' $FN_i$ ', ' $FTN_i$ ' and ' $FT_i$ '. Here, ' $FN_i$ ' corresponds to the number of genes present in a given network motif ID; ' $FTN_i$ ' is the sum of frequencies for all the genes occurring in a given network motif ID and ' $FT_i$ ' is defined as the ratio of number of genes for a particular network motif ID and the sum of frequencies for all genes in a given network motif. For a given network motif ID say ' $n_i$ ', where  $i=1,2,3,\dots,n$ ; ' $FT_i$ ' is given by:

$$FT_i = \frac{FN_i}{FTN_i}$$

Each ' $FT_i$ ' value for a particular network motif ID provides the magnitude of all genes involved in a particular network motif. From these novel deliberated parameters, it was observed; the lower ' $FT_i$ ' value proves to be more statistically significant as it signified greater involvement of a few genes that explains complex interactions among different nodes in a given motif. Thus, the applied methodology comprises of both top-down and bottom-up approaches for detecting the key players in complicated pathways. Using the top-down approach, first the entire disease pathways were partitioned into smaller sub-graphs with small functional modules and then the involved nodes were identified and annotated. The current study lay upon the intended pipeline which has been described in the form of a flowchart in **Figure 5.1**. These network motifs are already proved to be an important aspect to recognize the modularity and to solve large-scale structure of complicated biological networks.

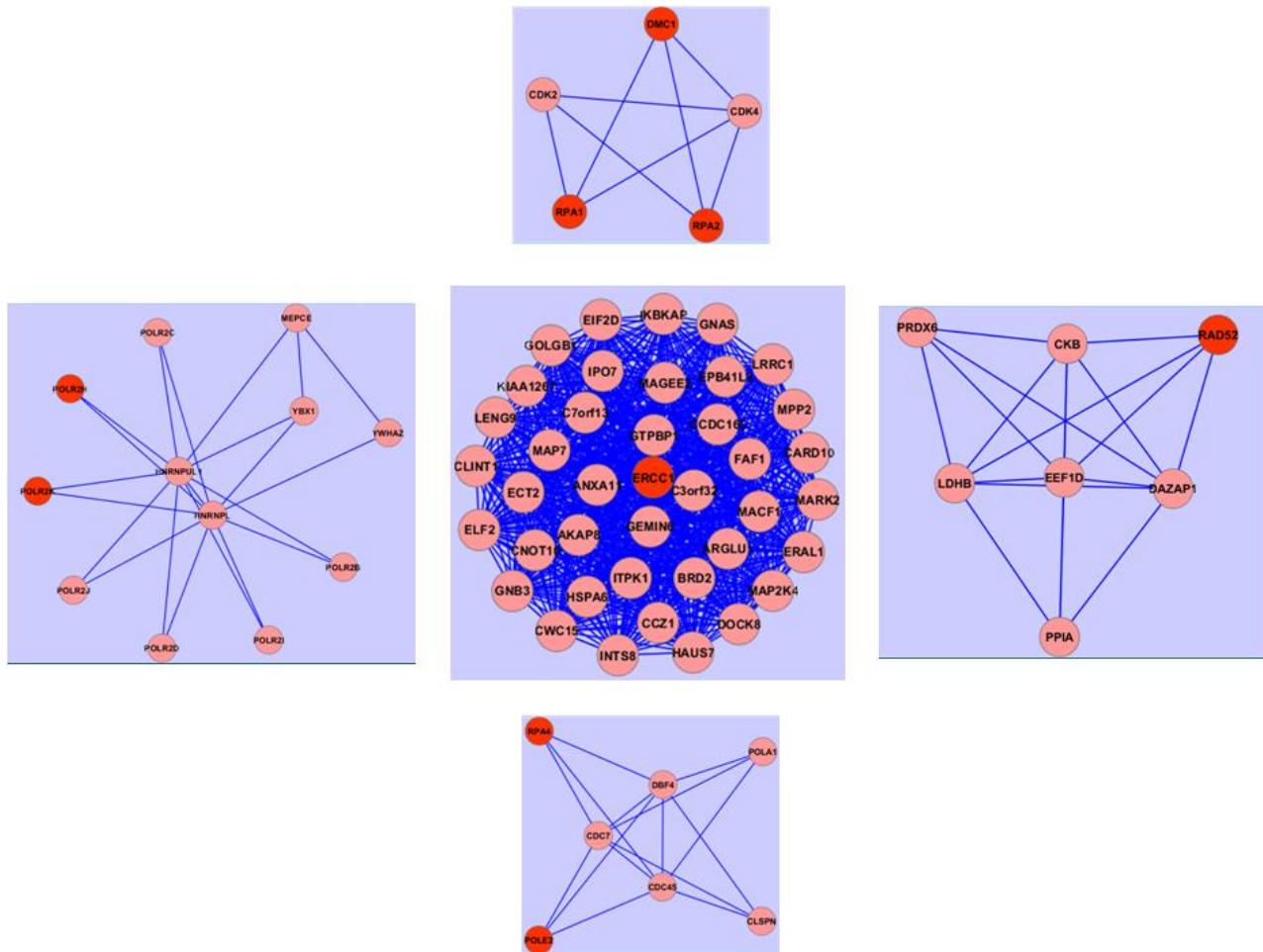


**Figure 5.1** A flow diagram for the applied methodology to detect key players from diseases involved in DNA repair.

### 5.3 RESULTS AND DISCUSSION

For examining the complex biological processes concerned with diseases related to DNA repair, a systems biology approach integrating several biological components and other influencing factors is essential to understand. A systems level understanding of the disease helps unravel several crucial network components and many regulatory elements in a coordinated manner. Using the integrative approach, the perceptive of complexity hidden in biological phenomenon is extensively simplified. Identification of candidate markers from DNA repair pathways associated to numerous types of cancers and other diseases is still a convoluted process and requires comprehensive examination of proteins and involved interactions. Since, analyzing the intact complex DNA repair protein interaction network is tedious task therefore we performed analysis to detect crucial network components i.e. network motifs (over-represented sub-graphs) that may assist in effortless understanding of the underlying biological processes.

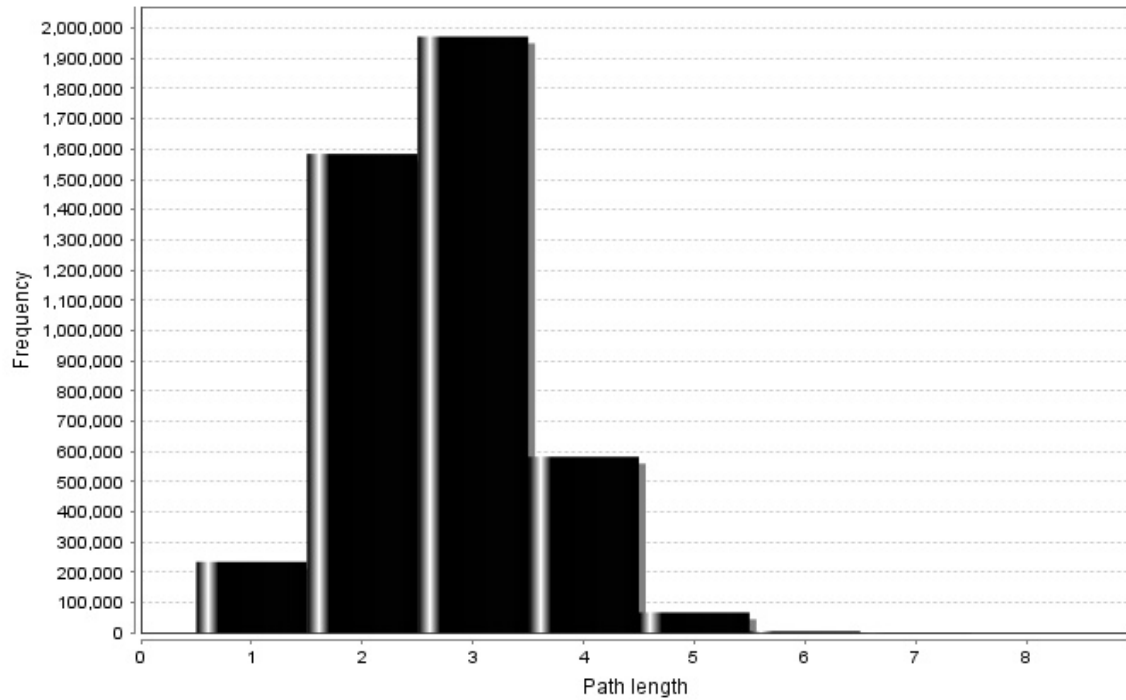
To meticulously comprehend the DNA repair process and implicated pathways, the DNA repair proteins along with their interacting partners were subjected to clustering in order to deduce crucial patterns as demonstrated in **Figure 5.2**. Various other proteins encompassing key interactions were also elucidated from the analysis where red coloured nodes represent the input DNA repair proteins and the pink coloured nodes describe the interacting partners. The intricate DNA repair pathway was then subjected to network analysis from the Cytoscape tool and vital statistical inferences were generated. **Figures 5.3-5.5** present few graphs obtained via scrupulous pathway level analysis covering important aspects such as shortest path length, neighborhood connectivity, indegree distribution, closeness centrality, average clustering coefficient, betweenness centrality and stress centrality. The shortest path length graph clearly depicts the small-world properties of the analyzed network [17]. The highest peak was obtained for the path length 3, indicating the effectiveness of dividing the entire pathway into small functional modules with probably 4 nodes. The neighborhood connectivity provides the average connectivity of all neighbors of  $n$  where the direction of edges is ignored. The graph illustrating both in and out neighbors portray that as the number of neighbors increase, the connectivity among them also becomes complex. The In degree graph shows that when there are huge number of nodes, the incoming interactions are comparatively low as compared to fewer number of nodes where the incoming nodes are observed to some extent.



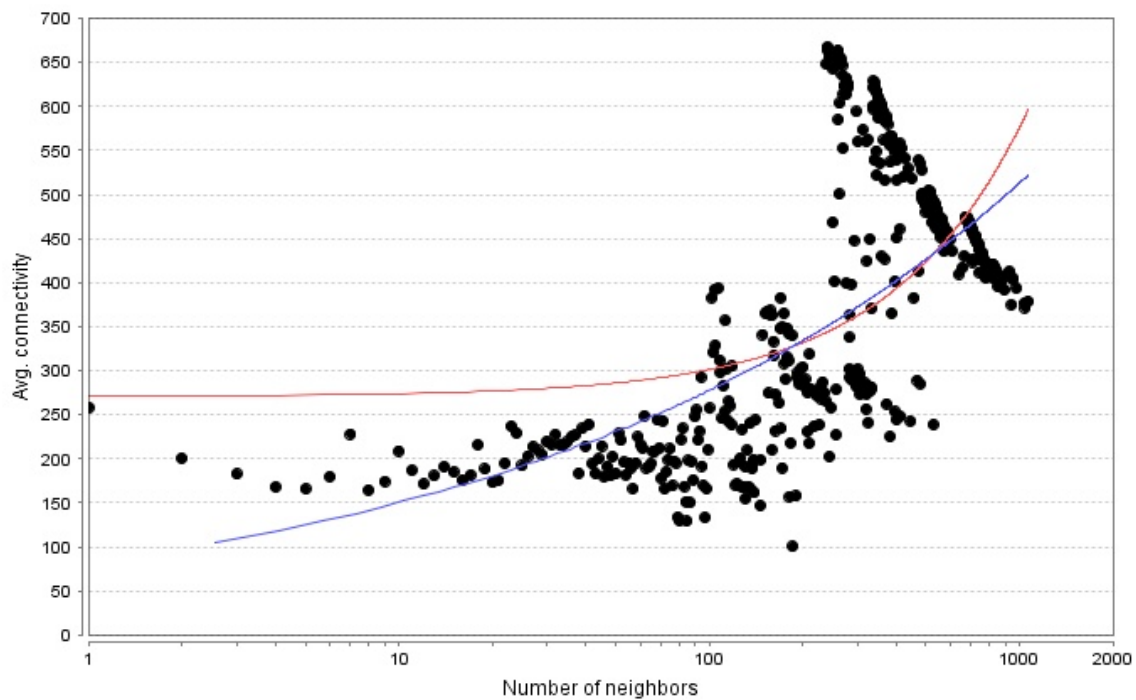
**Figure 5.2** The one to one, one to many and many to many interacting partners attained from clustering; revealed hub nodes as well as important connections.

The closeness centrality shows the rate of information spread from one node to other accessible nodes in the network. The average clustering coefficient is basically a measure of the degree of nodes clustering in a graph. The clusters are observed more closely once the node size increases where the clustering is performed on the basis of physical and biochemical similarities. The betweenness centrality parameter of a node not only reflects the management of interactions of other nodes in the network with this node rather is highly valuable when dealing with dense biological networks. The graph evidently depicts the betweenness centrality measure for all the nodes implicated in DNA repair pathway. The attained stress centrality graph shows an exponential decline when the number of nodes increases illustrating less availability for connecting nodes.



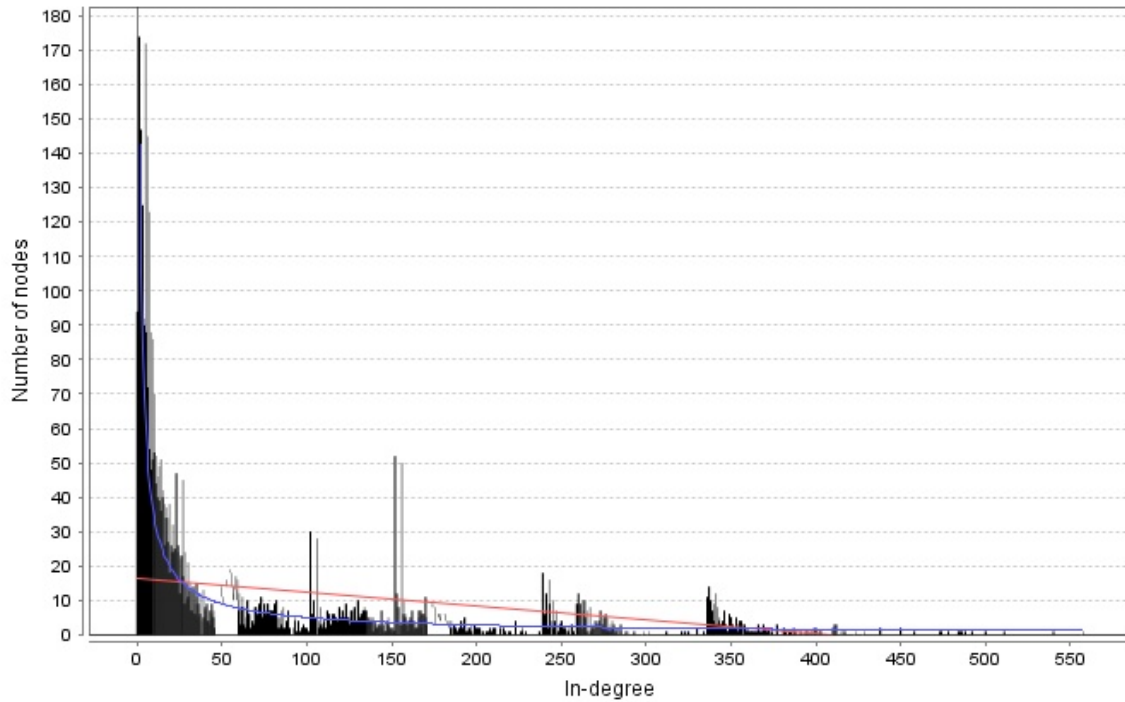


### a. Shortest path length

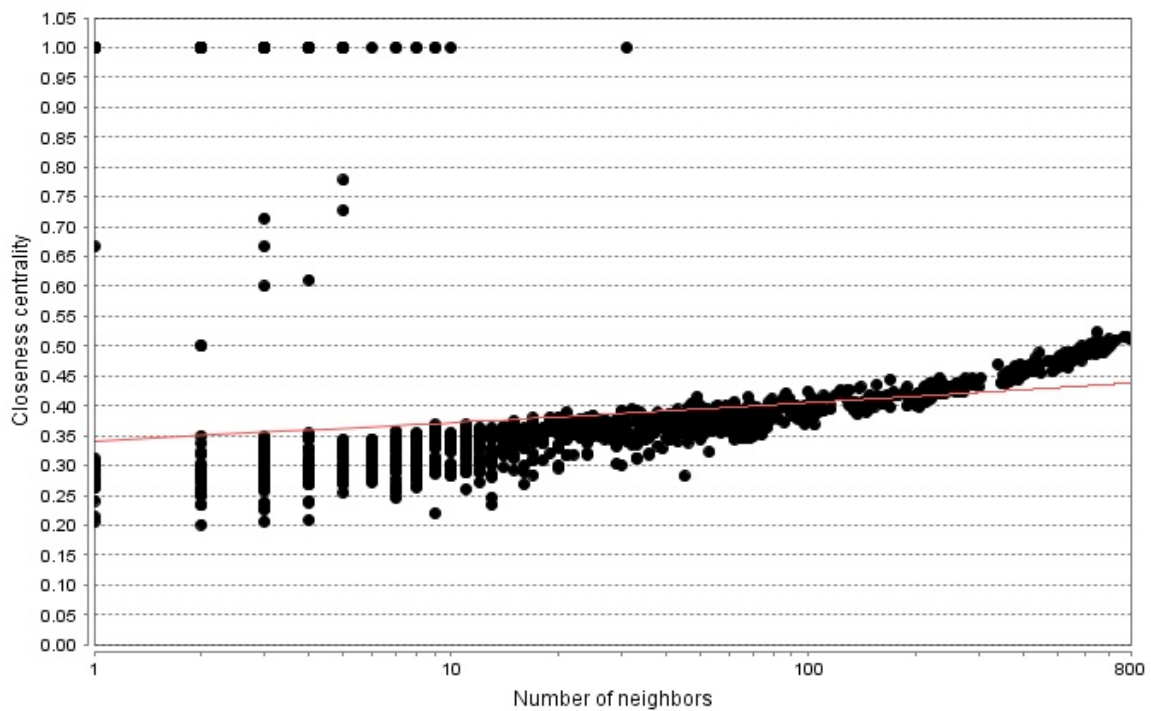


### b. Neighborhood connectivity (In and Out)

**Figure 5.3** The statistical inferences from DNA repair pathway comprising of shortest path length and neighborhood connectivity.

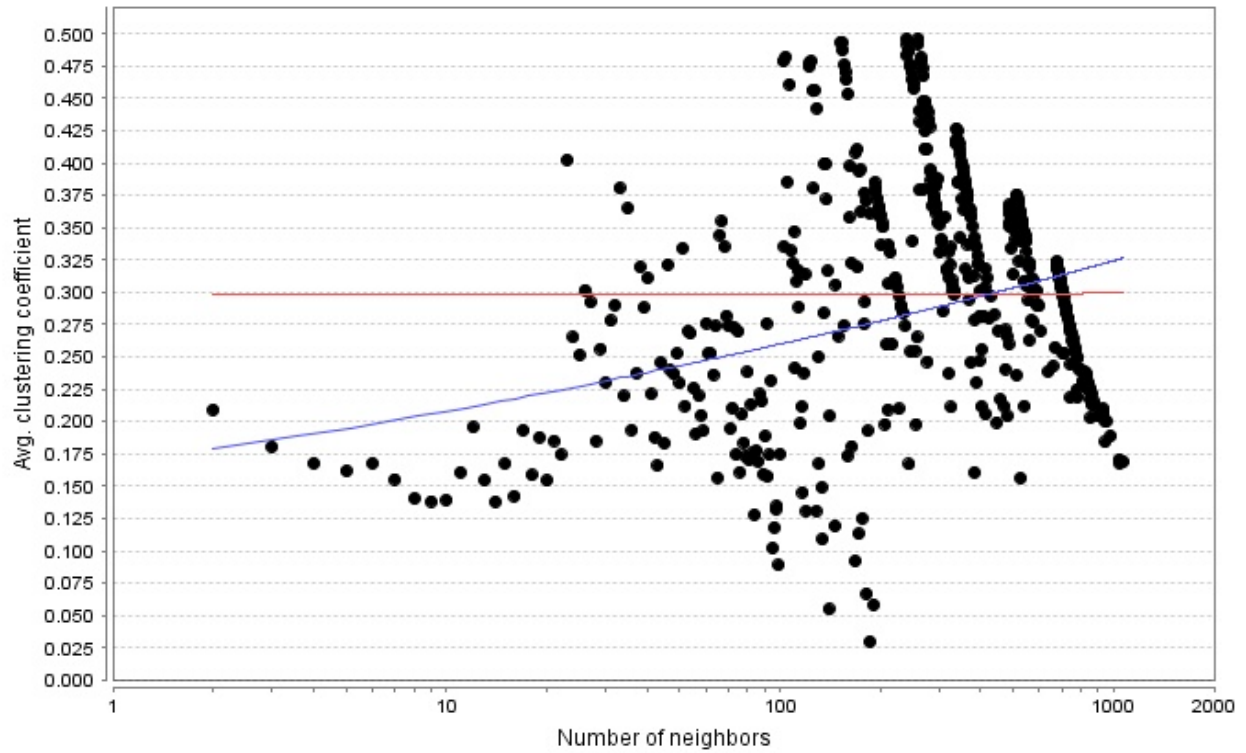


### a. Indegree distribution

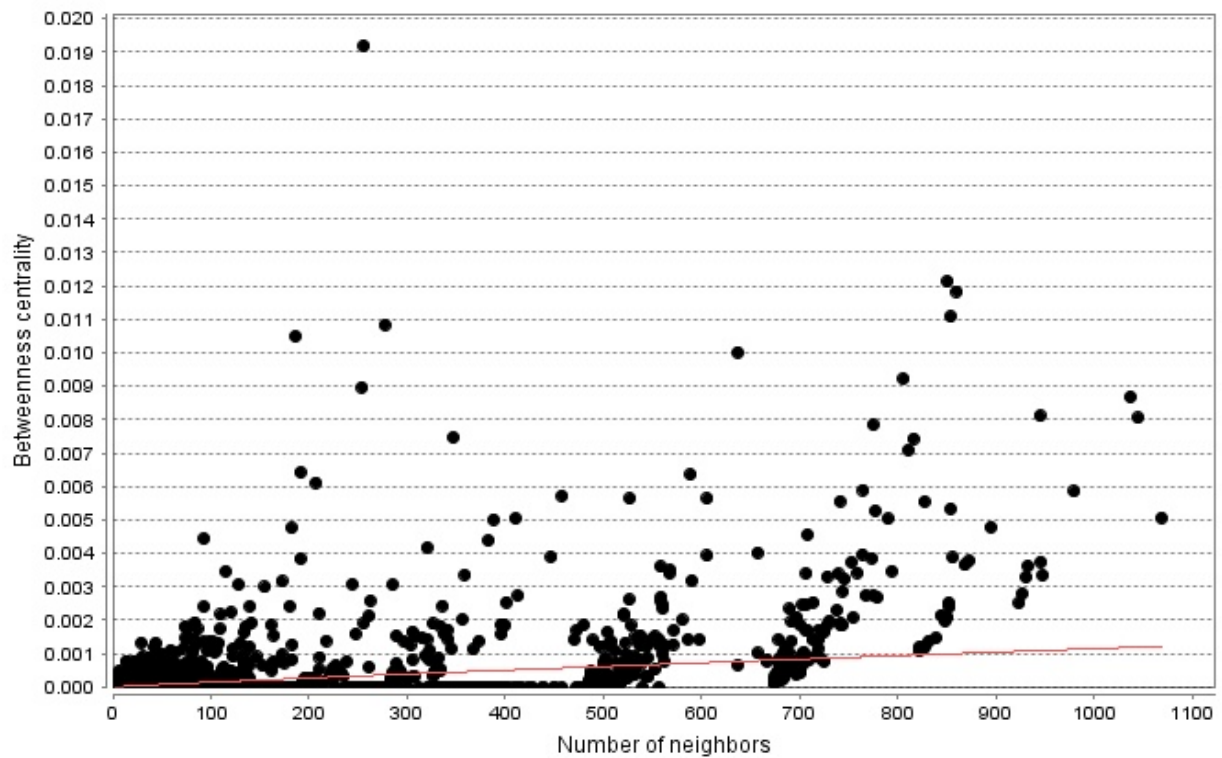


### b. Closeness centrality

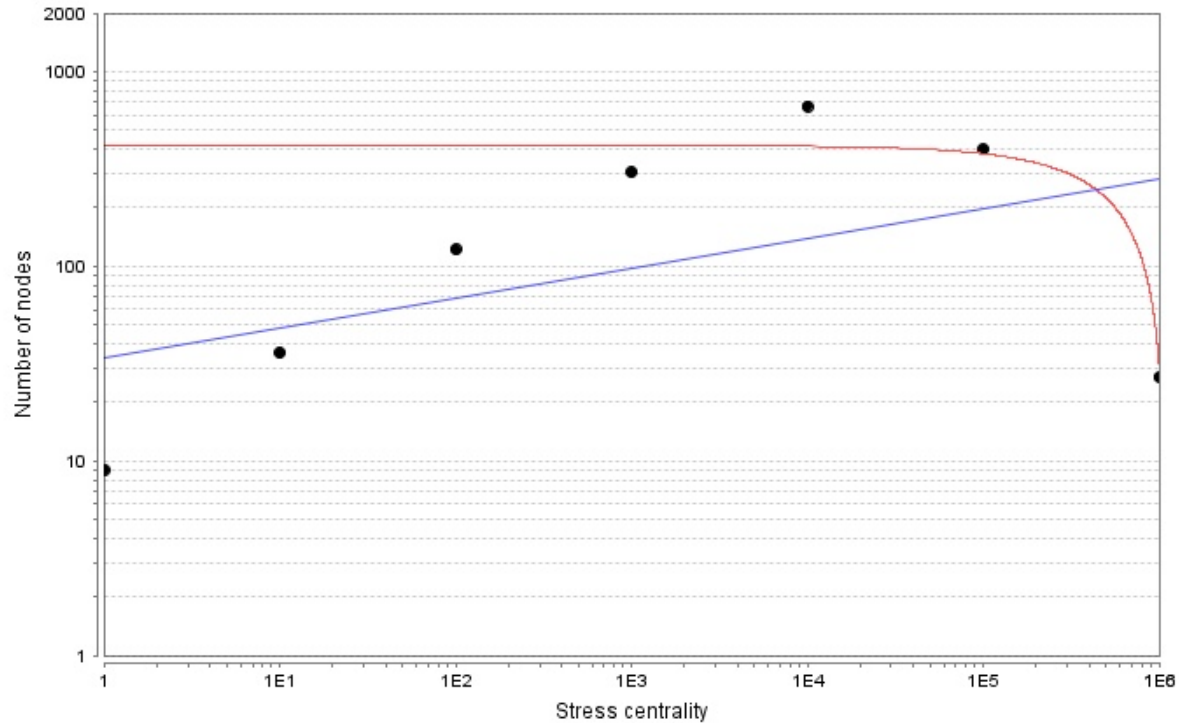
**Figure 5.4** The Indegree distribution and Closeness centrality as observed in DNA repair pathway.



**a. Average clustering coefficient**



**b. Betweenness centrality**

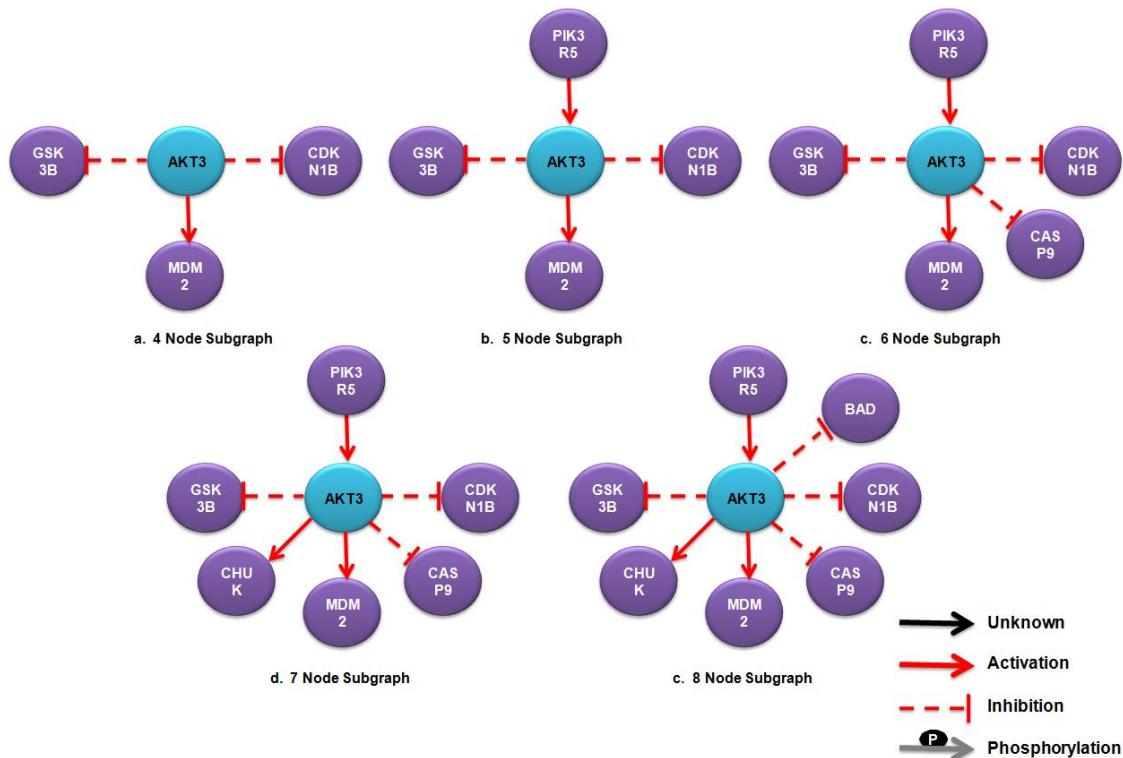


### c. Stress centrality

**Figure 5.5** The average clustering coefficient, betweenness centrality and stress centrality generated from DNA repair proteins and elucidated pathway.

For the comprehensive pathway level analyses, 13 DNA repair associated diseases, i.e. Fanconi anemia, Glioma, Non-small cell lung cancer, Colorectal cancer, Acute myeloid leukemia, Brain tumor, Chronic myeloid leukemia, Endometrial cancer, Necrosis, Pancreatic cancer, Prostate cancer, Renal cell carcinoma and Small-cell lung cancer were evaluated for their association in obscure biological networks. Here, vital network components i.e. small connected sub-networks occurring in significantly higher frequencies in a network than would be expected for a given random network were detected in each of 13 diseases. These patterns or motifs are considerably overrepresented and characterize certain essential functional aspects associated with diseases and underlying mechanisms. Several motifs ranging from 3-8 sub-graph nodes were generated and annotated for the 13 disease pathways and a few corresponding to prostate cancer pathway have been depicted in **Figure 5.6**. We searched for 3-node motifs but due to their absence in most of the diseases, we removed them from the final analyses. Annotation of these graphs was based on a statistical criterion using mean-frequencies, standard deviation,  $z$ -scores

and  $p$ -values. Further, SP was also calculated for estimating the importance of each motif for 13 diseases as illustrated in **Table 5.1** for Prostate cancer pathway.



**Figure 5.6** A few classified network motifs from the prostate cancer pathway.

**Table 5.1** Network motifs from prostate cancer pathway with their respective standard statistical parameters

Network motif ID (Adjacency matrix)	z-score	p-value	Significance Profile
000000000011100	3.18	0.001	0.42
000000000001110	2.97	0.004	0.39
0000100001000110	5.40	0.006	0.71
0000100010001100	2.07	0.011	0.27
0000100010000110	2.29	0.015	0.3
000000000000000000111100	3.66	0.004	0.31
0000000000000000000011110	2.85	0.008	0.24
000000000000101100000110	6.19	0.001	0.52
0000010000010000000101100	5.15	0.003	0.44
0000010000010000100001100	6.17	0.001	0.52
0000000010010001000001100	3.98	0.018	0.34
000000000000000000000000001111100	3.68	0.004	0.17
0000000000000000000000000000111110	2.95	0.007	0.14
0000000000000000000000001000101111000	7.79	0.001	0.36
00000000000000000000000010011000100100	2.82	0.014	0.13
0000000000000000100000101110000001010	7.62	0.001	0.36
000000000000100000000010011000000110	6.31	0.005	0.29
00000000000000100000010010000101000	2.25	0.028	0.1
000000000000010000000010010000101000	2.03	0.048	0.09
000000001000100000000001000100011000	10.47	0.001	0.49









Different identified patterns include SIM, MIM, feedback loops, auto-regulation, FFL and other putative regulatory motifs. All the detected network motifs have their own importance in biological networks and have imperative functions to execute; as in case of SIM motif, several genes are controlled by a single master gene and the master gene is known to be auto-regulatory. Whereas, in MIM motif (a generalization of SIM), a single gene is being controlled by multiple genes. This kind of biologically meaningful information is associated with all network motif types and correlated with their functional significance. These motifs are an important parameter to analyze since it not only describes the local properties of a network but are also functionally significant as they occur frequently and distinctively in a biological network. A wide variety of vital 4-8 node sub-graphs for all 13 DNA repair associated diseases has been described in **Table 5.2**.

**Table 5.2** A list of DNA repair associated diseases subjected to pathway level analysis

S. No	Diseases	4-node motifs	5-node motifs	6-node Motifs	7-node motifs	8-node Motifs
1	Acute myeloid leukemia	2	8	26	88	253
2	Chronic myeloid leukemia	-	2	6	15	33
3	Brain tumor	2	6	25	85	270
4	Colorectal cancer	3	3	8	17	29
5	Endometrial cancer	2	2	3	4	4
6	Fanconi anemia	-	-	1	3	7
7	Glioma	2	6	28	80	263
8	Non-small cell lung cancer	2	10	17	30	50
9	Necrosis	5	10	18	41	76
10	Pancreatic cancer	-	-	3	7	10
11	Prostate cancer	3	5	13	48	99
12	Renal cell carcinoma	3	4	7	14	14
13	Small cell lung cancer	4	13	21	35	54

As mentioned in **Table 5.1**, the statistical implication of these motifs was ensured by available measures like  $z$ -score,  $p$ -value and SP. The calculated SP for all motifs in 13 diseases revealed a similar pattern of graph as obtained in **Figure 4.7** of previous chapter. Therefore, our approach of reducing the disease pathways into smaller sub-graphs and subsequently identifying key players proves quite valuable. Based upon this SP profile analysis we suggest that network motifs with smaller node size (3 or 4) are more functionally allied towards their role in pathways

while motifs of larger size ( $\geq 5$  nodes) are less functional. It is believed that the observed trend might be similar in many such biological networks if analyzed. For an overview of the deliberated parameters applied on all 13 diseases, **Table 5.3** depicts the part of a table showing parameter's value for motifs in prostate cancer pathway. Further, the motif showing least ' $FT_i$ ' value i.e.  $0.003$  for motif ID ' $8bx$ ' was chosen for identifying the key players in the given motif. Here the patterns being referred are small connected sub-networks occurring in significantly higher frequencies in prostate cancer network than would be expected for a given random network. These patterns or motifs are considerably overrepresented and characterize essential functional aspects associated with disease specific pathways. This information was attained by mapping all the genes from 13 complex disease pathways onto the network motifs and then frequency of each gene for each network motif was calculated. Then the frequencies for all genes/proteins in the above mentioned motif (with least ' $FT_i$ ' values) were calculated and presented in **Table 5.4**. The proposed key candidate genes for prostate cancer identified via applied computational approaches revealed MDM2, CHUK, GSK3B, AKT3 and CDKN1B. In previous reports [29, 30], the regulatory functions of MDM2 and AKT3 are well explored for prostate cancer. The apoptotic response of prostate cancer to androgen deprivation is strongly influenced by MDM2 expression [29] and up-regulation of AKT3 [31] contributes to androgen-independent prostate cancer whereas the role of other three genes in the disease is not well understood.

The current study anticipated for identification of vital components in pursuit of reducing the complexity hidden in intricate disease pathways and their associated biological processes. Thus, the applied system-component level integrative approach to investigate these sub-structures in human DNA repair pathways is alleged to unravel the ambiguity regarding factors causing numerous precarious diseases related to DNA repair. Identification of crucial network motifs will help systems biologists to find key components from whole pathways and analyze their behaviour against different experimental conditions. Consequently, we made an effort to identify certain other genes that may potentially impact meticulous understanding of major repair allied diseases. Many important genes as revealed in **Table 5.4** were components of numerous cellular and signalling processes and were observed to contribute maximum complexity in the major carcinogenesis. These genes illustrated higher frequencies and numerous interactions

among nodes and are proposed to be vital for cancer progression. Here, the biological disease pathway complexity has been reduced to a few key genes that may be explored further for their putative roles in the disease.

**Table 5.3 Values of the designed parameters for each recurrent motif in prostate cancer pathway**

Network Motif ID (Adjacency matrix)	Symbols	$FTN_i$	$FN_i$	$FTi$
'0000100001000110'	4a	36	12	0.333
'0000000000011100'	4b	444	20	0.045
'0000000000001110'	4c	336	10	0.03
'0000000000001110'	4f	336	10	0.03
'000000000000101100000110'	5a	180	12	0.067
'0000010000010000100001100'	5b	45	13	0.289
'0000010000010000000101100'	5c	45	13	0.289
'00000000000000000000111100'	5d	1260	13	0.01
'00000000000000000000111100'	5i	1260	13	0.01
'00000000000000000000011110'	5j	630	10	0.016
'000000000000100000000010011000000110'	6u	192	16	0.083
'000000000000000000000000000000001111100'	6v	2268	13	0.006
'00000000000000000000000010011000100100'	6x	264	16	0.061
'0000000000000000100000010010000101000'	6y	162	23	0.142
'0000000000000000000001000001000100000010'	7e	49	13	0.265
'0000000000000000000001000001000100000010'	7f	63	15	0.238
'000000000000000000000010100000010000000100011001000'	7r	112	16	0.143
'000000000000000000000000000000001000000100010011110000'	7s	588	13	0.022
'00000000000000000000000000000000000000000000000001001111010'	7au	1960	14	0.007
'0000000000000000000001010000000101001000010'	7bj	63	15	0.238
'00000000000000000000000000000000000000000000000001000010001110100'	7cn	1176	14	0.012
'000000000000000000000000000000000000000000000000011001110010'	7cq	392	12	0.031
'000000000000000000000000000000000000000000000000010011000001000010100'	8l	192	17	0.089
'000000000000000000000000000000000000000000000000010100000001000000000110010000010'	8m	64	15	0.234
'00000000000000000000000000000000000000000000000001000000000000001000001010000'	8av	616	23	0.037
'00000000000000000000000000000000000000000000000001000000000010010010011001000'	8bw	896	17	0.019
<b>'000000000000000000000000000000000000000000000000010000000000111110100'</b>	<b>8bx</b>	<b>6720</b>	<b>17</b>	<b>0.003</b>
'000000000000000000000000000000000000000000000000010000000110000000010000100001100'	8ev	40	11	0.275
'0000000000000000000000000000000000000000000000000100000010000000000010000010110011000'	8ew	280	16	0.057
'00000000000000000000000000000000000000000000000001000000100001100011100000'	8ex	672	14	0.021
'00000000000000000000000000000000000000000000000001000001000110011110000'	8fk	1008	13	0.013
'0000000000000000000000000000000000000000000000000100001100011100100'	8fl	448	13	0.029
'00000000000000000000000000000000000000000000000001000000100000000000100010010000'	8fm	168	16	0.095
'00000000000000000000000000000000000000000000000001000001000000000000011100000000011000'	8ge	72	13	0.181
'000000000000000000000000000000000000000000000000010000000100000010000000000110000'	8gf	72	16	0.222
'0000000000000000000000000000000000000000000000000100100000100011000001001000'	8gh	64	15	0.234

**Table 5.4 List of identified key proteins in DNA repair diseases using newly designed parameters**

S. No	DNA repair associated diseases	No. of proteins in pathway	No. of key proteins	Key proteins *	Experimentally validated proteins (Pubmed IDs)
1.	Fanconi Anemia	57	8	<b>RAD51, FAN1, BRCA2, FANCM, FANCG, FANCI, FANCD2, FANCA</b>	RAD51 (12239151, 12483114), BRCA2 (19530235, 14559878), FANCM (18285517), FANCG and FANCA (11050007), FANCD2 (25455269)
2.	Glioma	80	4	<b>HRAS, PIK3R5, PRKCA, EGFR</b>	EGFR and HRAS (10072878, 8017863)
3.	Non-Small Cell Lung Cancer	56	8	<b>FOXO3, BAD, CASP9, ERBB2, PDPK1, GRB2, AKT3, PIK3CA</b>	BAD (15870947), CASP9 (17291493), ERBB2 (15902485), PDPK1 (14614329), PIK3CA (16170026)
4.	Colorectal Cancer	44	5	<b>KRAS, ARAF, PIK3R5, AKT3, RALGDS</b>	KRAS (19515263, 15069679, 10545700), AKT3 (18813315)
5.	Acute Myeloid Leukemia	44	5	<b>FLT3, CHUK, MTOR, AKT3, PIK3R5</b>	FLT3 (15778081), MTOR (16939811)
6.	Brain Tumor	80	4	<b>PIK3R5, HRAS, EGFR, PRKCA</b>	HRAS (20425820), EGFR (17236582)
7.	Chronic Myeloid Leukemia	58	3	ABL1, AKT3, PIK3R5	-
8.	Endometrial Cancer	43	6	<b>HRAS, SOS1, GRB2, EGFR, AKT3, PIK3R5</b>	EGFR (7978930)
9.	Necrosis	142	7	<b>BRCA2, FANCD2, FANCI, FANCM, FANCA, FANCG, RAD51</b>	BRCA2 (9665145), RAD51 (11782381), FANCA and FANCG (15299030)
10.	Pancreatic Cancer	60	5	<b>KRAS, PIK3R5, AKT3, ARAF, RALGDS</b>	KRAS (18075308)
11.	Prostate Cancer	60	5	<b>MDM2, CHUK, GSK3B, AKT3, CDKN1B</b>	MDM2 (15176048), AKT3 (16721361)
12.	Renal Cell Carcinoma	64	4	<b>ARNT, CREBBP, EPAS1, EGLN2</b>	EPAS1 (11301389)
13.	Small Cell Lung Cancer	51	5	<b>AKT3, PIK3R5, CHUK, NFKB1A, NFKB1</b>	-

\* Key proteins (in bold) with experimental evidences for their implication in DNA repair associated disorders

The study reveals important markers and a few novel genes that are believed to associate with DNA repair diseases. The genes reported in study via the novel applied methodology can be studied broadly for its association in respective diseases. Moreover, the novel parameters designed present the dependence of an entire system on a few key genes, proteins and metabolites for examining the statistical significance. Hence, the proposed genes from comprehensive theoretical and computational analysis implicated in diseases may serve as imperative therapeutic targets. There is an imperative need to apply this approach on other diseases as well to identify crucial network components and candidate markers. It is believed that besides key genes proposed in this study, we provide novel methodology to analyze small components of large and complex biological networks. Further, investigating and targeting these proposed genes for experimental validations, instead being spellbound by the complicated pathway will certainly endow valuable insight in a well-timed systematic understanding of the disease. It is anticipated that the performed analysis on available biological data and the proposed pipeline would serve as a useful accompaniment for analyzing candidate markers in human repair specific disease pathways. This study will be of utmost use to researchers who are focused in developing therapeutic targets for precarious diseases like multi-system defects, cancers, skin diseases and neurodegenerative disorders implicated in DNA repair system.

#### **5.4 CONCLUSION**

Neither the study based on DNA repair pathways connecting vital biological entities in complicated interacting pathways has been performed nor the functional enrichment of network motifs from repair specific disease networks been hitherto executed. Thus, the novel parameters designed in study comprehend the dependence of an entire system on a few key genes, proteins and metabolites for examining the statistical significance in 13 DNA repair associated disorders. On the whole, a holistic approach was practiced that includes various aspects of molecular data, biomarkers, networks and pathways for uncovering the intricacy in disease specific pathways and then confining the search to only a small number of proteins or network components that may answer diverse biological queries concerning diseases. Investigating these proteins would undoubtedly provide gratifying insights to the disease and underlying mechanisms. Additionally, this *in silico* approach could be applied to other diseases in quest for identifying candidate biomarkers and will aid experimental biologists to identify novel targets for diseases. These

computational techniques are not only time efficient but also cost effective and provides essential clues to experimental biologists for carrying their experiments and validating our observations.

## REFERENCES

- [1] J. H. J. Hoeijmakers, "Genome maintenance mechanisms for preventing cancer," *Nature*, vol. 411, pp. 366-374, 2001.
- [2] A. L. Jackson and L. A. Loeb, "The contribution of endogenous sources of DNA damage to the multiple mutations in cancer," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 477, pp. 7-21, 2001.
- [3] T. Lindahl, "DNA repair enzymes," *Annual review of biochemistry*, vol. 51, pp. 61-87, 1982.
- [4] C. E. Lilley, R. A. Schwartz, and M. D. Weitzman, "Using or abusing: viruses and the cellular DNA damage response," *Trends in microbiology*, vol. 15, pp. 119-126, 2007.
- [5] M. C. Moraes, J. B. Neto, and C. F. Menck, "DNA repair mechanisms protect our genome from carcinogenesis," *Front Biosci*, vol. 17, pp. 1362-1388, 2012.
- [6] J. Knoch, Y. Kamenisch, C. Kubisch, and M. Berneburg, "Rare hereditary diseases with defects in DNA-repair," *European Journal of Dermatology*, vol. 22, pp. 443-455, 2012.
- [7] S. V. Boychuk and B. R. Ramazanov, "DNA Repair System Defects—Role in Oncogenesis and Cancer Therapy," *Kazanskiy meditsinskiy zhurnal*, vol. 95, pp. 307-314, 2014.
- [8] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, pp. 450-461, 2007.
- [9] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of *Escherichia coli*," *Nature genetics*, vol. 31, pp. 64-68, 2002.
- [10] W. Kim, M. Li, J. Wang, and Y. Pan, "Biological network motif detection and evaluation," *BMC systems biology*, vol. 5, p. S5, 2011.
- [11] M. E. Smoot, K. Ono, J. Ruscheinski, P. L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, pp. 431-432, 2011.
- [12] G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC bioinformatics*, vol. 4, p. 2, 2003.

- 
- [13] Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, pp. 282-284, 2008.
- [14] S. Maslov and K. Sneppen, "Specificity and stability in topology of protein networks," *Science*, vol. 296, pp. 910-913, 2002.
- [15] M. E. J. Newman, "A measure of betweenness centrality based on random walks," *Social networks*, vol. 27, pp. 39-54, 2005.
- [16] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, pp. 101-113, 2004.
- [17] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-442, 1998.
- [18] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, "Hierarchical organization of modularity in metabolic networks," *science*, vol. 297, pp. 1551-1555, 2002.
- [19] J. Yoon, A. Blumer, and K. Lee, "An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality," *Bioinformatics*, vol. 22, pp. 3106-3108, 2006.
- [20] U. Brandes, "A faster algorithm for betweenness centrality\*," *Journal of Mathematical Sociology*, vol. 25, pp. 163-177, 2001.
- [21] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research*, vol. 28, pp. 27-30, 2000.
- [22] L. Matthews, G. Gopinath, M. Gillespie, M. Caudy, D. Croft, B. de Bono, P. Garapati, J. Hemish, H. Hermjakob, and B. Jassal, "Reactome knowledgebase of human biological pathways and processes," *Nucleic acids research*, vol. 37, pp. D619-D622, 2009.
- [23] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic acids research*, vol. 34, pp. D535-D539, 2006.
- [24] S. Wernicke and F. Rasche, "FANMOD: a tool for fast network motif detection," *Bioinformatics*, vol. 22, pp. 1152-1153, 2006.
- [25] F. Schreiber and H. Schwöbbermeyer, "MAVisto: a tool for the exploration of network motifs," *Bioinformatics*, vol. 21, pp. 3572-3574, 2005.

- [26] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon, "Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs," *Bioinformatics*, vol. 20, pp. 1746-1758, 2004.
- [27] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, pp. 1129-1164, 1991.
- [28] S. Wernicke, "Efficient detection of network motifs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 3, pp. 347-359, 2006.
- [29] Z. Mu, P. Hachem, S. Agrawal, and A. Pollack, "Antisense MDM2 oligonucleotides restore the apoptotic response of prostate cancer cells to androgen deprivation," *The Prostate*, vol. 60, pp. 187-196, 2004.
- [30] T. Kirkegaard, C. J. Witton, J. Edwards, K. V. Nielsen, L. B. Jensen, F. M. Campbell, T. G. Cooke, and J. M. S. Bartlett, "Molecular alterations in AKT1, AKT2 and AKT3 detected in breast and prostatic cancer by FISH," *Histopathology*, vol. 56, pp. 203-211, 2010.
- [31] K. Nakatani, D. A. Thompson, A. Barthel, H. Sakaue, W. Liu, R. J. Weigel, and R. A. Roth, "Up-regulation of Akt3 in Estrogen Receptor-deficient Breast Cancers and Androgen-independent Prostate Cancer Lines," *Journal of Biological Chemistry*, vol. 274, pp. 21528-21532, 1999.



*“If we knew what it was we were doing, it would not be called research, would it?”*

*-Albert Einstein*

# OVERALL CONCLUSIONS AND FUTURE PROSPECTS

## **CONCLUSIONS**

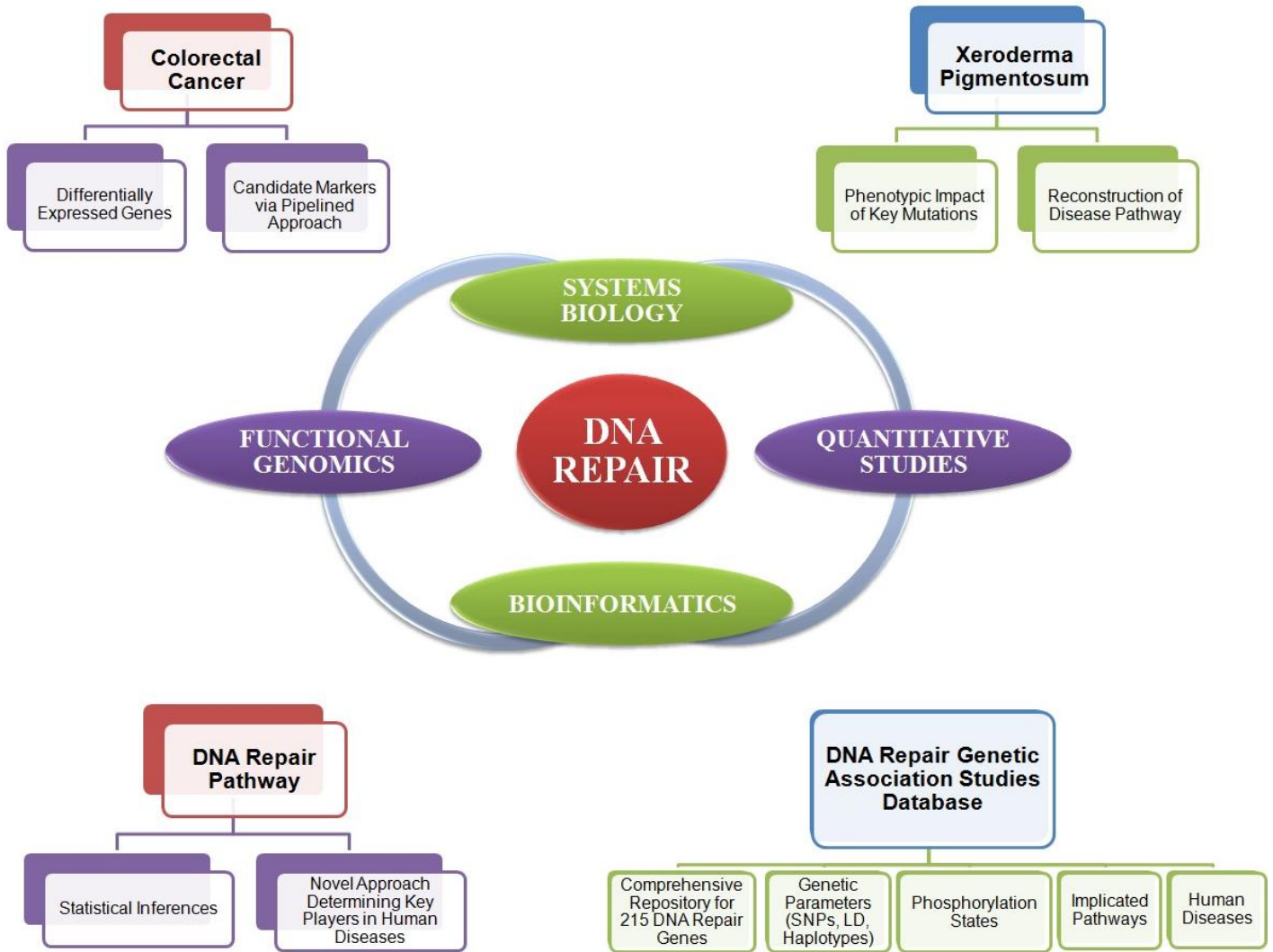
We broadly analyzed DNA repair genes/proteins, genetic variations, PPIs, intricate networks and pathways concerned with human DNA repair system. The critical role of DNA repair mechanisms in plethora of diseases such as oncogenesis, genetic abnormalities and pre-mature aging is deliberated through computational means. The overall study aimed in widespread assessment of DNA repair associated diseases and engrossed mechanisms as evident from the created database. The overall top-down approach has been utilized for deciphering candidate markers for XP, CRC and other important DNA repair related diseases. Various biomolecules such as genes, TFs, proteins, vital regulatory elements and interactions among them have been elucidated at systems level for better understanding of the bioprocesses concerned with DNA repair. Future prospects and the practical applications of our approach are also discussed briefly to provide new directions for the DNA repair related research.

Important findings of this thesis are summarized below and presented in a connected manner in **Figure 6.1**.

- The developed database on human DNA repair system, named DR-GAS is a wide-ranging catalog for DNA repair genes that includes information on quantitative genetic parameters like LD, haplotypes, SNPs, disease allied information and their phosphorylation states for 215 DNA repair genes. This repository is unique and first of its kind model since there is no such resource available till date for DNA repair system which provides appropriate classification of DNA repair genes in associated mechanisms as presented by DR-GAS (i.e. 16 major classes along with quantitative genetic parameters and phosphorylation sites). This database will assist researchers to study the repair genes in depth and will provide useful insight for future analysis and studies. It also presents insights into some well-known mutations in these DNA repair genes that may be further analyzed by researchers at length. This database will be of utmost use to the scientists focused in dealing with DNA damages, repair and concerned in developing therapeutic targets for precarious diseases implicated in DNA repair system.
- The systems level vision on XP facilitated a broad perspective on the disease and explicated the role of mutations in 8 DNA repair genes. These mutations and their impact on the structure and caught up interactions have been analyzed through simulation studies. A

putative reconstructed model for the XP pathway is also proposed in the study which focuses on both NER and TLS mechanisms and allied aberrations. The applied approach is estimated to be constructive in figuring out unknown details immersed in XP which indeed will assist in applying suitable therapeutic strategies for skin cancer and other cancers predisposed to XP.

- In our next objective, an integrative approach comprising of various aspects of biological data such as gene expression, network components and pathways was applied for uncovering the intricacy in CRC. Various differentially expressed genes in the diseased state along with TFs, promoter elements and other regulatory factors are identified. The intact CRC pathway was then analyzed for important patterns governing the severity of disease using available standard statistical constraints. In the study, a few novel parameters specific for the analysis of vital patterns in CRC pathway were designed to deduce the importance of these patterns or motifs. On the basis of these parameters, 5 key genes i.e. KRAS, ARAF, PIK3R5, RALGDS and AKT3 illustrating high statistical significance in CRC were revealed. Analyzing these genes for their association in CRC and disease progression, rather than analyzing pathway as a whole would definitely confine the search to only a few genes or network components. The whole is better than sum of its parts, so we applied this approach to analyze pathways for many diseases linked to DNA repair systems.
- As mentioned, the final objective focussed on the extensive analysis of DNA repair proteins, PPIs and implicated pathways. The DNA repair proteins mapped to human interactome along with the statistical analysis of complex network revealed key neighboring interactions and valuable inferences from the generated graphs. The 13 DNA repair linked diseases were also subjected to pathway-level analysis and for determination of key players; the novel parameters deliberated in previous study were utilized. This approach facilitated identification of major contributing agents underlying major malignancies and genetic disorders. In the study, various proteins in disease pathways ranging from 43 to 142 implicated proteins have been reduced to 3 to 8 proteins in respective 13 analyzed diseases through computationally applied novel methodology. Instead studying all the proteins in disease pathways, we supplied major candidates that may be experimentally verified and thus can serve as important therapeutic targets for the 13 diseases.



**Figure 6.1** Representation of the overall applied approach in the fulfillment of proposed objectives in the thesis.

## FUTURE PROSPECTS

- It is anticipated that this web based comprehensive resource would serve as a valuable accompaniment for analyzing DNA repair systems for human and will also contribute scientific knowledge towards better understanding of other mammalian repair systems. An agile approach is implemented for the design of DR-GAS so that expansion and updation could be done on regular basis to provide state-of-the-art information to the scientific community. DR-GAS will be updated on regular basis and is believed that this resource will

not only assist molecular biologists but also therapeutic developers and other scientific community to encounter biologically meaningful information.

- The proposed mutations in 8 DNA repair genes leading to the development of XP can be experimentally tested and the consequences of these damages could be observed. The analysis of these mutations at structural level and their deleterious impact in influencing disease pathway provides insights in the progression of XP. Identification of these mutations in 8 repair genes of a patient can serve as plausible biomarkers for the disease.
- The study on differential expression and pathway-level analysis of CRC through standard as well as novel deliberated parameters bestowed 5 key genes. Further, investigating and targeting these proposed genes for experimental validations will certainly endow valuable insight in opportune understanding of the disease. Also, this *in silico* approach could be applied to other diseases in quest for identifying biomarkers and will assist the scientific community to identify novel candidate markers. In addition, extra detailed implementation of the concepts of graph theory needs to be applied and hence established.
- The projected key candidates for the 13 DNA repair diseases may be authenticated experimentally thus saving time and resources; contributing to timely systematic understanding of complex diseases. The novel parameters deliberated for identifying key candidates in DNA repair associated disease pathways could be applied on other diseases too having known biological pathways. In order to achieve a wide perspective and apply knowledge from these designed parameters to any disease pathway, the pipeline needs to be automated. Therefore, in future a tool could be planned on the basis of proposed pipeline to attain key signatures in intricate biological pathways.
- A study of all disease pathways needs to be done at an instance in order to search out the probable targets that will facilitate prediction of the same center of interest targets for related and unrelated diseases. This kind of approach will prove valuable in cases with multiple disorders such as if a patient suffers from liver cancer, arthritis and asthma altogether. In this

scenario, analyzing all conditions simultaneously will eventually give out common targets for curing the disease or preventing it from getting more malicious.

- In a scenario when we accomplish comprehensive knowledge regarding disease pathways and have annotated all the nodes and edges of the pathway; we can successfully discover suitable targets in pathways with perfection. All the projected biomolecules (Chapters 2-5) would serve as potential candidates for experimental validations and further assist in therapeutic interventions. The biologically meaningful information generated through computational analyses would be of utmost use to scientific community after experimental verifications.

*“Life is like riding a bicycle. To keep your balance, you must keep moving.”*

*-Albert Einstein*

# APPENDIX

## Appendix A

### 1. Haplotype blocks identified in XP related genes

Gene (Gene IDs)	Blocks	Number of Markers	Important (Putative) Markers <sup>1</sup>
DDB2 (1643)	1	14	3
ERCC2 (2068)	1, 2, 3	3, 2, 9	3, 2, 3
ERCC3 (2071)	1	21	3
ERCC4 (2072)	1	18	2
ERCC5 (2073)	1	32	4
POLH (5429)	1	3	1
XPA (7507)	1	10	3
XPC (7508)	1	30	3

### 2. Important phosphorylation sites in the eight NER and TLS genes

Proteins	Protein ID	Position	Sequence	Score	Phosphorylated Residue
DDB2	NP_000098	13	TQKTSEIVL	0.739	*S*
DDB2	NP_000098	24	RNKRSRSPL	0.834	*S*
DDB2	NP_000098	26	KRSRSPLEL	0.997	*S*
DDB2	NP_000098	45	GSGPSRRCD	0.97	*S*
DDB2	NP_000098	50	RRCDSCLW	0.681	*S*
DDB2	NP_000098	82	LGRASWPSV	0.795	*S*
DDB2	NP_000098	116	RRATSLAWH	0.987	*S*
DDB2	NP_000098	131	VAVGSKGGD	0.64	*S*
DDB2	NP_000098	175	QFYASSMEG	0.881	*S*
DDB2	NP_000098	176	FYASSMEGT	0.585	*S*
DDB2	NP_000098	310	TDQKSEIRV	0.979	*S*
DDB2	NP_000098	316	IRVYSASQW	0.872	*S*
DDB2	NP_000098	318	VYSASQWDC	0.67	*S*
DDB2	NP_000098	379	FDGNSGKMM	0.53	*S*
DDB2	NP_000098	392	DPSSGSISS	0.529	*S*
DDB2	NP_000098	395	SSGISSLNE	0.507	*S*
DDB2	NP_000098	419	ILIWSQEEA	0.804	*S*
DDB2	NP_000098	9	KRPETQKTS	0.918	*T*
DDB2	NP_000098	115	DRRATSLAW	0.892	*T*
DDB2	NP_000098	180	SMEGTTRLQ	0.895	*T*
DDB2	NP_000098	266	SVDQTVKIW	0.561	*T*
DDB2	NP_000098	305	ARLLTTDQK	0.595	*T*
DDB2	NP_000098	338	FQHLTPIKA	0.609	*T*
DDB2	NP_000098	365	FKSCTPYEL	0.794	*T*
DDB2	NP_000098	371	YELRTIDVF	0.504	*T*
DDB2	NP_000098	406	PMGDTLASA	0.608	*T*
DDB2	NP_000098	367	SCTPYELRT	0.78	*Y*
ERCC2	NP_000391	23	PEQFSYMRE	0.643	*S*

<sup>1</sup> The putative markers comprise of those vital markers whose population frequencies are  $\geq 0.1$



ERCC2	NP_000391	111	LALSSRKNL	0.921	*S*
ERCC2	NP_000391	222	ADLVSKELA	0.902	*S*
ERCC2	NP_000391	296	LREASAARE	0.84	*S*
ERCC2	NP_000391	322	AVPGSIRTA	0.772	*S*
ERCC2	NP_000391	381	ERLRSLLHT	0.672	*S*
ERCC2	NP_000391	436	ILHFSCMDA	0.621	*S*
ERCC2	NP_000391	453	ERFQSVIIT	0.622	*S*
ERCC2	NP_000391	462	SGTLSPLDI	0.995	*S*
ERCC2	NP_000391	505	QVAISSKFE	0.984	*S*
ERCC2	NP_000391	506	VAISSKFET	0.836	*S*
ERCC2	NP_000391	551	STVASWYEQ	0.992	*S*
ERCC2	NP_000391	740	LSLLSLEQL	0.919	*S*
ERCC2	NP_000391	746	EQLESEETL	0.624	*S*
ERCC2	NP_000391	122	HPEVTPLRF	0.884	*T*
ERCC2	NP_000391	255	LTRRTLDRC	0.538	*T*
ERCC2	NP_000391	427	DRTPTIANP	0.672	*T*
ERCC2	NP_000391	571	LFIETQDGA	0.609	*T*
ERCC2	NP_000391	577	DGAETSVAL	0.854	*T*
ERCC2	NP_000391	672	IRGKTDYGL	0.722	*T*
ERCC2	NP_000391	749	ESEETLKRI	0.842	*T*
ERCC2	NP_000391	542	FFTSYQYME	0.563	*Y*
ERCC2	NP_000391	584	ALEKYQEAC	0.667	*Y*
ERCC2	NP_000391	625	FGVPYVYTQ	0.643	*Y*
ERCC2	NP_000391	674	GKTDYGLMV	0.657	*Y*
ERCC3	NP_000113	14	DKKKSRRKH	0.998	*S*
ERCC3	NP_000113	49	QVDESGTKV	0.948	*S*
ERCC3	NP_000113	144	LRKLSKTGV	0.88	*S*
ERCC3	NP_000113	202	RLRNSEGEA	0.994	*S*
ERCC3	NP_000113	216	ETFTSKSAI	0.757	*S*
ERCC3	NP_000113	218	FTSKSAISK	0.522	*S*
ERCC3	NP_000113	221	KSAISKTAE	0.978	*S*
ERCC3	NP_000113	226	KTAESSGGP	0.968	*S*
ERCC3	NP_000113	227	TAESSGGPS	0.813	*S*
ERCC3	NP_000113	231	SGGPSTSRV	0.902	*S*
ERCC3	NP_000113	242	PQGKSDIPM	0.544	*S*
ERCC3	NP_000113	301	FRNDSVNP	0.811	*S*
ERCC3	NP_000113	323	YQEKSLRKM	0.696	*S*
ERCC3	NP_000113	370	NSAVSVEQW	0.995	*S*
ERCC3	NP_000113	382	FKMWSTIDD	0.775	*S*
ERCC3	NP_000113	387	TIDDSQICR	0.733	*S*
ERCC3	NP_000113	394	CRFTSDAKD	0.785	*S*
ERCC3	NP_000113	420	TTKRWEAE	0.988	*S*
ERCC3	NP_000113	515	WCPMSPEFY	0.938	*S*
ERCC3	NP_000113	585	YGPTSQGER	0.989	*S*
ERCC3	NP_000113	614	VGDTSFDP	0.504	*S*
ERCC3	NP_000113	632	SHGGSRRQE	0.951	*S*
ERCC3	NP_000113	664	YSLVSQDTQ	0.742	*S*
ERCC3	NP_000113	673	EMAYSTKRQ	0.988	*S*
ERCC3	NP_000113	704	DLAFSTKEE	0.995	*S*
ERCC3	NP_000113	738	GSRSSQASR	0.962	*S*
ERCC3	NP_000113	741	SSQASRRFG	0.994	*S*
ERCC3	NP_000113	749	GTMSSMSGGA	0.651	*S*
ERCC3	NP_000113	751	MSSMSGADD	0.996	*S*

ERCC3	NP_000113	763	MEYHSSRSK	0.86	*S*
ERCC3	NP_000113	766	HSSRSKAPS	0.846	*S*
ERCC3	NP_000113	770	SKAPSKHVH	0.991	*S*
ERCC3	NP_000113	51	DESGTKVDE	0.879	*T*
ERCC3	NP_000113	72	KDDHTSRPL	0.663	*T*
ERCC3	NP_000113	111	VCRPTHVHE	0.587	*T*
ERCC3	NP_000113	133	VGLQTSBIT	0.837	*T*
ERCC3	NP_000113	215	TETFTSKSA	0.778	*T*
ERCC3	NP_000113	232	GGPSTSRVT	0.986	*T*
ERCC3	NP_000113	265	EEEETQTVS	0.626	*T*
ERCC3	NP_000113	356	TAACTVRKR	0.758	*T*
ERCC3	NP_000113	383	KMWSTIDDS	0.776	*T*
ERCC3	NP_000113	393	ICRFTSDAK	0.704	*T*
ERCC3	NP_000113	417	LGHTTKRSW	0.921	*T*
ERCC3	NP_000113	456	RRVLTIVQA	0.97	*T*
ERCC3	NP_000113	527	VAIKTKKRI	0.811	*T*
ERCC3	NP_000113	613	KVGDTSFDL	0.768	*T*
ERCC3	NP_000113	667	VSQDTQEMA	0.711	*T*
ERCC3	NP_000113	674	MAYSTKRQR	0.807	*T*
ERCC3	NP_000113	756	GADDTVYME	0.84	*T*
ERCC3	NP_000113	19	RKRHYEED	0.852	*Y*
ERCC3	NP_000113	56	KVDEYGAKD	0.731	*Y*
ERCC3	NP_000113	93	FSPVYKYAQ	0.669	*Y*
ERCC3	NP_000113	252	LFDFYEQMD	0.649	*Y*
ERCC3	NP_000113	501	QNGYIAKV	0.555	*Y*
ERCC3	NP_000113	758	DDTVYMEYH	0.825	*Y*
ERCC4	NP_005227	94	NEITSNSRY	0.971	*S*
ERCC4	NP_005227	96	ITSNSRYEV	0.827	*S*
ERCC4	NP_005227	125	DRIPSDLIT	0.633	*S*
ERCC4	NP_005227	246	CHNPSLEVE	0.921	*S*
ERCC4	NP_005227	253	VEDLSLENA	0.981	*S*
ERCC4	NP_005227	299	LQYLSQYDC	0.753	*S*
ERCC4	NP_005227	312	NLLESLRAT	0.862	*S*
ERCC4	NP_005227	352	DAKMSKKEK	0.996	*S*
ERCC4	NP_005227	358	KEKISEKME	0.995	*S*
ERCC4	NP_005227	418	DRTCSQLRD	0.713	*S*
ERCC4	NP_005227	444	FEKDSKAE	0.99	*S*
ERCC4	NP_005227	458	RKEDSSKRI	0.996	*S*
ERCC4	NP_005227	459	KEDSSKRIR	0.983	*S*
ERCC4	NP_005227	465	RIRKSHKRP	0.998	*S*
ERCC4	NP_005227	479	KERASTKER	0.998	*S*
ERCC4	NP_005227	519	RREISSSPE	0.914	*S*
ERCC4	NP_005227	520	REISSSPES	0.978	*S*
ERCC4	NP_005227	521	EISSSPESC	0.995	*S*
ERCC4	NP_005227	524	SSPESCPEE	0.99	*S*
ERCC4	NP_005227	613	IYGGSTEEQ	0.939	*S*
ERCC4	NP_005227	639	REKASMVVP	0.918	*S*
ERCC4	NP_005227	697	SELPSLIHR	0.545	*S*
ERCC4	NP_005227	728	VERKSISDL	0.996	*S*
ERCC4	NP_005227	768	SKPFLTSR	0.657	*S*
ERCC4	NP_005227	785	SNDISSKLT	0.933	*S*
ERCC4	NP_005227	805	LWCPSPHAT	0.516	*S*
ERCC4	NP_005227	835	ITADSETLP	0.697	*S*

ERCC4	NP_005227	841	TLPSEKYN	0.987	*S*
ERCC4	NP_005227	880	LAALSQDEL	0.928	*S*
ERCC4	NP_005227	910	AEVVSCKGKG	0.717	*S*
ERCC4	NP_005227	89	PRRVTNEIT	0.907	*T*
ERCC4	NP_005227	110	VIFATSRIL	0.565	*T*
ERCC4	NP_005227	265	PFDKTIRHY	0.831	*T*
ERCC4	NP_005227	316	SLRATEKAF	0.867	*T*
ERCC4	NP_005227	369	EGEETKKEL	0.992	*T*
ERCC4	NP_005227	385	WEALTEVLK	0.562	*T*
ERCC4	NP_005227	480	ERASTKERT	0.974	*T*
ERCC4	NP_005227	484	TKERTLKKK	0.714	*T*
ERCC4	NP_005227	586	DAELTFVRQ	0.565	*T*
ERCC4	NP_005227	669	VSTDTRKAG	0.837	*T*
ERCC4	NP_005227	719	DYILTPEMC	0.765	*T*
ERCC4	NP_005227	770	PFSLTSRGA	0.961	*T*
ERCC4	NP_005227	71	AEEYFINQ	0.979	*Y*
ERCC4	NP_005227	98	SNSRYEVYT	0.923	*Y*
ERCC4	NP_005227	101	RYEVYTQGG	0.691	*Y*
ERCC4	NP_005227	514	VEEGYRREI	0.586	*Y*
ERCC4	NP_005227	564	CSDPYALTR	0.882	*Y*
ERCC4	NP_005227	577	VEPRYVVLY	0.972	*Y*
ERCC4	NP_005227	619	EEQRYLTAL	0.883	*Y*
ERCC4	NP_005227	716	EVGDYILTP	0.828	*Y*
ERCC4	NP_005227	751	SMSRYKRP	0.68	*Y*
ERCC4	NP_005227	844	ESEKYNPGP	0.923	*Y*
ERCC5	NP_000114	18	GRQVSPEAL	0.997	*S*
ERCC5	NP_000114	49	RHGNSIENP	0.972	*S*
ERCC5	NP_000114	99	KDLASSDSR	0.779	*S*
ERCC5	NP_000114	100	DLASSDSRK	0.964	*S*
ERCC5	NP_000114	102	ASSDSRKT	0.992	*S*
ERCC5	NP_000114	125	TAFRSKRDE	0.995	*S*
ERCC5	NP_000114	156	EEKHSSEEE	0.998	*S*
ERCC5	NP_000114	157	EKHSSEED	0.998	*S*
ERCC5	NP_000114	194	SEDFSSLPP	0.622	*S*
ERCC5	NP_000114	230	SDDFSQYQL	0.949	*S*
ERCC5	NP_000114	259	NQQHSGHIR	0.633	*S*
ERCC5	NP_000114	283	RRVSEDTS	0.972	*S*
ERCC5	NP_000114	310	ESLPSSSKM	0.645	*S*
ERCC5	NP_000114	311	SLPSSSKMH	0.99	*S*
ERCC5	NP_000114	324	DVKSSPCEK	0.994	*S*
ERCC5	NP_000114	341	ATPPSPRTL	0.961	*S*
ERCC5	NP_000114	355	ALLGSSSEE	0.871	*S*
ERCC5	NP_000114	356	LLGSSSEEE	0.988	*S*
ERCC5	NP_000114	357	LGSSSEEL	0.995	*S*
ERCC5	NP_000114	384	EGSISPRTL	0.98	*S*
ERCC5	NP_000114	389	PRTLSAIKR	0.609	*S*
ERCC5	NP_000114	423	MRINSSTEN	0.981	*S*
ERCC5	NP_000114	424	RINSSTENS	0.994	*S*
ERCC5	NP_000114	428	STENSDEGL	0.985	*S*
ERCC5	NP_000114	449	TLASSSVNS	0.754	*S*
ERCC5	NP_000114	450	LASSSVNSA	0.895	*S*
ERCC5	NP_000114	453	SSVNSAEH	0.997	*S*
ERCC5	NP_000114	460	EHVASTNEG	0.977	*S*

ERCC5	NP_000114	470	EPTDSVPKE	0.548	*S*
ERCC5	NP_000114	489	AFPISDESM	0.709	*S*
ERCC5	NP_000114	492	ISDESMIKD	0.967	*S*
ERCC5	NP_000114	511	VVRHSDAPG	0.98	*S*
ERCC5	NP_000114	526	LTPASPTCT	0.862	*S*
ERCC5	NP_000114	532	TCTNSVSKN	0.588	*S*
ERCC5	NP_000114	534	TNSVSKNET	0.707	*S*
ERCC5	NP_000114	554	CPYESKFDS	0.896	*S*
ERCC5	NP_000114	559	KFDSSLLSS	0.852	*S*
ERCC5	NP_000114	562	SSLLSSDDE	0.993	*S*
ERCC5	NP_000114	563	SLLSSDDET	0.994	*S*
ERCC5	NP_000114	573	CKPNSASEV	0.959	*S*
ERCC5	NP_000114	582	IGPVSLQET	0.871	*S*
ERCC5	NP_000114	591	SSIVSVPSE	0.798	*S*
ERCC5	NP_000114	645	MEIDSEESE	0.946	*S*
ERCC5	NP_000114	648	DSEESDSDG	0.997	*S*
ERCC5	NP_000114	650	EEESDGSF	0.906	*S*
ERCC5	NP_000114	653	ESDGSFIEV	0.841	*S*
ERCC5	NP_000114	659	IEVQSVISD	0.638	*S*
ERCC5	NP_000114	662	QSVISDEEL	0.99	*S*
ERCC5	NP_000114	678	SKPPSEQGE	0.993	*S*
ERCC5	NP_000114	705	LRDNSERDD	0.997	*S*
ERCC5	NP_000114	724	DAEDSLHEW	0.911	*S*
ERCC5	NP_000114	749	AQQNSLKAQ	0.708	*S*
ERCC5	NP_000114	804	TDQTSGTIT	0.731	*S*
ERCC5	NP_000114	811	ITDDSDIWL	0.951	*S*
ERCC5	NP_000114	944	VVDDSKGSF	0.777	*S*
ERCC5	NP_000114	947	DSKGSFLWG	0.647	*S*
ERCC5	NP_000114	1012	KRIKSQRLN	0.99	*S*
ERCC5	NP_000114	1032	EAAASEIEA	0.994	*S*
ERCC5	NP_000114	1038	IEAVSVAME	0.66	*S*
ERCC5	NP_000114	1069	EESSSLKRK	0.99	*S*
ERCC5	NP_000114	1076	RKRLSDSKR	0.995	*S*
ERCC5	NP_000114	1078	RLSDSKRKN	0.988	*S*
ERCC5	NP_000114	1096	CLSESSDGS	0.823	*S*
ERCC5	NP_000114	1100	SSDGSSSED	0.993	*S*
ERCC5	NP_000114	1101	SDGSSSEDA	0.998	*S*
ERCC5	NP_000114	1102	DGSSSEDAE	0.935	*S*
ERCC5	NP_000114	1125	EPKTSASDS	0.995	*S*
ERCC5	NP_000114	1127	KTSASDSQN	0.894	*S*
ERCC5	NP_000114	1129	SASDSQNSV	0.878	*S*
ERCC5	NP_000114	1132	DSQNSVKEA	0.998	*S*
ERCC5	NP_000114	1147	ATTSSSSDS	0.994	*S*
ERCC5	NP_000114	1148	TTSSSSDSD	0.929	*S*
ERCC5	NP_000114	1149	TSSSSDSD	0.998	*S*
ERCC5	NP_000114	1151	SSSDSDDDG	0.997	*S*
ERCC5	NP_000114	105	DSRKTTEKL	0.937	*T*
ERCC5	NP_000114	106	SRKTTEKLL	0.973	*T*
ERCC5	NP_000114	212	MKEFTKRRR	0.836	*T*
ERCC5	NP_000114	217	KRRRTLFEA	0.99	*T*
ERCC5	NP_000114	299	IQAKTVAEV	0.602	*T*
ERCC5	NP_000114	331	EKLKTEKEP	0.922	*T*
ERCC5	NP_000114	338	EPDATPPSP	0.658	*T*

ERCC5	NP_000114	411	DDVQTGGPG	0.51	*T*
ERCC5	NP_000114	468	GREPTDSVP	0.684	*T*
ERCC5	NP_000114	523	GRELTPASP	0.97	*T*
ERCC5	NP_000114	567	SDETKCKP	0.535	*T*
ERCC5	NP_000114	623	QEQQTTESA	0.655	*T*
ERCC5	NP_000114	673	EFPETSKPP	0.953	*T*
ERCC5	NP_000114	912	NPHDTKVKK	0.95	*T*
ERCC5	NP_000114	973	NRTKTDESL	0.62	*T*
ERCC5	NP_000114	1055	AKRKTQKRG	0.927	*T*
ERCC5	NP_000114	1063	GITNTLEES	0.706	*T*
ERCC5	NP_000114	1083	KRKNTCGGF	0.725	*T*
ERCC5	NP_000114	1145	GGATTSSSS	0.766	*T*
ERCC5	NP_000114	144	ENDLYVLPP	0.988	*Y*
ERCC5	NP_000114	289	DTSHYILIK	0.535	*Y*
ERCC5	NP_000114	783	FGIPYIQAP	0.617	*Y*
ERCC5	NP_000114	835	FVEYYQYVD	0.729	*Y*
ERCC5	NP_000114	837	EYYQYVDFH	0.731	*Y*
POLH	NP_955452	7	LLRRSGKRR	0.995	*S*
POLH	NP_955452	13	KRRRSESGS	0.998	*S*
POLH	NP_955452	15	RRSESGSDS	0.998	*S*
POLH	NP_955452	17	SESGSDSFS	0.964	*S*
POLH	NP_955452	19	SGSDSFGS	0.919	*S*
POLH	NP_955452	21	SDSFGSGG	0.99	*S*
POLH	NP_955452	23	SFSGSGGDS	0.995	*S*
POLH	NP_955452	28	GGDSSASPQ	0.763	*S*
POLH	NP_955452	30	DSSASPQFL	0.921	*S*
POLH	NP_955452	40	GSVLSPPPG	0.798	*S*
POLH	NP_955452	118	SAPTSAGKT	0.89	*S*
POLH	NP_955452	176	MGSTSPSRH	0.994	*S*
POLH	NP_955452	178	STSPSRHFS	0.87	*S*
POLH	NP_955452	182	SRHFSSLDI	0.987	*S*
POLH	NP_955452	244	ITRKSASCQ	0.984	*S*
POLH	NP_955452	246	RKSASCQAD	0.667	*S*
POLH	NP_955452	254	DLASSLSNA	0.907	*S*
POLH	NP_955452	301	KVGNISYDS	0.99	*S*
POLH	NP_955452	375	LVKPSECPP	0.623	*S*
POLH	NP_955452	399	RRLPSGLDS	0.99	*S*
POLH	NP_955452	448	STLSSGVNL	0.922	*S*
POLH	NP_955452	502	NSEKSKGIA	0.915	*S*
POLH	NP_955452	530	EVTGSMIRA	0.623	*S*
POLH	NP_955452	547	VASTSQDMH	0.842	*S*
POLH	NP_955452	562	FLAASMKEG	0.997	*S*
POLH	NP_955452	597	EFIQSTEAS	0.783	*S*
POLH	NP_955452	622	TLSSSLSPA	0.88	*S*
POLH	NP_955452	624	SSSLSPADT	0.996	*S*
POLH	NP_955452	679	KLPTSMKRV	0.991	*S*
POLH	NP_955452	800	LVRVLLNA	0.719	*S*
POLH	NP_955452	813	VLYASGFHT	0.628	*S*
POLH	NP_955452	841	VPFKSARKA	0.669	*S*
POLH	NP_955452	914	SLTHSESEV	0.942	*S*
POLH	NP_955452	929	SQTKSSYKK	0.63	*S*
POLH	NP_955452	930	QTKSSYKKL	0.994	*S*
POLH	NP_955452	940	SKNKSNTIF	0.597	*S*

POLH	NP_955452	947	IFSDSYIKH	0.967	*S*
POLH	NP_955452	962	DLNKSREHT	0.959	*S*
POLH	NP_955452	993	RKRASLDIN	0.971	*S*
POLH	NP_955452	1004	KPGASQNEG	0.985	*S*
POLH	NP_955452	1011	EGKTSDDKV	0.931	*S*
POLH	NP_955452	1032	LNFNSEKMS	0.791	*S*
POLH	NP_955452	1036	SEKMSRSFR	0.586	*S*
POLH	NP_955452	1038	KMSRSFRSW	0.826	*S*
POLH	NP_955452	1041	RSFRSWKRR	0.995	*S*
POLH	NP_955452	1051	HLKRSRDSS	0.957	*S*
POLH	NP_955452	1054	RSRDSSPLK	0.986	*S*
POLH	NP_955452	1055	SRDSSPLKD	0.996	*S*
POLH	NP_955452	1076	LSNPSLCD	0.991	*S*
POLH	NP_955452	1094	EFRNSGPFA	0.99	*S*
POLH	NP_955452	1102	AKNVSLSGK	0.556	*S*
POLH	NP_955452	1104	NVSLSGKEK	0.998	*S*
POLH	NP_955452	1152	CQATSVVSE	0.929	*S*
POLH	NP_955452	1155	TSVVSEKGR	0.971	*S*
POLH	NP_955452	1200	LRKQSHEQT	0.994	*S*
POLH	NP_955452	1276	AGAFSKSEG	0.987	*S*
POLH	NP_955452	1278	AFSKSEGQH	0.943	*S*
POLH	NP_955452	1402	MKQSSDSHG	0.986	*S*
POLH	NP_955452	1404	QSSDSHGVD	0.984	*S*
POLH	NP_955452	1414	LTPESPIFH	0.979	*S*
POLH	NP_955452	1436	KNEVSVTDS	0.993	*S*
POLH	NP_955452	1486	SLNMSDSSL	0.678	*S*
POLH	NP_955452	1493	LLFDSFSDD	0.927	*S*
POLH	NP_955452	1519	SEVTSNHFS	0.831	*S*
POLH	NP_955452	1559	SIIFSEMDS	0.973	*S*
POLH	NP_955452	1587	HTVVSPRAL	0.967	*S*
POLH	NP_955452	1628	RQNHSFIWS	0.878	*S*
POLH	NP_955452	1639	SFDLSPGLQ	0.962	*S*
POLH	NP_955452	1651	DKVSSPLEN	0.992	*S*
POLH	NP_955452	1683	QEVISNLET	0.936	*S*
POLH	NP_955452	1703	NEVKSКИEM	0.971	*S*
POLH	NP_955452	1726	PRKESNIVD	0.75	*S*
POLH	NP_955452	1743	PIPTSASKL	0.767	*S*
POLH	NP_955452	1776	YLFGPSDI	0.994	*S*
POLH	NP_955452	1786	NHDLSPGSR	0.924	*S*
POLH	NP_955452	1789	LSPGSRNGF	0.516	*S*
POLH	NP_955452	1797	FKDNSPISD	0.959	*S*
POLH	NP_955452	1800	NSPISDTSF	0.945	*S*
POLH	NP_955452	1820	TPASSSSES	0.982	*S*
POLH	NP_955452	1821	PASSSSES	0.985	*S*
POLH	NP_955452	1822	ASSSSESLS	0.682	*S*
POLH	NP_955452	1824	SSSESLSII	0.944	*S*
POLH	NP_955452	1826	SESLSIIDV	0.991	*S*
POLH	NP_955452	1851	KKRFSISLA	0.952	*S*
POLH	NP_955452	1864	RSLTSSKTA	0.983	*S*
POLH	NP_955452	1865	SLTSSKTAT	0.952	*S*
POLH	NP_955452	1879	KQASSPQEI	0.995	*S*
POLH	NP_955452	1924	EQKHSEISA	0.981	*S*
POLH	NP_955452	1956	LRKESDKEC	0.995	*S*

POLH	NP_955452	1985	SLEQSYEDP	0.998	*S*
POLH	NP_955452	2000	LDPDSQEPT	0.979	*S*
POLH	NP_955452	2038	LNAGSEHSG	0.66	*S*
POLH	NP_955452	2041	GSEHSGRYR	0.531	*S*
POLH	NP_955452	2047	RYRASVESI	0.983	*S*
POLH	NP_955452	2080	VEMPSQYCL	0.735	*S*
POLH	NP_955452	2101	AECESQKHI	0.896	*S*
POLH	NP_955452	2126	GHSFSFTSS	0.636	*S*
POLH	NP_955452	2129	FSFTSSDDI	0.944	*S*
POLH	NP_955452	2154	KNQGSKCTL	0.855	*S*
POLH	NP_955452	2160	KTLGSTRRG	0.983	*S*
POLH	NP_955452	2178	GRQFSTSKD	0.921	*S*
POLH	NP_955452	2233	IYPVQSHT	0.635	*S*
POLH	NP_955452	2267	LVGESPPSQ	0.966	*S*
POLH	NP_955452	2270	ESPPSQAVG	0.908	*S*
POLH	NP_955452	2313	PFSISMRHA	0.803	*S*
POLH	NP_955452	2373	IEPESVGDD	0.996	*S*
POLH	NP_955452	2397	MGAKSLGEQ	0.986	*S*
POLH	NP_955452	2418	DSFKSRYTG	0.803	*S*
POLH	NP_955452	2476	IVQGSAADI	0.966	*S*
POLH	NP_955452	2517	QTGLSRKRK	0.99	*S*
POLH	NP_955452	2569	AVKLSVKLK	0.888	*S*
POLH	NP_955452	2581	KIGASWGEL	0.978	*S*
POLH	NP_955452	190	IAVCTIERA	0.502	*T*
POLH	NP_955452	421	HAGLTFEER	0.653	*T*
POLH	NP_955452	488	KGVDTVGES	0.938	*T*
POLH	NP_955452	628	SPADTLDIF	0.843	*T*
POLH	NP_955452	721	KRFFTSVLV	0.53	*T*
POLH	NP_955452	871	RKGLTEREA	0.988	*T*
POLH	NP_955452	935	YKKLTSKNK	0.818	*T*
POLH	NP_955452	1010	NEGKTSDDK	0.715	*T*
POLH	NP_955452	1023	FSQKTKKAP	0.774	*T*
POLH	NP_955452	1204	SHEQTSTIT	0.816	*T*
POLH	NP_955452	1326	FYLDTQSEK	0.6	*T*
POLH	NP_955452	1397	KTVGTMKQS	0.726	*T*
POLH	NP_955452	1411	VDILTPESP	0.634	*T*
POLH	NP_955452	1454	QTQETVKPV	0.917	*T*
POLH	NP_955452	1469	KRTPTGVEG	0.918	*T*
POLH	NP_955452	1673	NRKNTELNE	0.609	*T*
POLH	NP_955452	1863	IRSLTSSKT	0.929	*T*
POLH	NP_955452	1940	DPSLTLKDR	0.959	*T*
POLH	NP_955452	2179	RQFSTSKDV	0.97	*T*
POLH	NP_955452	2204	WRITNAIT	0.806	*T*
POLH	NP_955452	2239	SHTATGRIT	0.773	*T*
POLH	NP_955452	2243	TGRITFTEP	0.563	*T*
POLH	NP_955452	2430	FMTETVKNC	0.624	*T*
POLH	NP_955452	64	TVPDYERDK	0.803	*Y*
POLH	NP_955452	84	VLEKYHSFG	0.765	*Y*
POLH	NP_955452	156	EKKYYLQSL	0.686	*Y*
POLH	NP_955452	171	KVDGYMGST	0.69	*Y*
POLH	NP_955452	303	GNSIYDSSM	0.638	*Y*
POLH	NP_955452	553	DMHTYAACT	0.541	*Y*
POLH	NP_955452	741	INQKYGCNR	0.555	*Y*

POLH	NP_955452	931	TKSSYKKL	0.753	*Y*
POLH	NP_955452	948	FSDSYIKHS	0.937	*Y*
POLH	NP_955452	1226	AVSSYINRD	0.579	*Y*
POLH	NP_955452	1298	KTGTYYTNNK	0.724	*Y*
POLH	NP_955452	1323	EDSFYLDLDTQ	0.941	*Y*
POLH	NP_955452	1498	FSDDYLVKE	0.933	*Y*
POLH	NP_955452	1914	RDAYYFSLQ	0.51	*Y*
POLH	NP_955452	1986	LEQSYEDPK	0.97	*Y*
POLH	NP_955452	2420	FKSRYTGIN	0.858	*Y*
POLH	NP_955452	2544	DELLYEVAE	0.88	*Y*
XPA	NP_000371	23	ELPASVRA	S 0.965	*S*
XPA	NP_000371	27	SVRASIER	K 0.997	*S*
XPA	NP_000371	49	ARPYATA	A 0.900	*S*
XPA	NP_000371	232	RAVRSSVW	K 0.679	*S*
XPA	NP_000371	233	AVRSSVWK	R 0.968	*S*
XPA	NP_000371	239	WKRETIVH	Q 0.700	*T*
XPA	NP_000371	102	MEFDYVIC	E 0.952	*Y*
XPA	NP_000371	116	FMDSYLMN	H 0.956	*Y*
XPA	NP_000371	257	EDDMYRKT	C 0.873	*Y*
XPA	NP_000371	270	HELTYEKM	-0.922	*Y*
XPC	NP_001139241	18	RELRSQKSK	0.974	*S*
XPC	NP_001139241	21	RSQKSKAKS	0.825	*S*
XPC	NP_001139241	25	SKAKSKARR	0.99	*S*
XPC	NP_001139241	49	KSLLSKVSQ	0.909	*S*
XPC	NP_001139241	61	KRGCSPGG	0.975	*S*
XPC	NP_001139241	94	DEALSDGDD	0.996	*S*
XPC	NP_001139241	122	MNEDSNEEE	0.996	*S*
XPC	NP_001139241	129	EEEESENDW	0.992	*S*
XPC	NP_001139241	142	TRERSEKIK	0.992	*S*
XPC	NP_001139241	188	NNICSQPDL	0.567	*S*
XPC	NP_001139241	235	NAELSASEQ	0.986	*S*
XPC	NP_001139241	237	ELSASEQDN	0.977	*S*
XPC	NP_001139241	254	FAIYSARDD	0.987	*S*
XPC	NP_001139241	298	GKKPSKERL	0.994	*S*
XPC	NP_001139241	309	DPGGSSETS	0.833	*S*
XPC	NP_001139241	310	PGGSSETSS	0.569	*S*
XPC	NP_001139241	313	SSETSSQVL	0.596	*S*
XPC	NP_001139241	326	KPKTSKGTK	0.845	*S*
XPC	NP_001139241	343	TCRPSAKGK	0.996	*S*
XPC	NP_001139241	357	RKKRSKPSS	0.955	*S*
XPC	NP_001139241	360	RSKPSSSEE	0.997	*S*
XPC	NP_001139241	361	SKPSSSEED	0.998	*S*
XPC	NP_001139241	362	KPSSSEEDE	0.998	*S*
XPC	NP_001139241	389	RRVASRVSY	0.957	*S*
XPC	NP_001139241	392	ASRVSYKEE	0.998	*S*
XPC	NP_001139241	397	YKEESGSDE	0.996	*S*
XPC	NP_001139241	399	EESGSDEAG	0.992	*S*
XPC	NP_001139241	404	DEAGSGSDF	0.996	*S*
XPC	NP_001139241	406	AGSGSDFEL	0.617	*S*
XPC	NP_001139241	411	DFELSSGEA	0.978	*S*
XPC	NP_001139241	412	FELSSGEAS	0.979	*S*
XPC	NP_001139241	416	SGEASDPSD	0.842	*S*
XPC	NP_001139241	419	ASDPSDEDS	0.998	*S*

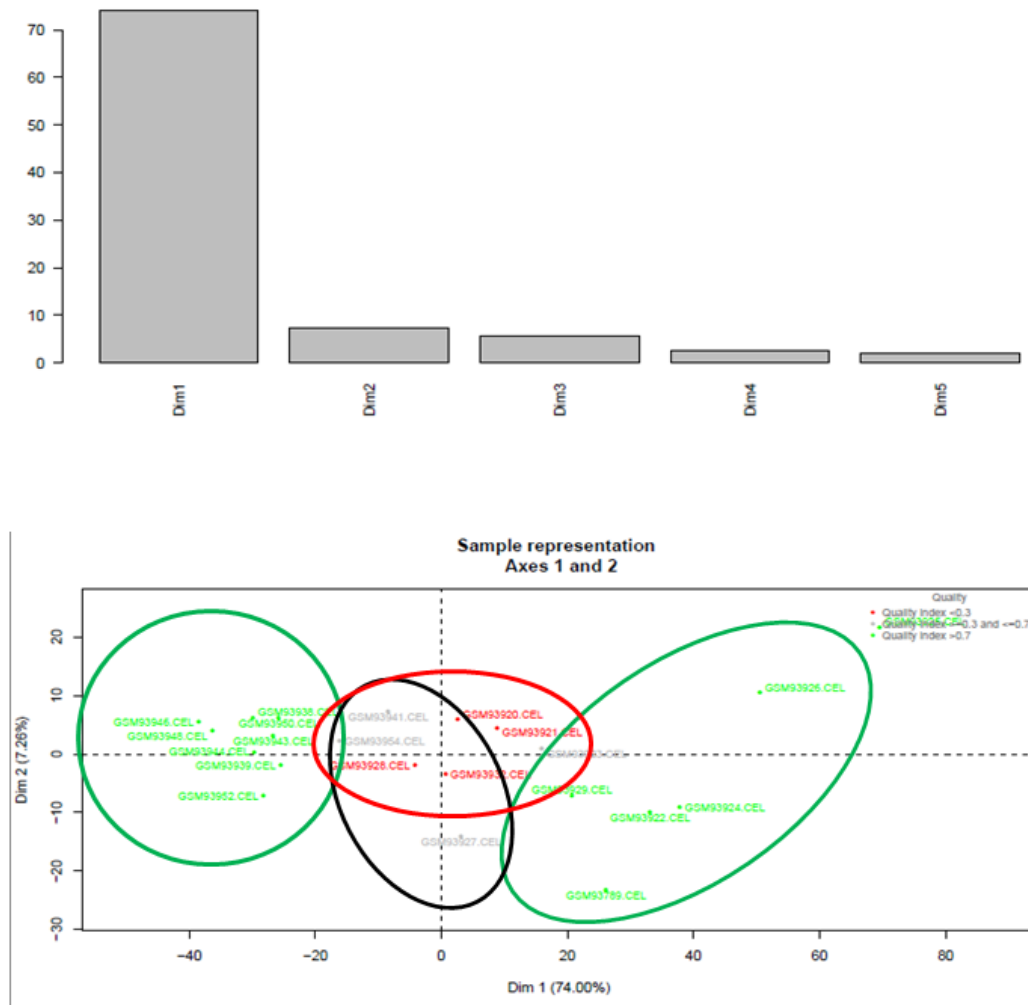


XPC	NP_001139241	423	SDEDSEPGP	0.913	*S*
XPC	NP_001139241	443	TKAGSKSAS	0.994	*S*
XPC	NP_001139241	445	AGSKSASRT	0.985	*S*
XPC	NP_001139241	447	SKSASRTHR	0.64	*S*
XPC	NP_001139241	453	THRGSHRKD	0.997	*S*
XPC	NP_001139241	464	LPAASSSSSS	0.873	*S*
XPC	NP_001139241	466	AASSSSSSSS	0.874	*S*
XPC	NP_001139241	467	ASSSSSSSK	0.912	*S*
XPC	NP_001139241	468	SSSSSSSKR	0.843	*S*
XPC	NP_001139241	469	SSSSSSKRG	0.998	*S*
XPC	NP_001139241	470	SSSSSKRGK	0.995	*S*
XPC	NP_001139241	478	KKMCSDGEK	0.866	*S*
XPC	NP_001139241	487	AEKRSIAGI	0.749	*S*
XPC	NP_001139241	575	RPYQSPFMD	0.544	*S*
XPC	NP_001139241	674	VKGFSNRAR	0.832	*S*
XPC	NP_001139241	832	YGPKEAAA	0.978	*S*
XPC	NP_001139241	846	GGGLSSDEE	0.948	*S*
XPC	NP_001139241	847	GGLSSDEEE	0.987	*S*
XPC	NP_001139241	855	EGTSSQAEA	0.917	*S*
XPC	NP_001139241	866	ILAASWPQN	0.918	*S*
XPC	NP_001139241	895	KAAASHLFP	0.66	*S*
XPC	NP_001139241	79	VAKVTVKSE	0.709	*T*
XPC	NP_001139241	117	KRGATMNED	0.747	*T*
XPC	NP_001139241	168	VHEDTHKVH	0.933	*T*
XPC	NP_001139241	205	PARFTRVLP	0.76	*T*
XPC	NP_001139241	303	KERLTADPG	0.872	*T*
XPC	NP_001139241	325	TKPKTSKGT	0.911	*T*
XPC	NP_001139241	339	FAKGTCRPS	0.821	*T*
XPC	NP_001139241	376	QEKATQRRP	0.696	*T*
XPC	NP_001139241	449	SASRTHRGS	0.751	*T*
XPC	NP_001139241	545	VRDVTQRYD	0.861	*T*
XPC	NP_001139241	556	WMTVTRKCR	0.622	*T*
XPC	NP_001139241	569	WWAETLRPY	0.949	*T*
XPC	NP_001139241	652	HSRDTWLKK	0.861	*T*
XPC	NP_001139241	853	EEEGTSSQA	0.812	*T*
XPC	NP_001139241	886	GPKKTKREK	0.983	*T*
XPC	NP_001139241	152	EFETYLRRA	0.818	*Y*
XPC	NP_001139241	216	VDYYLSNL	0.821	*Y*
XPC	NP_001139241	524	TCYKYATKP	0.703	*Y*
XPC	NP_001139241	604	AIGLYKNHP	0.542	*Y*
XPC	NP_001139241	623	YEAIPETA	0.901	*Y*
XPC	NP_001139241	667	GEVPYKMKV	0.893	*Y*

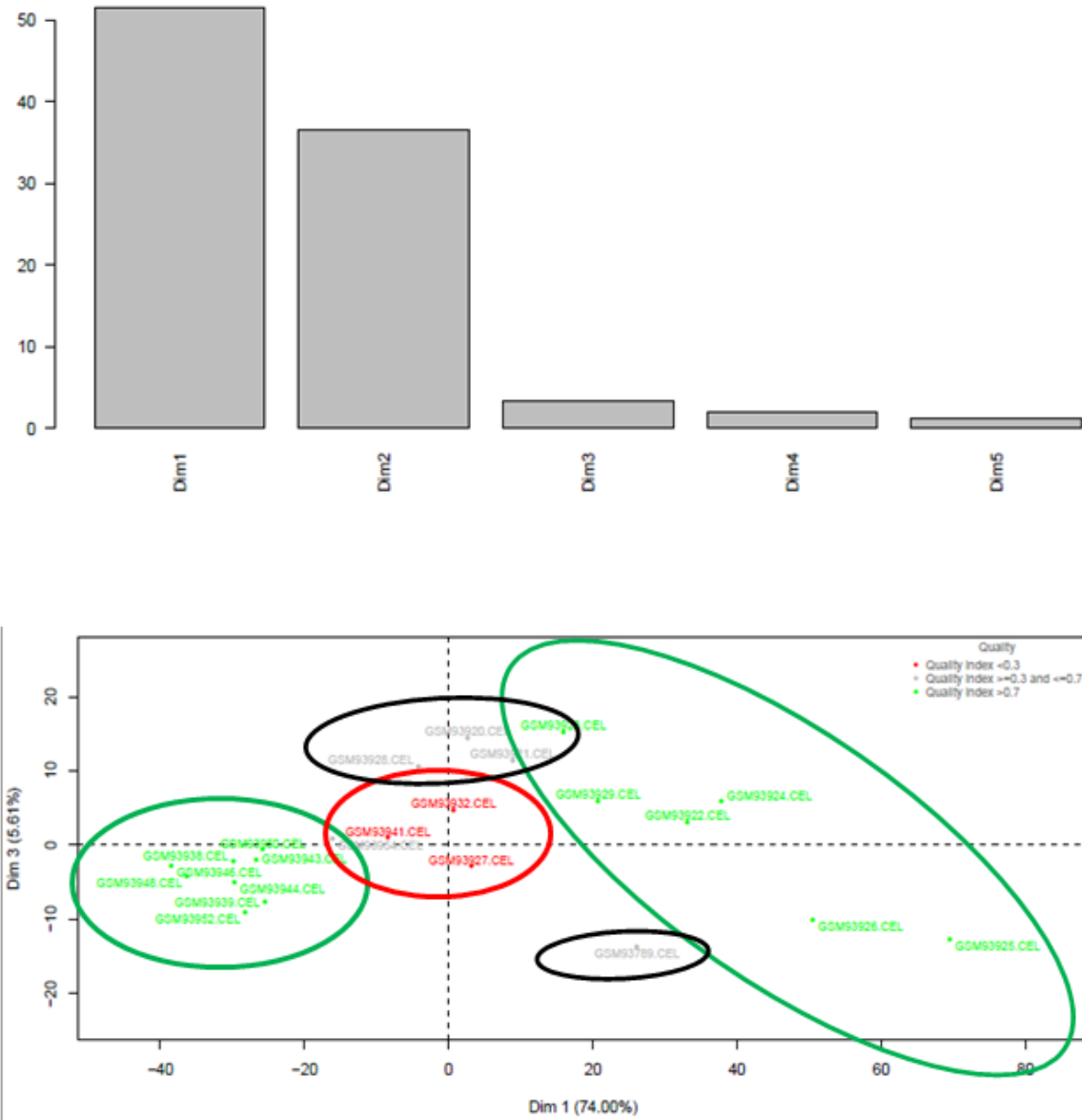
**3. The cycles of energy minimization carried out for each mutation in the XP proteins**

Genes	Mutations	Energy Minimization Method	
		Steepest Descent	Conjugate Gradient
DDB2	K244E	30	330
	A293T	30	150
	D307Y	20	230
POLH	M14V	30	60
	G209V	30	60
	K231N	30	60
	C321W	30	60
XPA	C108F	30	150

## Appendix B



**Figure 1.** The component level analysis performed by PCA experiments.



**Figure 2.** The components obtained from PCA analysis.

## 3. A table listing all the identified TFs in early CRC progression with their frequency of occurrence

S. No	Transcription Factor	Frequency	Importance
1	HNF4	31.80%	0.32
2	NR2F1	19.43%	0.50
3	DR1	17.31%	0.04
4	PPARG	14.49%	0.04
5	HNF1	14.13%	0.36
6	HNF4_DR1	13.78%	0.17
7	PPAR_DR1	13.43%	0.13
8	HNF4ALPHA	12.01%	0.30
9	PAX4	12.01%	0.18
10	ER	10.60%	0.08
11	COUPTF	9.19%	0.01
12	TAL1	8.48%	0.12
13	GATA1	8.48%	0.02
14	LEF1TCF1	8.13%	0.04
15	NFY	8.13%	0.07
16	AREB6	7.77%	0.02
17	SF1	7.77%	0.09
18	CEBPB	7.77%	0.08
19	STAT5A	7.42%	0.08
20	TCF4	7.42%	0.02
21	NKX25	7.42%	0.06
22	PPARA	7.42%	0.00
23	T3R	7.07%	0.06
24	AR	7.07%	0.04
25	STAT	7.07%	0.04
26	LMO2COM	6.71%	0.02
27	GATA3	6.71%	0.03
28	TAL1BETAITF2	6.71%	0.04
29	IPF1	6.71%	0.07
30	TAL1BETA47	6.36%	0.03
31	FXR_IR1	6.01%	0.10
32	NF1	6.01%	0.04
33	SRF	5.65%	0.03
34	RORA1	5.65%	0.05
35	GATA	5.65%	0.02
36	HNF3ALPHA	5.30%	0.07
37	IRF1	5.30%	0.06
38	NANOG	5.30%	0.07
39	STAT1	5.30%	0.01
40	TAL1ALPHA47	4.95%	0.04
41	HNF3	4.59%	0.01
42	MEF3	4.59%	0.06
43	CEBPDELTA	4.24%	0.06
44	EGR	4.24%	0.07
45	MEIS1	4.24%	0.01
46	ZIC3	4.24%	0.06
47	PR	4.24%	0.04
48	CACCCBINDINGFACTOR_Q	3.89%	0.01
49	ATF3	3.89%	0.03
50	E2A	3.89%	0.01
51	GEN_INI3	3.89%	0.04
52	DEAF1	3.89%	0.02
53	GC	3.89%	0.00

54	HAND1E47	3.89%	0.07
55	S8	3.53%	0.04
56	MYOGENIN	3.53%	0.00
57	CHX10	3.53%	0.03
58	FXR	3.53%	0.01
59	NFKB	3.53%	0.00
60	CDPCR3HD	3.53%	0.02
61	PBX1	3.18%	0.01
62	OLF1	3.18%	0.04
63	ZTA	3.18%	0.03
64	ZIC2	3.18%	0.06
65	ATF6	3.18%	0.01
66	GATA6	3.18%	0.06
67	HSF1	3.18%	0.02
68	GEN_INI2	3.18%	0.01
69	IK2	3.18%	0.01
70	STAT5B	3.18%	0.01
71	KAISO	3.18%	0.05
72	SPZ1	2.83%	0.01
73	TTF1	2.83%	0.02
74	BLIMP1	2.83%	0.01
75	CMAF	2.83%	0.03
76	TFIIA	2.83%	0.03
77	PAX6	2.47%	0.02
78	CDPCR1	2.47%	0.01
79	CMYB	2.47%	0.00
80	DEC	2.47%	0.04
81	RSRFC4	2.47%	0.01
82	PU1	2.47%	0.03
83	CIZ	2.47%	0.03
84	STRA13	2.12%	0.00
85	E2F1DP1	2.12%	0.02
86	PEA3	2.12%	0.01
87	NFKAPPAB	2.12%	0.02
88	MOVOB	2.12%	0.01
89	RP58	2.12%	0.03
90	MAZR	2.12%	0.00
91	FOXD3	2.12%	0.02
92	HFH8	1.77%	0.02
93	BRCA	1.77%	0.01
94	TCF11MAFG	1.77%	0.01
95	DR3	1.41%	0.01
96	IK1	1.41%	0.01
97	TFE	1.06%	0.01
98	PBX	1.06%	0.00
99	NGFIC	1.06%	0.00
100	MRF2	1.06%	0.00
101	E2F1DP1RB	1.06%	0.00
102	ROAZ	0.71%	0.00
103	ATF4	0.71%	0.00
104	NCX	0.71%	0.00
105	NMYC	0.35%	0.00
106	MINI19	0.35%	0.00
107	MYC	0.35%	0.00
108	LDSPOLYA	0.35%	0.00

## 4. The number of TFs in a particular gene with their relative positions

Gene	Locus	Type	No. of TFBS	Candidate transcription factor binding sites (Relative positions)
CPT2	chr1:53380877-53452455	intergenic	3	VJUN(66) CREBP1CJUN(66) SREBP1(135)
JUN	chr1:58938345-59535239	promoter	2	ZNF219(136) ZNF219(143)
HOOK1	chr1:60001596-60131561	promoter	6	TATA(100) MYOGENIN(124) LBP1(125) E2A(128) CDP(153) NKX25(174)
CTH--PTGER3	chr1:70614848-71089046--chr1:70677843-71301448	intergenic	10	IPF1(28) TCF11(30) TITF1(35) NKX25(36) NKX25(270) HOXA3(331) HMGYI(406) NKX61(426) IRF1(471) OTX(523)
LPHN2	chr1:81039512-83230281	promoter	6	MYOGNF1(61) MEIS1BHOXA9(129) IPF1(132) SOX5(140) TEF1(147) SOX5(203)
PRKACB	chr1:84237459-84536603	promoter	2	CREB(21) SOX5(99)
TGFBR3	chr1:91764540-92185938	intron	19	RUSH1A(33) RP58(80) HLF(155) VBP(155) HOXA3(159) HOXA4(159) SRY(161) SOX5(161) FOXO4(162) VJUN(174) PAX2(180) TCF11(181) FREAC2(231) STAT3(446) OCT4(467) XPF1(467) PAX2(633) STAT1(781) HNF3B(786)
TMED5	chr1:93375721-93420209	UTR3	7	TATA(93) MMEF2(95) IK1(116) YY1(155) HLF(220) VBP(220) VBP(239)
TMED5	chr1:93375721-93420209	intron	6	DTYPEPA(50) TITF1(196) CRX(301) SOX(309) SRY(309) CREB(329)
FNBP1L	chr1:93602459-93799844	promoter	1	E12(27)
SLC35A3	chr1:100163244-100276352	UTR3	3	STAT5A(1) MEIS1BHOXA9(234) TATA(286)
SLC35A3	chr1:100163244-100276352	intergenic	13	TST1(158) TEF(231) TATA(521) TCF4(535) TEF1(540) TBP(546) SRF(597) NKX61(626) CDP(627) CDP(644) STAT5A(685) STAT6(685) TATA(744)
BCL2L15	chr1:114215883-114239134	intergenic	14	CBF(121) PEBP(121) PXR(129) VJUN(132) CREB(132) HMGYI(154) HOXA4(155) CBF(177) PEBP(177) PADS(179) STAT6(222) PPARG(247) CBF(295) PADS(297)
DCLRE1B	chr1:114248065-114273485	UTR3	14	HOXA4(93) SOX9_B1(95) ZNF219(126) HFH3(169) HNF3(170) HNF3ALPHA(171) SRY(171) E2A(201) MYOGENIN(203) LBP1(204) SREBP1(256) CREB(261) NF1(264) MYOGNF1(269)
OLFML3	chr1:114322004-114433346	promoter	2	SRF(41) STAT5A(57)
PEA15	chr1:158438311-158452124	promoter	2	HSF1(35) SREBP1(70)
PRRX1	chr1:168789233-169171244	promoter	7	TEF(62) SRY(99) OTX(130) TEF1(198) HOXA4(205) S8(206) CHX10(206)
NPHS2	chr1:177786299-177826775	promoter	1	HAND1E47(4)
FAM129A	chr1:182990362-183281169	intron	23	TCF4(177) CDP(192) STAT5A(206) OCT4(208) HNF3ALPHA(211) OTX(280) OCT4(295) MTATA(317) NKX61(331) S8(332) CDP(336) IK2(415) IK1(416) IK3(416) STAT1(417) STAT3(417) NFKAPPAB(418) IK1(421) IK3(421) IK2(421) RBPIK(422) HNF3(541) HAND1E47(719)
PTGS2	chr1:184696869-185064477	intergenic	15	STAT1(236) STAT4(236) STAT6(236) NFKAPPAB65(265) NFKAPPAB(265) STAT1(267) NKX25(294) VJUN(331) CREB(331) CREBP1CJUN(331) HMGYI(415) HAND1E47(439) GLI(484) TST1(546) PEBP(643)
PRELP	chr1:201586915-201729807	intergenic	9	ZNF219(5) NF1(155) E12(174) HAND1E47(304) LBP1(347) MYOGENIN(347) HNF3B(355) TAL1BETAE47(476) TAL1BETAITF2(476)
PRELP	chr1:201586915	promoter	1	PAX3(90)

	-201729807			
NFASC	chr1:202921303 -203275004	intergenic	19	HOXA3(112) VJUN(137) CHX10(142) HOXA3(150) SOX5(242) SMAD4(316) RP58(338) GEN_INI(396) HNF4_DR1(431) PPAR_DR1(431) HNF4_DR1(438) PPAR_DR1(438) PPARG(440) SMAD(501) SMAD4(505) SREBP1(591) LBP1(603) MYOGENIN(603) E2A(605)
PFKFB2	chr1:205291082 -205327928	UTR3	7	IK2(52) STAT3(54) STAT4(83) SRF(182) HAND1E47(280) HNF3(294) PAX2(351)
FLVCR1	chr1:211087657 -211190588	intron	5	NCX(2) IRF1(5) SRY(106) TEF(185) RBPJK(284)
PTPN14	chr1:212577132 -212843138	intron	9	FOXO4(17) PPARG(89) RUSH1A(91) SRF(131) PAX2(176) HMGYIY(240) FREAC2(243) SRY(246) HFH4(248)
DEGS1	chr1:222413880 -222480420	intergenic	1	SRF(768)
ENAH	chr1:223683294 -224032106	promoter	2	YY1(64) NKX61(95)
EGLN1	chr1:229557004 -229729523	promoter	9	HNF3(92) FREAC2(92) STAT1(118) STAT3(118) YY1(146) E12(200) E2A(201) MYOGENIN(201) HAND1E47(231)
KIAA1804	chr1:231498090 -231816304	intergenic	7	P53(30) STAT1(79) STAT3(79) IK3(81) S8(136) CHX10(136) NF1(187)
ACTN2	chr1:234834454 -235023153	intron	3	CREB(115) VBP(115) NCX(307)
GREM2	chr1:238705408 -239005439	promoter	5	PPARG(30) PPAR_DR1(30) STAT1(91) HMGYIY(92) NFKAPPAB65(93)
AKRIC3	chr10:5056239- 5186370	intergenic	5	NKX61(101) TCF4(143) YY1(181) YY1(188) YY1(251)
VIM	chr10:17283728 -17402575	promoter	1	GCNF(11)
VIM	chr10:17283728 -17402575	intergenic	7	HLF(259) VBP(259) TATA(290) CDP(299) OTX(357) PXR(443) TCF4(510)
PBLD	chr10:69661887 -69763564	promoter	1	SRF(117)
LGII	chr10:95452328 -95643067	promoter	1	MYOGNF1(102)
ALDH18A1	chr10:97311181 -97412635	intergenic	6	SMAD4(35) YY1(77) CBF(320) PAX2(332) TBP(359) NKX25(392)
FRAT2	chr10:99071681 -99106252	intergenic	11	IPF1(86) PAX2(105) STAT1(115) STAT3(115) STAT5A(115) PPARG(137) SRF(147) GFII(149) TAL1BETA47(358) TAL1BETAITF2(358) RP58(359)
HTRA1	chr10:12420772 7-124310166	promoter	3	NKX25(44) MYOGNF1(67) NF1(72)
IGF2, INS, INS-IGF2	chr11:1935974- 2141603	promoter	6	SRF(57) TAL1BETA47(82) TAL1BETAITF2(82) STAT1(118) CREB(157) IPF1(188)
CD81	chr11:2296008- 2378536	UTR3	3	GLI(48) TCF11(202) P53(224)
DCHS1	chr11:6597277- 6659248	promoter	1	ZTA(7)
LYVE1	chr11:10519351 -10549874	promoter	3	CBF(181) PEBP(182) STAT5A(202)
EHF	chr11:34491911 -34860381	promoter	3	DTYPEPA(96) NFKAPPAB65(190) NFKAPPAB(190)
MS4A8B	chr11:60066975 -60280611	intergenic	4	STAT5A(159) P53(252) P53_DECAMER(252) NRF1(253)
C11orf54	chr11:93113247 -93151633	intron	2	NF1(72) IRF1(193)
NCAM1	chr11:11161052 2-112690460	promoter	6	NKX61(26) SRY(28) SOX5(28) FOXO4(29) FREAC2(31) HFH8(31)
ZBTB16	chr11:11336764 4-113670215	promoter	3	MTATA(21) ZNF219(149) ZNF219(156)



NNMT	chr11:11362662-2-113774926	intergenic	15	CBF(86) NKX61(232) LBP1(322) STAT3(385) TAL1BETAE47(402) OTX(404) DTYPEPA(513) PAX9(843) MYOGNF1(855) S8(856) SRF(857) NKX61(858) NCX(858) NKX25(858) NFKAPPAB65(866)
GNB3	chr12:6818433-6828121	UTR3	4	TEF1(100) IK3(102) SRF(199) MAZR(265)
MFAP5	chr12:8656716-8741781	UTR5	7	SOX9_B1(155) TEF1(219) NFKAPPAB65(238) NFKAPPAB(238) HMGYIY(239) GEN_INI(259) TEF1(261)
MGP	chr12:14887711-14958197	intergenic	8	MEIS1BHOXA9(11) NKX25(148) S8(150) FOXP3(175) TGIF(216) NF1(235) HFH8(368) IPF1(389)
BCAT1	chr12:24629307-25037186	intergenic	17	SMAD(119) SRY(140) HLF(310) STAT1(365) SREBP1(382) IPF1(403) STAT(426) NKX25(496) NKX25(566) HLF(571) VBP(571) TATA(609) STAT6(617) IK1(619) IK3(619) XBP1(637) HAND1E47(779)
TMTC1	chr12:29543996-30673185	intron	9	NKX25(21) HEN1(145) CMAF(171) FREAC2(233) SOX(286) PAX3(293) HOXA4(434) SRY(538) SOX5(538)
SLC38A4	chr12:45052835-45755537	intergenic	15	CREBP1CJUN(70) FOXO4(162) HNF3ALPHA(162) SRY(162) HNF3(163) HFH3(164) HOXA3(165) GF11(186) HMGYIY(266) HNF3(314) DR4(440) NKX25(440) TATA(551) SREBP1(570) NKX25(570)
AMIGO2	chr12:45506067-45895910	promoter	4	SRF(59) GEN_INI(75) CDP(121) CBF(239)
TUBA1A	chr12:47811573-47945112	promoter	3	RBPIK(65) STAT5A(165) STAT6(165)
PRPH	chr12:47953524-48003217	promoter	3	E2A(20) E12(22) DR4(79)
TENC1	chr12:51722331-51744657	promoter	2	TEF1(117) GEN_INI(119)
METTL7B	chr12:54317988-54364659	intergenic	3	XPF1(126) YY1(169) TAL1BETAITF2(169)
MSRB3	chr12:63928435-64504502	intergenic	1	VJUN(1468)
MSRB3	chr12:63928435-64504502	promoter	10	PAX2(54) NCX(120) PPARG(175) HNF4_DR1(175) PPAR_DR1(175) PBX1(199) NKX25(312) NKX25(321) RP58(427) XPF1(449)
TMEM19	chr12:70360701-70434917	intron	5	NKX25(6) CRX(15) P53(66) P53_DECAMER(66) SRF(74)
CSRP2	chr12:75772009-75936314	promoter	10	GEN_INI(84) CREB(85) VJUN(87) CREBP1CJUN(87) TEF1(105) IRF1(120) PAX2(185) E2F4DP1(196) SOX9_B1(292) GATA3(330)
DCN	chr12:90029789-91061033	promoter	2	HFH8(65) HNF3ALPHA(67)
GLTP	chr12:10875559-8-108821618	UTR3	8	HMEF2(87) PAX2(95) OTX(99) MMEF2(116) HOXA4(126) SOX9_B1(128) GEN_INI(135) TATA(201)
HSPB8	chr12:11808532-8-118256865	intergenic	21	STAT1(15) E2A(174) IPF1(190) PBX1(198) HAND1E47(245) CBF(247) MMEF2(253) GF11(384) GF11(399) NKX61(419) SOX9_B1(421) SRY(421) FOXO3(423) PXR(468) TEF(469) HFH4(485) TBP(578) HMGYIY(580) NKX25(641) HOXA4(642) NF1(646)
CRYL1	chr13:19704462-20038363	intron	7	TCF11(43) SOX9_B1(62) NFE2(122) HOXA3(362) NKX25(362) HLF(365) VBP(365)
DCLK1	chr13:35145061-35640341	promoter	7	PAX3(39) ZTA(39) NF1(132) NFKAPPAB65(159) NFKAPPAB(159) HSF1(201) HSF2(201)
SPG20	chr13:35687593-35904598	intergenic	3	XBP1(85) P53(85) P53_DECAMER(85)
LHFP	chr13:38522514-39127796	promoter	3	SMAD(70) TAL1BETAE47(107) YY1(107)
GPC6	chr13:92317546-93889303	promoter	4	HOXA4(64) FOXO4(149) FREAC2(151) HFH8(151)
COL4A1	chr13:10923693-4-109758941	intergenic	8	SREBP1(181) DTYPEPA(245) YY1(246) NF1(253) TST1(379) SOX5(423) SOX9_B1(511) SOX9_B1(578)

EFS	chr14:22891943 -22911759	promoter	2	SMAD(37) HAND1E47(39)
ERO1L	chr14:52174661 -52243667	UTR3	2	PAX3(38) CRX(133)
ERO1L	chr14:52174661 -52243667	intergenic	2	TCF11(6) DTYPEPA(101)
PPM1A	chr14:59701895 -59968900	UTR5	3	CREB(2) VJUN(3) CREBP1CJUN(3)
PPM1A	chr14:59701895 -59968900	intron	16	PAX3(136) TBP(240) TEF(357) ZTA(438) HLF(441) SRF(449) GF11(469) SOX(487) TAL1BETAE47(531) NKX61(577) SRF(580) SOX9_B1(651) SOX5(651) TCF4(661) XFD2(740) SRY(752)
PPM1A	chr14:59701895 -59968900	intergenic	23	CHX10(175) S8(177) PPARG(193) HNF4_DR1(193) HMG1Y(200) LBP1(215) MYOGENIN(215) PPARG(235) PPAR_DR1(235) ZTA(242) OTX(251) PITX2(254) CREB(297) VJUN(299) HLF(299) VBP(299) SRF(309) NCX(311) NKX61(312) P53(402) P53_DECAMER(402) HSF1(416) TEF1(485)
RHOJ	chr14:62638539 -62849295	UTR5	22	NFE2(16) PAX9(22) CREB(29) VJUN(30) CREBP1CJUN(30) TGIF(198) LDSPOLYA(230) STAT1(231) XPF1(258) STAT3(260) SOX9_B1(294) STAT6(304) SRF(322) HOXA3(325) YY1(327) STAT1(353) STAT4(353) STAT6(353) YY1(421) NFKAPPAB(545) PAX2(574) PXR(578)
MPP5	chr14:66765040 -66874328	intron	5	TATA(114) PPARG(166) MMEF2(183) NFKAPPAB65(441) NFKAPPAB(441)
PRIMA1	chr14:93243517 -93455001	intergenic	2	HOXA4(516) HOXA3(517)
SERPINA3	chr14:94129205 -94176345	intergenic	4	TITF1(65) TATA(146) MMEF2(148) HMEF2(150)
CLMN	chr14:94693516 -94953582	intergenic	4	YY1(91) TATA(182) P53(559) P53_DECAMER(559)
GREM1	chr15:30776627 -30853617	promoter	1	PAX2(109)
MEIS2	chr15:34961901 -36014578	promoter	3	MAZR(24) NFE2(97) IRF1(158)
CILP	chr15:63264688 -63337458	promoter	2	NKX25(204) DR4(265)
COX5A	chr15:72986871 -73034399	intron	4	GEN_INI(56) VJUN(59) P53(127) OTX(137)
CHRNA3	chr15:76674063 -76703568	UTR3	2	ZTA(97) ZTA(128)
FAHD1	chr16:1815744- 1830173	intron	4	NF1(63) MYOGNF1(68) NKX25(82) MMEF2(138)
C16orf73	chr16:1823990- 1901258	UTR5	6	TAL1BETAE47(115) TAL1BETAITF2(115) RP58(116) SRF(165) TST1(186) NKX25(191)
C16orf73	chr16:1823990- 1901258	intergenic	26	PPARG(18) HNF4_DR1(18) PPAR_DR1(18) HAND1E47(139) SMAD(141) PPARG(151) TAL1BETAE47(157) TAL1BETAITF2(157) E12(157) STAT1(163) IK1(165) IK3(165) IK2(165) CDP(171) XPF1(192) OCT4(193) SRY(197) PEBP(227) NKX61(249) HFH4(285) OTX(411) XFD2(413) PBX1(487) TCF4(547) STAT6(553) PITX2(606)
NDE1	chr16:15644514 -15726490	promoter	1	HSF1(149)
MYH11	chr16:15704495 -15866811	intron	5	LBP1(111) SRF(143) HOXA3(145) MTATA(176) SRF(176)
MYH11	chr16:15704495 -15866811	UTR5	3	NCX(45) TATA(160) MTATA(161)
NETO2	chr16:45565134 -45746539	intron	4	FOXO4(13) FOXO3(14) HSF1(226) SRY(239)
CES1	chr16:54295240 -54437599	intergenic	10	STAT1(70) TAL1BETAITF2(90) SMAD4(96) PAX9(101) SRF(127) SRF(156) YY1(158) TAL1BETAE47(158) IK3(208) NF1(215)
NQO1	chr16:68296909 -68332678	UTR3	4	FREAC2(62) FOXO4(65) STAT1(143) STAT3(143)

NQO1	chr16:68296909 -68332678	intron	5	NF1(53) ZTA(59) YY1(60) NF1(70) NF1(124)
SCO1	chr17:10500349 -10542496	intergenic	15	IPF1(63) IPF1(72) E2A(161) MYOGENIN(163) E12(164) LBP1(165) E2A(167) CBF(180) MYOGENIN(200) HOXA3(203) LBP1(216) MYOGENIN(217) VJUN(297) SRF(310) GFI1(311)
CPD	chr17:25685231 -25828460	intron	7	STAT5A(29) STAT6(29) TCF11(42) ATATA(65) HNF6(251) HLF(443) VBP(443)
CPD	chr17:25685231 -25828460	UTR3	22	E2F4DP1(116) IPF1(178) HNF6(188) S8(200) HOXA4(201) NKX25(202) NF1(398) SRY(482) PBX1(489) SOX9_B1(490) SRY(490) SOX5(490) SRY(574) PITX2(613) CREBP1CJUN(618) NKX25(709) PBX1(733) YY1(783) HLF(808) VBP(808) HLF(842) PAX3(901)
CCL2	chr17:29507941 -29621336	intergenic	5	PBX1(111) PPARG(177) HOXA3(206) HOXA4(206) NKX25(217)
CCL2	chr17:29507941 -29621336	promoter	1	NCX(10)
CCL2	chr17:29507941 -29621336	promoter	1	STAT4(58)
KRT24	chr17:36075009 -36154523	intergenic	5	NFKAPPAB65(140) NFKAPPAB(140) S8(158) NKX25(160) SOX9_B1(194)
COL1A1	chr17:45608311 -45706827	promoter	2	SREBP1(37) IK3(60)
TOM1L1	chr17:49334502 -50394309	promoter	3	NFKAPPAB65(4) NFKAPPAB(4) PADS(95)
COX11	chr17:50384266 -50402538	UTR3	3	GCNF(140) NKX25(142) STAT3(243)
SLC16A5	chr17:70573739 -70617653	UTR5	8	PPARG(61) GLI(102) E2A(179) MYOGENIN(181) HEN1(182) E12(183) SMAD4(203) PEBP(215)
ACOX1	chr17:71449189 -71508873	UTR3	3	TAL1BETAE47(2) TAL1BETAITF2(2) NKX61(101)
RAB31	chr18:9604601- 9874625	promoter	3	VJUN(65) CREB(67) SRF(98)
TUBB6	chr18:12267603 -12319075	intergenic	3	TEF(203) TEF1(254) SREBP1(274)
DSC2	chr18:26876814 -26963189	intergenic	6	SRF(101) TATA(103) RBPJK(135) HMGYIY(159) SMAD4(447) HAND1E47(452)
DSC2	chr18:26876814 -26963189	UTR3	7	MTATA(105) SMAD(166) FOXO4(260) GATA3(304) TGIF(412) IPF1(415) TCF11(417)
MEP1B	chr18:27965612 -28094968	intron	2	TAL1BETAITF2(350) P53(372)
RAB27B	chr18:50417753 -50719214	promoter	2	HLF(14) NKX61(19)
CDH19	chr18:61699197 -63311230	intergenic	16	STAT6(161) PITX2(162) CRX(163) IPF1(165) STAT5A(195) STAT6(195) SRF(198) YY1(200) ATATA(269) OTX(270) HOXA4(284) NKX61(290) SOX9_B1(305) MEIS1BHOXA9(317) CDP(411) OTX(718)
HSPB6	chr19:40937302 -40941111	UTR5	4	TATA(43) MTATA(49) SRF(129) XBP1(151)
AXL	chr19:46405976 -46460228	promoter	1	P53(126)
MYADM	chr19:59019662 -59077277	promoter	1	LDSPOLYA(32)
PEG3, ZIM2	chr19:61876135 -62322739	promoter	4	NCX(81) SRF(122) TEF1(186) HMGYIY(236)
RHOQ	chr2:46600698- 46665329	UTR3	13	NFKAPPAB65(120) NKX25(196) GFI1(214) FOXO4(220) SRY(220) RP58(346) SRF(422) SRF(430) HFH4(470) HNF3(471) HNF3ALPHA(472) NKX61(739) PBX1(744)
ANXA4	chr2:69724606- 69910321	intergenic	5	STAT5A(41) CRX(108) GEN_INI(254) SRF(347) TATA(347)
ANXA4	chr2:69724606- 69910321	UTR3	9	IPF1(79) GATA3(81) STAT5A(97) STAT6(97) NKX61(103) E2A(120) HOXA4(162) SRF(220) OTX(269)

ACTG2	chr2:73943558-74007447	intergenic	9	DR4(35) RBPJK(51) IK2(52) IK1(53) STAT3(54) SRF(136) CREB(160) VJUN(162) CREBPCJUN(162)
ACTG2	chr2:73943558-74007447	intron	3	PBX1(33) XFD2(56) TCF11(137)
RETSAT	chr2:85409873-85436642	UTR3	2	HMGYIY(31) TATA(36)
RETSAT	chr2:85409873-85436642	promoter	1	TITF1(139)
MAL	chr2:94906464-95116673	intergenic	4	NKX25(48) SOX9_B1(78) SOX5(78) VBP(378)
SLC9A2	chr2:102516953-102699677	promoter	9	CBF(33) CHX10(48) NCX(48) HOXA4(49) P53(60) GFII1(68) GFII1(83) STAT3(95) STAT5A(95)
SLC35F5	chr2:114119975-114364003	intergenic	16	IK3(265) NKX25(273) NF1(304) NFKAPPAB65(304) NFKAPPAB(304) IRF1(309) SRF(312) HAND1E47(350) PITX2(355) OTX(357) XBP1(360) TCF4(390) STAT(485) STAT5A(488) STAT5A(495) STAT(501)
GYPC	chr2:126130333-127521990	intergenic	11	IRF1(150) PAX2(172) HOXA4(292) PEBP(302) DTYPEPA(303) HOXA4(307) OTX(328) CDP(331) IPF1(406) NFE2(434) IRF1(489)
PLEKHB2	chr2:131567501-131690843	intergenic	2	ZTA(65) TCF4(293)
HNMT	chr2:138151780-138975819	intergenic	1	IRF1(1137)
SCN7A	chr2:166880386-167467935	promoter	4	HNF6(70) HNF6(150) CDP(153) PBX1(153)
KLHL23, PHOSPHO2	chr2:170258321-170363627	intergenic	6	HSF1(39) HSF2(39) HNF3(273) IRF1(297) SMAD(389) PAX3(422)
SDPR	chr2:192260048-192522838	promoter	3	NKX25(101) HOXA4(102) NKX25(123)
IDH1	chr2:208763034-208839224	UTR3	9	MEIS1BHOXA9(222) RUSH1A(262) FOXO4(292) SRY(292) HMGYIY(295) NKX61(296) NCX(296) HLF(299) TITF1(322)
DES	chr2:219960910-220007922	intergenic	25	IPF1(102) TCF11(104) OTX(104) HNF3B(127) FREAC2(128) HFH8(128) HNF3ALPHA(130) FOXO3(149) FOXO4(149) FREAC2(150) HFH3(150) NFE2(181) STAT3(192) ZTA(233) STAT5A(252) HAND1E47(272) DTYPEPA(283) NKX61(284) DTYPEPA(291) SMAD(303) TATA(320) PPARG(364) HNF4_DR1(364) PBX1(408) MEIS1BHOXA9(432)
SCG2	chr2:223628759-224326950	intergenic	14	OCT4(81) GEN_INI(101) NFE2(131) MEIS1BHOXA9(216) OCT4(218) S8(284) SRF(285) HOXA4(285) HOXA3(286) NKX25(286) TITF1(362) NKX25(363) IPF1(431) S8(431)
UGT1A1, UGT1A10, UGT1A3, UGT1A4, UGT1A5, UGT1A6, UGT1A7, UGT1A8, UGT1A9	chr2:234134757-234407062	intergenic	16	HNF4_DR1(13) HNF4_DR1(188) SRF(228) TAL1BETAE47(251) TAL1BETAITF2(251) IK1(329) RBPJK(422) HFH4(505) PAX2(505) SRY(507) PEBP(623) PADS(625) XPF1(631) GATA3(653) TATA(735) SRF(736)
CXCR7	chr2:237080948-237654833	promoter	6	E2A(25) MYOGENIN(25) HEN1(26) E12(27) SRY(90) HFH4(92)
CXCR7	chr2:237080948-237654833	UTR5	4	ZNF219(115) MEIS1BHOXA9(141) TATA(185) STAT6(265)
ADAM33	chr20:3592139-3615595	intron	5	ZTA(70) TEF1(89) HSF1(91) HSF2(91) TEF1(102)
THBD	chr20:22965452-23007990	UTR3	13	HNF3(62) HNF3ALPHA(64) CREB(82) CREBPCJUN(83) CDP(127) HNF3ALPHA(130) HNF6(130) HFH3(132) HFH8(132) TEF1(244) HSF1(345) HSF2(345) GEN_INI(352)
ADAMTS1	chr21:26867477-27211824	intergenic	12	PAX3(83) S8(292) HMGYIY(301) HOXA4(302) MEIS1BHOXA9(343) SRF(472) STAT(491) HSF1(496) HSF2(496) TEF1(498) IPF1(621) TATA(674)
ADAMTS1	chr21:26867477	UTR5	2	TATA(191) SRF(193)

	-27211824			
ADARB1	chr21:45227226-45507511	intergenic	7	HAND1E47(98) NKX25(117) YY1(124) XPF1(152) PEBP(346) PADS(348) GATA3(366)
TIMP3	chr22:31224886-31998840	UTR3	11	RP58(109) STAT5A(220) NF1(239) PEBP(290) STAT5A(328) STAT6(328) HFH3(365) HFH4(365) HFH8(365) HOXA4(407) TEF1(510)
RBMS3	chr3:28541702-30622977	promoter	3	HAND1E47(28) TAL1BETAE47(49) TAL1BETAITF2(49)
ITGA9	chr3:37452781-37878671	intron	13	SRF(78) GATA3(80) MEIS1BHOXA9(99) TCF4(134) GATA3(307) TCF11(347) PPAR_DR1(473) RBPJK(517) CREB(600) VJUN(601) CREBP1CJUN(601) SOX9_B1(680) SOX9_B1(690)
ABHD5	chr3:43639455-44352123	intergenic	21	YY1(136) NKX25(197) YY1(202) CDP(270) PAX3(294) HOXA3(310) SREBP1(314) YY1(346) ATATA(358) OCT4(365) TCF11(414) SRY(513) HFH8(515) HSF1(559) SRY(619) GFII(659) MTATA(661) YY1(674) VBP(787) PAX2(829) STAT6(837)
CLEC3B	chr3:45028443-45097998	intergenic	4	HMGYIY(16) TITF1(104) NKX25(181) E2F4DP1(228)
MUSTN1	chr3:52840929-52848933	intergenic	5	SRF(96) PITX2(126) SOX9_B1(147) TEF(159) TATA(180)
FILIP1L	chr3:100997862-101460159	UTR5	13	VBP(94) GLI(294) SREBP1(518) TAL1BETAITF2(601) TITF1(617) HEN1(701) MYOGENIN(703) E12(704) LBP1(704) NF1(740) TAL1BETAITF2(899) E12(900) GCNF(1061)
FILIP1L	chr3:100997862-101460159	intergenic	17	HOXA4(77) YY1(95) HAND1E47(95) TAL1BETAE47(95) TAL1BETAITF2(95) MEIS1BHOXA9(115) CBF(180) SRY(264) TEF1(571) HLF(654) VBP(654) PBX1(699) OTX(700) PBX1(736) GFII(738) HMGYIY(756) LBP1(936)
GPR128	chr3:101779325-101906919	intron	3	NCX(100) TGIF(144) XFD2(270)
GPR128	chr3:101779325-101906919	intron	10	TITF1(27) TGIF(61) XBP1(64) SMAD(171) SOX(320) SRY(365) MMEF2(388) HAND1E47(412) ATATA(441) HFH8(442)
ABI3BP	chr3:101950501-102427567	promoter	1	PAX3(181)
CCDC80	chr3:113785882-114017031	intron	10	SRF(130) NFKAPPAB65(198) HMGYIY(199) YY1(304) TAL1BETAE47(312) LBP1(312) STAT1(599) STAT3(599) STAT5A(599) STAT(599)
BOC	chr3:114221281-114564279	promoter	4	STAT5A(129) STAT(129) HSF1(133) SMAD4(158)
FSTL1	chr3:121550880-121797812	intergenic	5	IK3(289) TAL1BETAITF2(301) NKX25(472) YY1(625) NCX(676)
WWTR1	chr3:150704063-150941200	intergenic	10	YY1(119) HAND1E47(150) CDP(163) FREAC2(169) TEF(309) PAX2(322) PBX1(322) TATA(392) SRY(517) TST1(613)
WWTR1	chr3:150704063-150941200	intergenic	23	P53(16) PPARG(81) PPAR_DR1(88) PPARG(90) OTX(249) CDP(260) GATA3(260) TBP(265) XFD2(266) VJUN(301) TATA(359) FOXO3(384) FOXO4(384) TEF(386) HLF(388) VBP(388) HMEF2(580) MMEF2(581) TITF1(590) NKX25(590) RP58(688) GFII(732) RP58(746)
WWTR1	chr3:150704063-150941200	promoter	1	NFKAPPAB(23)
P2RY14	chr3:152403696-152494496	UTR5	2	HMX1(104) GATA3(125)
CCDC50	chr3:192482859-192661629	intron	13	HMGYIY(77) NFKAPPAB65(78) NFE2(86) PEBP(129) PADS(131) CMAF(201) CBF(243) PEBP(243) NF1(264) TCF4(308) CRX(322) TCF4(352) PITX2(383)
APOD	chr3:196751372-196933385	intergenic	9	HSF1(11) HSF1(18) PPARG(93) NKX25(186) EIS1BHOXA9(205) S8(210) TBP(273) TATA(276) STAT(391)
APOD	chr3:196751372-196933385	promoter	10	HMGYIY(96) RP58(157) HAND1E47(158) TAL1BETAE47(158) TAL1BETAITF2(158) HFH3(166) HNF3ALPHA(168) IRF1(175) OCT4(181) STAT1(211)
BDH1	chr3:198509879-198875658	promoter	9	PPARG(55) HNF4_DR1(57) PPAR_DR1(57) HNF3B(79) HFH4(80) HNF3(81) SOX(82) HNF3ALPHA(82) SRY(82)

UCHL1	chr4:40911258-41057560	intergenic	11	PITX2(10) GATA3(338) SRF(343) XPF1(572) TEF(592) HNF3ALPHA(596) HFH3(598) HFH8(598) XFD2(598) HNF3B(599) FOXO4(600)
FRYL	chr4:48193364-48527729	UTR3	28	NCX(84) HAND1E47(95) PXR(128) GFII(141) STAT(178) HMX1(236) NKX25(239) GCNF(240) NRF1(240) STAT5A(309) PAX3(558) PEBP(705) STAT(771) TATA(876) YY1(893) CREBP1CJUN(962) IK2(997) NFKAPPAB65(1001) HMEF2(1085) OTX(1087) P53(1111) P53_DECAMER(1111) NKX25(1194) RUSH1A(1200) SRY(1217) IRF1(1226) SOX9_B1(1255) SMAD4(1295)
SULT1B1	chr4:70548204-70722318	intergenic	2	HAND1E47(165) HNF3B(246)
SULT1B1	chr4:70548204-70722318	intergenic	2	DTYPEPA(94) XPF1(129)
ADH1B	chr4:100435011-100476600	intergenic	1	TGIF(175)
ANK2	chr4:113798453-114593016	promoter	3	TEF(74) P53(79) HSF1(115)
UGT8	chr4:115120359-115968318	promoter	4	PITX2(20) DTYPEPA(42) ATATA(78) OCT4(131)
C4orf33	chr4:130232923-131252648	promoter	1	TAL1BETAITF2(4)
EDNRA	chr4:148086560-148757981	promoter	3	STAT1(4) STAT(4) NF1(19)
SFRP2	chr4:154901557-155374956	intergenic	23	ATATA(121) CRX(203) OTX(204) MYOGNF1(301) E2F4DP1(301) NF1(305) YY1(313) SOX5(426) FOXO4(428) FREAC2(429) NKX61(506) PBX1(507) PAX3(545) CREB(545) GEN_INI(546) TCF11(550) TATA(568) XBP1(607) CBF(613) TATA(622) NKX25(766) E2F4DP1(828) NF1(832)
ETFDH	chr4:159811537-159849486	intron	2	CDP(22) TCF4(30)
PDLIM3	chr4:186630829-186743139	promoter	1	TITF1(29)
SRD5A1	chr5:6685070-6767705	promoter	1	XBP1(17)
C7	chr5:40892557-41033689	promoter	7	NKX25(160) STAT5A(215) STAT(218) RUSH1A(222) LBP1(248) HAND1E47(286) GCNF(312)
NLN	chr5:65052352-65258137	intron	14	PAX3(149) PAX2(151) MMEF2(194) HMEF2(196) HLF(240) VBP(240) TCF4(255) FOXO4(263) HFH8(265) XBP1(301) CREB(362) VBP(363) SMAD4(544) HOXA4(668)
OCLN	chr5:68775709-68891043	intergenic	5	SOX9_B1(225) SRY(225) TITF1(304) NKX25(304) CBF(427)
SERF1B	chr5:68925252-69381082	intergenic	16	FREAC2(110) HFH8(111) XFD2(111) SRY(113) HFH4(115) TBP(225) MTATA(346) OCT4(383) OTX(392) SOX9_B1(429) TCF11(429) VJUN(435) CREBP1CJUN(435) CREB(437) HNF6(475) S8(677)
SERF1B	chr5:69780407-70256500	intergenic	16	FREAC2(110) HFH8(111) XFD2(111) SRY(113) HFH4(115) TBP(225) MTATA(346) OCT4(383) OTX(392) SOX9_B1(429) TCF11(429) VJUN(435) CREBP1CJUN(435) CREB(437) HNF6(475) S8(677)
F2RL2	chr5:75947064-75956450	promoter	3	HMGIIY(22) OTX(101) TEF(168)
DPYSL3	chr5:146708746-146950862	promoter	7	SRF(264) TATA(266) NCX(421) NF1(503) OCT4(534) OCT4(666) HMEF2(684)
PDE6A	chr5:149207465-149319970	UTR5	5	P53_DECAMER(102) SMAD(107) OTX(159) IPF1(166) PPARG(202)
STK38	chr6:36566293-36670127	UTR3	10	GLI(446) NF1(458) XPF1(558) ZTA(603) HSF1(763) STAT(767) STAT5A(770) HSF1(773) PPAR_DR1(822) STAT3(837)
DAAM2	chr6:39801200-39980873	intergenic	16	TBP(49) NF1(102) NFKAPPAB65(260) NFKAPPAB(260) PXR(323) TITF1(343) NKX25(344) PAX2(345) NF1(407) ATATA(434) TCF4(451) PPARG(452) HNF6(484) TEF(492)

				STAT5A(520) XPF1(561)
DAAM2	chr6:39801200-39980873	promoter	2	NF1(78) MYOGNF1(83)
DAAM2	chr6:39801200-39980873	UTR5	4	TGIF(25) HMX1(107) NKX25(108) E12(237)
RCAN2	chr6:46246679-46625276	promoter	3	XBP1(115) RBPJK(146) STAT1(149)
COL12A1	chr6:74859330-76004106	intron	23	MYOGNF1(238) PITX2(266) STAT5A(274) STAT1(282) STAT5A(282) SREBP1(383) TAL1BETAE47(414) TAL1BETAITF2(414) RP58(415) HNF3B(424) NKX61(424) HNF3(425) HNF3ALPHA(427) OCT4(435) NKX61(441) HOXA4(622) NKX25(623) RP58(719) YY1(753) TAL1BETAE47(812) TAL1BETAITF2(812) RP58(813) HOXA4(921)
MYO6	chr6:76486390-76687768	intron	13	SREBP1(216) MTATA(241) SREBP1(306) HOXA3(347) HOXA4(347) NKX25(360) TAL1BETAE47(392) TAL1BETAITF2(392) E12(392) YY1(392) TGIF(407) MYOGENIN(409) LBPI(411)
FYN	chr6:112034024-112481965	intergenic	19	NKX25(31) YY1(139) SRY(188) CREB(377) VJUN(379) CREBP1CJUN(379) PAX3(383) PAX2(385) PBX1(410) HNF3(452) HNF3ALPHA(455) SRY(455) HFH4(457) HNF3B(458) HX10(663) SOX9_B1(743) STAT1(806) GFII(946) FOXO3(985)
C6orf204	chr6:118745578-119263572	intergenic	3	HNF3B(964) SRF(971) HOXA3(975)
PLN	chr6:118974693-118988279	UTR3	4	YY1(45) ZTA(46) IRF1(122) TITF1(248)
PKIB	chr6:122833379-123142183	intron	4	GFII(123) TGIF(315) TITF1(500) XPF1(597)
CTGF	chr6:132254340-132658737	promoter	2	PXR(95) DTYPEPA(331)
C6orf98	chr6:152529594-152532298	UTR3	3	FREAC2(314) SRY(317) FOXP3(370)
SYNE1	chr6:152466144-153060717	intron	11	FREAC2(90) PPARG(144) HNF4_DR1(146) PITX2(194) HNF6(230) PBX1(233) TITF1(266) SRY(279) STAT3(357) TCF11(360) GFII(421)
VIP	chr6:153089775-153331523	promoter	15	STAT5A(141) STAT1(145) STAT4(145) STAT6(145) STAT3(149) STAT5A(149) STAT(149) STAT6(152) NKX25(154) STAT5A(156) CBF(160) CREB(178) TGIF(189) SMAD(210) HAND1E47(211)
VIP	chr6:153089775-153331523	UTR5	6	ATATA(215) TCF11(229) CREB(285) VJUN(287) CREBP1CJUN(287) TATA(335)
QKI	chr6:163656519-164914803	promoter	1	DTYPEPA(26)
SCIN	chr7:12243427-12693002	intergenic	8	MTATA(132) NKX25(299) TCF4(352) HOXA4(357) TAL1BETAE47(367) TAL1BETAITF2(367) GEN_INI(394) NKX25(571)
SNX13	chr7:17355302-18032607	promoter	4	IRF1(106) SOX9_B1(156) SOX5(156) GATA3(242)
SCRN1	chr7:29912325-30019342	intergenic	5	NF1(186) PAX3(205) GFII(220) HNF3B(228) SRF(228)
ZNRF2	chr7:30168932-30430664	intergenic	9	MAZR(130) TEF(210) SMAD(327) STAT(358) TGIF(530) CMAF(531) SOX(538) SRY(538) DTYPEPA(655)
SRI	chr7:87664415-87743276	intergenic	9	SREBP1(41) IRF1(74) STAT6(223) NRF1(231) GFII(237) TATA(308) GCNF(315) PXR(315) IRF1(362)
STEAP1	chr7:88804301-89678821	intergenic	22	HMGYI(201) NKX61(271) HOXA4(281) SRF(322) TCF4(332) STAT5A(546) TATA(551) PAX3(587) DR4(624) TEF1(633) HNF3ALPHA(645) HNF3(646) HFH4(659) OTX(664) PITX2(667) NF1(683) OTX(702) PBX1(742) PAX3(780) STAT5A(855) STAT(860) PXR(966)
AKAP9	chr7:91349365-91578504	UTR3	17	OCT4(50) GFII(54) SOX9_B1(54) IK3(82) ATATA(170) TEF1(228) TEF(234) TBP(267) HMX1(275) GATA3(286) HOXA3(429) DTYPEPA(438) VJUN(495) CREB(497) YY1(501)

				HSF1(525) MYOGNF1(537)
COL1A2	chr7:93471627-93977119	intergenic	17	STAT6(22) NFKAPPAB(23) SRF(72) HOXA3(85) GCNF(226) TCF4(283) HSF1(365) HSF2(365) HLF(376) IPF1(483) STAT1(538) STAT5A(538) MEIS1BHOXA9(718) IPF1(733) HNF6(799) PPARG(806) HOXA4(829)
CALD1	chr7:134016340-134483346	promoter	2	CHX10(206) HOXA4(207)
CLU	chr8:27459024-27547479	promoter	2	NF1(3) NF1(8)
RBPM5	chr8:30160647-30555553	promoter	2	NFKAPPAB(95) PAX9(99)
C8orf4	chr8:39993718-40507263	promoter	3	SOX9_B1(28) SRY(28) TEF(123)
SFRP1	chr8:40874519-41467196	intergenic	22	CBF(71) PEBP(72) PADS(164) PEBP(166) SOX(279) HOXA3(292) TATA(385) MEIS1BHOXA9(388) NCX(394) NKX61(395) HOXA4(395) SRF(442) TEF(445) TEF(471) GATA3(480) TALIBETAE47(532) TCF4(545) PADS(567) NKX25(625) NKX25(630) CBF(766) PADS(768)
SULF1	chr8:69893970-70747119	promoter	10	IPF1(94) SOX5(153) IRF1(266) VBP(271) MEIS1BHOXA9(288) HOXA4(430) S8(431) SOX5(495) SMAD4(577) SMAD(581)
TPD52	chr8:81105082-81561002	intergenic	22	GFI1(39) PADS(41) TALIBETAE47(132) TALIBETAITF2(132) NF1(148) SMAD(155) HAND1E47(156) SREBP1(236) DR4(369) NF1(417) SRF(495) TATA(496) RUSH1A(543) MTATA(549) NF1(618) NFKAPPAB65(643) TEF1(644) TALIBETAE47(736) TALIBETAITF2(736) E12(736) E2A(737) SREBP1(823)
FABP4	chr8:82536405-82599831	UTR3	5	HMGY1(81) HLF(146) VBP(146) SOX9_B1(264) SOX5(264)
GEM	chr8:95290022-95452022	promoter	2	OTX(79) NF1(142)
OSR2	chr8:99907087-100094669	intergenic	19	SREBP1(156) FREAC2(267) HFH8(268) FOXO4(270) RP58(287) STAT5A(443) STAT6(443) RBPJK(488) IK1(490) STAT3(491) TALIBETAE47(545) TALIBETAITF2(545) RP58(546) CHX10(836) S8(838) ATATA(838) NKX25(838) GATA3(856) XPF1(945)
COL14A1	chr8:121132222-121474533	intron	2	TCF11(288) NF1(377)
ANXA13	chr8:124734818-124844526	intron	14	SOX9_B1(68) MMEF2(74) SRF(142) TEF1(181) YY1(193) SOX9_B1(234) VJUN(247) TEF1(252) SMAD(287) NRF1(300) STAT5A(318) STAT6(318) RP58(336) TALIBETAE47(337)
KIAA1161	chr9:34333736-34369009	UTR3	6	YY1(187) SRF(188) MTATA(233) PPARG(278) HNF4_DR1(278) PPAR_DR1(278)
TPM2	chr9:35671385-35687219	promoter	3	GFI1(16) XPF1(33) RBPJK(41)
MAMDC2	chr9:71711006-72063692	intron	12	IRF1(192) YY1(280) TALIBETAE47(280) TALIBETAITF2(280) E12(280) PADS(430) LDSPOLYA(453) STAT5A(460) STAT6(460) CBF(508) STAT5A(548) NFE2(591)
ANXA1	chr9:74757814-75974291	promoter	2	STAT(22) CDP(93)
GAS1	chr9:88159211-88953226	promoter	6	IK1(51) TEF1(52) TALIBETAE47(70) E12(70) E2A(71) MYOGENIN(71)
AUH	chr9:92698993-93209340	intergenic	11	TALIBETAE47(54) TALIBETAITF2(54) NCX(60) HMGY1(62) NCX(93) TGIF(165) MEIS1BHOXA9(169) TGIF(205) TALIBETAE47(301) TALIBETAITF2(301) E12(301)
OGN	chr9:94127716-94215239	intron	3	TCF11(58) HSF1(130) OTX(192)
OGN	chr9:94127716-94215239	UTR5	2	NKX25(9) HMX1(10)
NIPSNAP3A	chr9:106497952-106566234	intergenic	1	SRF(111)
ANGPTL2	chr9:128687979	intergenic	13	SRF(118) PAX2(262) SRY(439) ATATA(462) OTX(463)



	-129066949			TGIF(487) RBPJK(520) HOXA3(561) PXR(727) HAND1E47(752) GATA3(772) SMAD(788) PAX3(791)
OLFM1	chr9:136952182 -137511463	promoter	3	SOX9_B1(45) SRY(45) SOX5(45)
GPM6B	chrX:13697432- 13934192	promoter	4	HMX1(170) DTYPEPA(193) HSF1(243) HSF2(243)
ACE2	chrX:15484647- 15555305	promoter	8	GEN_INI(20) MTATA(68) TATA(74) YY1(75) FOXO4(150) P53(151) P53_DECAMER(151) CDP(181)
AP1S2	chrX:15751696- 16051243	promoter	4	FOXO4(110) P53(111) CMAF(115) NRF1(142)
SRPX	chrX:37871586- 38013303	intron	9	HAND1E47(162) SREBP1(222) OTX(227) MEIS1BHOXA9(276) NKX25(279) GATA3(319) SRF(324) SREBP1(416) SOX(682)
MAOA	chrX:42400389- 43510720	intergenic	6	NKX25(165) STAT1(272) RBPJK(275) NKX25(631) TCF4(949) NF1(1022)
POF1B	chrX:84415033- 85002640	promoter	4	HAND1E47(52) HAND1E47(60) HOXA4(96) TBP(108)
TCEAL7	chrX:102453100 -102498021	promoter	2	TEF1(18) FOXO4(156)
TCEAL7	chrX:102453100 -102498021	promoter	1	OTX(191)
PLP1	chrX:102866579 -102966545	promoter	2	YY1(23) ZNF219(83)
PLP1	chrX:102866579 -102966545	promoter	2	P53_DECAMER(4) P53_DECAMER(94)
CHRD1	chrX:109586328 -110226227	promoter	4	SRF(43) RP58(104) TAL1BETAE47(105) TAL1BETAITF2(105)
FLNA	chrX:153212003 -153260929	promoter	1	SMAD4(27)

## 5. General description regarding TFs including their class and families.

S. No	Transcription Factor	JASPAR ID	Class	Family	Target Gene hits
1	HOXA5	MA0158.1	Helix-Turn-Helix	Homeo	447
2	Nkx2-5	MA0063.1	Helix-Turn-Helix	Homeo	435
3	ARID3A	MA0151.1	Helix-Turn-Helix	Arid	424
4	SRY	MA0084.1	Other Alpha-Helix	High Mobility Group	404
5	Pdx1	MA0132.1	Helix-Turn-Helix	Homeo	412
6	Foxd3	MA0041.1	Winged Helix-Turn-Helix	Forkhead	332
7	FOXI1	MA0042.1	Winged Helix-Turn-Helix	Forkhead	336
8	Sox5	MA0087.1	Other Alpha-Helix	High Mobility Group	383
9	FOXA1	MA0148.1	Winged Helix-Turn-Helix	Forkhead	384
10	FOXO3	MA0157.1	Winged Helix-Turn-Helix	Forkhead	388
11	NFATC2	MA0152.1	Ig-fold	Rel	412
12	Prrx2	MA0075.1	Helix-Turn-Helix	Homeo	402
13	TBP	MA0108.2	Beta-sheet	TATA-binding	315
14	CEBPA	MA0102.2	Zipper-Type	Leucine Zipper	368
15	Gfi	MA0038.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	383
16	NKX3-1	MA0124.1	Helix-Turn-Helix	Homeo	354
17	FOXD1	MA0031.1	Winged Helix-Turn-Helix	Forkhead	381
18	SOX9	MA0077.1	Other Alpha-Helix	High Mobility Group	332
19	Nobox	MA0125.1	Helix-Turn-Helix	Homeo	377
20	AP1	MA0099.2	Zipper-Type	Leucine Zipper	426
21	Foxa2	MA0047.2	Winged Helix-Turn-Helix	Forkhead	341
22	Sox17	MA0078.1	Other Alpha-Helix	High Mobility Group	384
23	Foxq1	MA0040.1	Winged Helix-Turn-Helix	Forkhead	265
24	Gata1	MA0035.2	Zinc-coordinating	GATA	379
25	IRF1	MA0050.1	Winged Helix-Turn-Helix	IRF	206
26	NFIL3	MA0025.1	Zipper-Type	Leucine Zipper	220
27	ELF5	MA0136.1	Winged Helix-Turn-Helix	Ets	428
28	MEF2A	MA0052.1	Other Alpha-Helix	MADS	237
29	Pou5f1	MA0142.1	Helix-Turn-Helix	Homeo	135
30	SPIB	MA0081.1	Winged Helix-Turn-Helix	Ets	443
31	HNF1B	MA0153.1	Helix-Turn-Helix	Homeo	154
32	HLF	MA0043.1	Zipper-Type	Leucine Zipper	190
33	SRF	MA0083.1	Other Alpha-Helix	MADS	55
34	Lhx3	MA0135.1	Helix-Turn-Helix	Homeo	190
35	YY1	MA0095.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	435
36	Hand1::Tcf2a	MA0092.1	Zipper-Type	Helix-Loop-Helix	377
37	RUNX1	MA0002.2	Ig-fold	Runt	376

38	STAT1	MA0137.2	Ig-fold	Stat	224
39	FEV	MA0156.1	Winged Helix-Turn-Helix	Ets	410
40	FOXF2	MA0030.1	Winged Helix-Turn-Helix	Forkhead	156
41	PBX1	MA0070.1	Helix-Turn-Helix	Homeo	144
42	Sox2	MA0143.1	Other Alpha-Helix	High Mobility Group	107
43	TEAD1	MA0090.1	Helix-Turn-Helix	Homeo	200
44	TAL1::TCF3	MA0091.1	Zipper-Type	Helix-Loop-Helix	236
45	Nkx3-2	MA0122.1	Helix-Turn-Helix	Homeo	408
46	SPI1	MA0080.2	Winged Helix-Turn-Helix	Ets	405
47	Pax4	MA0068.1	Helix-Turn-Helix	Homeo	8
48	REL	MA0101.1	Ig-fold	Rel	303
49	Ddit3::Cebpa	MA0019.1	Zipper-Type	Leucine Zipper	171
50	RELA	MA0107.1	Ig-fold	Rel	223
51	Tal1::Gata1	MA0140.1	Zipper-Type	Helix-Loop-Helix	192
52	NFE2L2	MA0150.1	Zipper-Type	Leucine Zipper	192
53	Nr2e3	MA0164.1	Zinc-coordinating	Hormone-nuclear Receptor	191
54	Stat3	MA0144.1	Ig-fold	Stat	311
55	IRF2	MA0051.1	Winged Helix-Turn-Helix	IRF	37
56	HNF1A	MA0046.1	Helix-Turn-Helix	Homeo	114
57	RORA_2	MA0072.1	Zinc-coordinating	Hormone-nuclear Receptor	125
58	Ar	MA0007.1	Zinc-coordinating	Hormone-nuclear Receptor	23
59	NR3C1	MA0113.1	Zinc-coordinating	Hormone-nuclear Receptor	83
60	Myb	MA0100.1	Helix-Turn-Helix	Myb	369
61	RREB1	MA0073.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	61
62	NR4A2	MA0160.1	Zinc-coordinating	Hormone-nuclear Receptor	364
63	Arnt::Ahr	MA0006.1	Zipper-Type	Helix-Loop-Helix	360
64	Evil	MA0029.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	64
65	CREB1	MA0018.2	Zipper-Type	Leucine Zipper	278
66	NR1H2::RXRA	MA0115.1	Zinc-coordinating	Hormone-nuclear Receptor	5
67	NR2F1	MA0017.1	Zinc-coordinating	Hormone-nuclear Receptor	119
68	MAX	MA0058.1	Zipper-Type	Helix-Loop-Helix	246
69	Pax6	MA0069.1	Helix-Turn-Helix	Homeo	42
70	USF1	MA0093.1	Zipper-Type	Helix-Loop-Helix	276
71	T	MA0009.1	Beta-Hairpin-Ribbon	T	66
72	ESR1	MA0112.2	Zinc-coordinating	Hormone-nuclear Receptor	14
73	EWSR1-FLI1	MA0149.1	Winged Helix-Turn-Helix	Ets	8
74	MYC::MAX	MA0059.1	Zipper-Type	Helix-Loop-Helix	109
75	TLX1::NFIC	MA0119.1	Helix-Turn-Helix::Other	Homeo::Nuclear Factor I-CCAAT-binding	38
76	RORA_1	MA0071.1	Zinc-coordinating	Hormone-nuclear Receptor	245
77	NF-kappaB	MA0061.1	Ig-fold	Rel	244
78	E2F1	MA0024.1	Winged Helix-Turn-Helix	E2F	241

79	HNF4A	MA0114.1	Zinc-coordinating	Hormone-nuclear Receptor	193
80	REST	MA0138.2	Zinc-coordinating	BetaBetaAlpha-zinc finger	7
81	RXRA::VDR	MA0074.1	Zinc-coordinating	Hormone-nuclear Receptor	14
82	EBF1	MA0154.1	Zipper-Type	Helix-Loop-Helix	325
83	TP53	MA0106.1	Zinc-coordinating	Loop-Sheet-Helix	0
84	Spz1	MA0111.1	Other	Other	194
85	RXR::RAR_DR5	MA0159.1	Zinc-coordinating	Hormone-nuclear Receptor	41
86	PPARG	MA0066.1	Zinc-coordinating	Hormone-nuclear Receptor	1
87	ESR2	MA0258.1	Zinc-coordinating	Hormone-nuclear Receptor	68
88	NFKB1	MA0105.1	Ig-fold	Rel	127
89	Myc	MA0147.1	Zipper-Type	Helix-Loop-Helix	244
90	CTCF	MA0139.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	114
91	PPARG::RXRA	MA0065.2	Zinc-coordinating	Hormone-nuclear Receptor	193
92	Arnt	MA0004.1	Zipper-Type	Helix-Loop-Helix	210
93	Esrrb	MA0141.1	Zinc-coordinating	Hormone-nuclear Receptor	277
94	MIZF	MA0131.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	67
95	Myf	MA0055.1	Zipper-Type	Helix-Loop-Helix	296
96	NHLH1	MA0048.1	Zipper-Type	Helix-Loop-Helix	180
97	Egr1	MA0162.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	182
98	Pax5	MA0014.1	Helix-Turn-Helix	Homeo	36
99	ZNF354C	MA0130.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	443
100	Mycn	MA0104.2	Zipper-Type	Helix-Loop-Helix	242
101	MZF1_5-13	MA0057.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	365
102	PLAG1	MA0163.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	57
103	MZF1_1-4	MA0056.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	439
104	ELK1	MA0028.1	Winged Helix-Turn-Helix	Ets	366
105	NFYA	MA0060.1	Other Alpha-Helix	NFY CCAAT-binding	172
106	INSM1	MA0155.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	216
107	ELK4	MA0076.1	Winged Helix-Turn-Helix	Ets	162
108	HIF1A::ARNT	MA0259.1	Zipper-Type	Helix-Loop-Helix	311
109	ZEB1	MA0103.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	439
110	Tcfcp2l1	MA0145.1	Other	CP2	294
111	Zfp423	MA0116.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	166
112	znf143	MA0088.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	26
113	GABPA	MA0062.2	Winged Helix-Turn-Helix	Ets	257
114	SP1	MA0079.2	Zinc-coordinating	BetaBetaAlpha-zinc finger	342
115	Klf4	MA0039.2	Zinc-coordinating	BetaBetaAlpha-zinc finger	374
116	Zfx	MA0146.1	Zinc-coordinating	BetaBetaAlpha-zinc finger	274

*“What you are is God’s gift to you, what you become is your gift to God.” -Unknown author*

# PUBLICATIONS AND PRESENTATIONS

## PAPERS IN INTERNATIONAL REFEREED JOURNALS:

**Manika Sehgal and Tiratha Raj Singh.** DR-GAS: a database of functional genetic variants and their phosphorylation states in human DNA repair systems. *DNA Repair (Amst)*, vol. 16, pp. 97-103, 2014. [ISSN: 1568-7864, **IF: 4.274**]

**Manika Sehgal and Tiratha Raj Singh.** Systems biology approach for mutational and site-specific structural investigation of DNA repair genes for xeroderma pigmentosum. *Gene*, vol. 543, pp. 108-117, 2014. [ISSN: 0378-1119, **IF: 2.196**]

**Manika Sehgal and Tiratha Raj Singh.** Computational Approach for the Identification of Plausible Biomarkers from Composite Networks and Gene Expression data Associated with Colorectal Cancer. *International Journal of Basic and Applied Biology*, vol. 1, pp. 62-66, 2014. [ISSN: 2349-5839]

**Manika Sehgal and Tiratha Raj Singh.** Identification and analysis of biomarkers for mismatch repair proteins: A bioinformatic approach. *Journal of Natural Science, Biology and Medicine*, vol. 3, pp. 139-146, 2012. [ISSN: 2229-7707]

**Manika Sehgal, Rajinder Gupta, Ahmed Moussa and Tiratha Raj Singh.** An integrative approach for mapping differentially expressed genes and network components using novel parameters to elucidate key regulatory genes in colorectal cancer. (Under Revision in PLoS ONE) [**IF: 3.5**]

## PRESENTATIONS IN NATIONAL AND INTERNATIONAL CONFERENCES:

**Manika Sehgal and Tiratha Raj Singh.** “Identification and analysis of biomarkers for mismatch repair proteins: A bioinformatic approach”. **Oral presentation** at ‘The International Interdisciplinary Science Conference (I-ISC, 2011) on Bioinformatics: An interface between Computer Science and Biology’ in *Jamia Millia Islamia, New Delhi, INDIA* from 15-17, November 2011.

**Manika Sehgal and Tiratha Raj Singh.** “Quantitative Genetic Analysis of DNA Repair Genes Implicated in Xeroderma Pigmentosum”. **Oral presentation** at ‘The Third International Federation for Information Processing (IFIP) International Conference on Bioinformatics’ in *Maulana Azad National Institute of Technology (MANIT), Bhopal, Madhya Pradesh, INDIA* from 23-26, September 2013.

**Manika Sehgal and Tiratha Raj Singh.** “Computational Approach for the Identification of Plausible Biomarkers from Composite Networks and Gene Expression data Associated with Colorectal Cancer”. **Oral presentation** at ‘World Congress on Stem Cell Research, Cancer Biology and Applied Biotechnology (Biotech-2014)’ in *Jawaharlal Nehru University, New Delhi, INDIA* from 3-4, May 2014.

**Manika Sehgal, Ankita Shukla and Tiratha Raj Singh.** “Functional Enrichment of Pathways Implicated in DNA Repair using Top Down Approach”. **Oral presentation** at ‘International Conference on Life Sciences, Informatics, Food and Environment (IC LIFE 2014)’ in *Jaypee Institute of Information Technology (JIIT), Noida, Uttar Pradesh, INDIA* from 29-30, August 2014.

**Manika Sehgal and Tiratha Raj Singh.** “Bioinformatics Applications to Resolve the Mystery of Diseases Associated with Human DNA Repair System”. **Oral presentation** at ‘35th IABMS Annual Conference on Environment and Health’ in *CSK Himachal Pradesh Agricultural University, Palampur, H.P., INDIA* from 14-16, November 2014.

**Manika Sehgal and Tiratha Raj Singh.** “Decoding the intricate biological pathways in quest of biomarkers implicated in human DNA repair system”. **Poster presentation** at ‘Recent Advances in Biological Sciences- Annual Conference of Society of Young Scientists (SYS)’ in *All India Institute of Medical Sciences (AIIMS), New Delhi, INDIA* on 10 December 2014.

## **WORKSHOPS ATTENDED:**

A “**Bioinformatics workshop**” organized jointly by **University of Nebraska**, Omaha, USA and **Jaypee University of Information Technology**, Solan, H.P., India from 8-10, May 2013.

National workshop on “*In silico* approaches for designing bioactive peptides” at **Institute of Microbial Technology (IMTECH)**, Chandigarh, India from 18-21, October 2011.

Workshop on “**Data Mining and Parallel Computing**” organized jointly by **University of Florida** and **Jaypee University of Information Technology** at Wahnaghat, Solan, H.P., India from 2-5, August 2011.

### **ACADEMIC ACHIVEMENTS:**

**Resource Person** in workshop on “The 19th Workshop on Pathways and System-Integrated Approaches” held during 16-19, September 2013 (Sponsored by the Dept. of Biotechnology, Govt. of India) at Bioinformatics Centre, Himachal Pradesh University (HPU), Shimla (H.P.), India.

**Resource Person** in Scripting Languages module for one month Summer Training held during June-July 2013 at Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology (JUIT), Wahnaghat, Solan (H.P.), India.